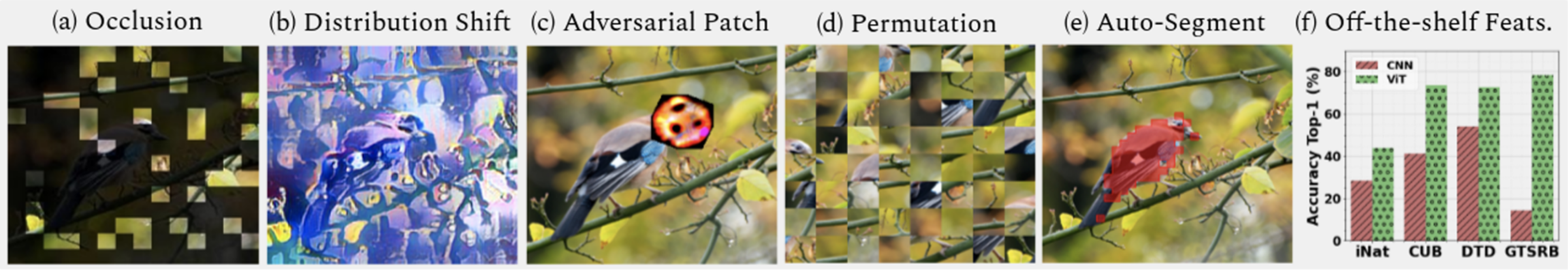
Intriguing Properties of Vision Transformers

Muzammal Naseer[1,2], kanchana Ranasinghe[3], Salman Khan[2], Munawar Hayat[4], Fahad Khan[2,5], Ming-Hsuan[6]

[1]Australian National University, Australia
[2]Mohamed bin Zayed University of Artificial Intelligence, UAE
[3]Stony Brook University
[4]Monash University, Australia
[5]Linkoping University, Sweden
[6]Google, USA



Summary

- Three ViT Families (ViT, Deit, T2T) vs CNN (ResNet-50)
- ViTs show better robustness against
  - Severe occlusions (upto 60% accuracy once 80% occluded)
  - Perturbations (permutations, adversarial noise, natural corruptions)
- ViTs are less biased towards local textures
- ViTs with shape bias can segment without pixel-level supervision

- Generalization
  - Off-the-shelf ViT features transfer well for few-shot and traditional classification
  - Better out of domain generalization

Vision Transformer (ViT)

- Image → Patches
- Tokens: Flattened Patches
- Multi-head self-attention blocks
- Each patch attends to all other patches

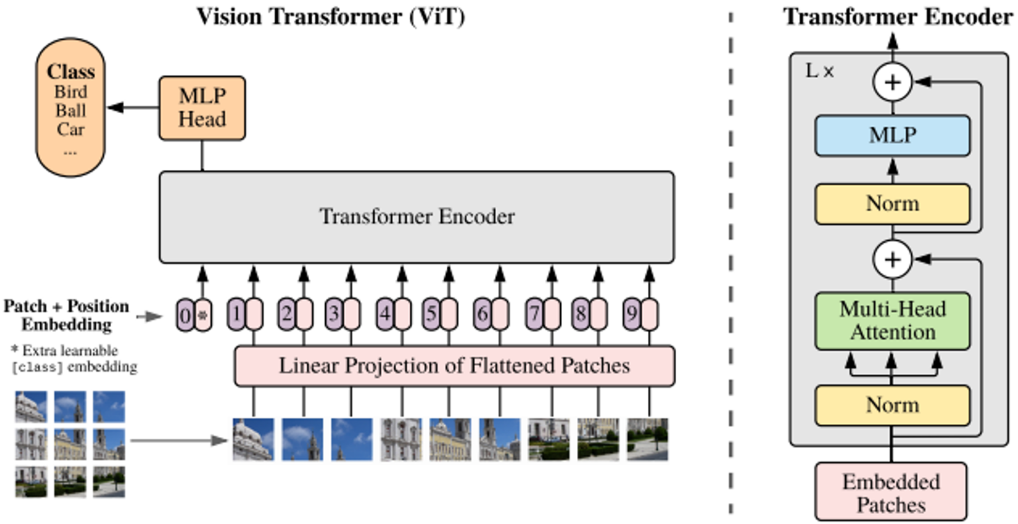


Fig. from Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929*.

Convolution vs Self-attention

- Compare ViTs with CNNs for **robustness** and **generalization**
    - occlusions, distributional shifts, adversarial and natural perturbations

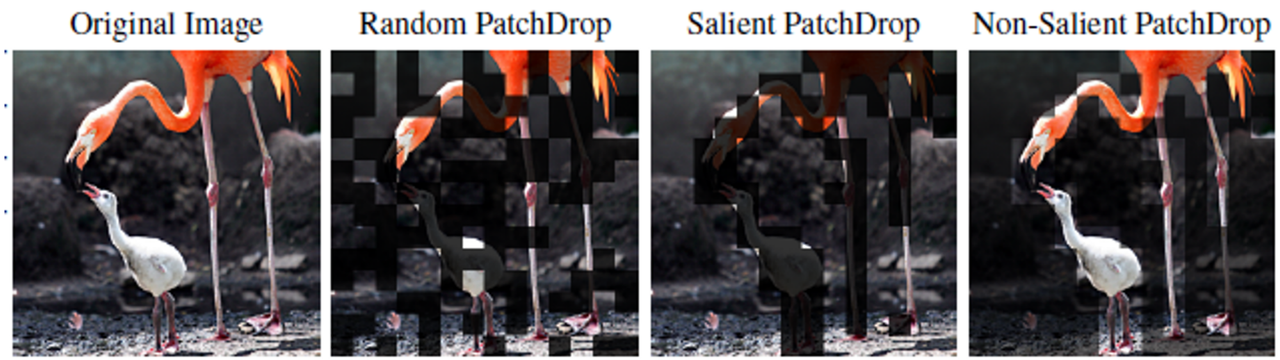
Convolution	Self-attention
Local-relationships (edges, contours)	Global interactions (b/w distant parts)
Content independent	Content Dependent
Designed to capture inductive biases	Designed to model relations in sequence

Are ViTs Robust to Occlusions?

An image: A sequence of N patches. Drop M patches.

Information Loss (IL) = M/N

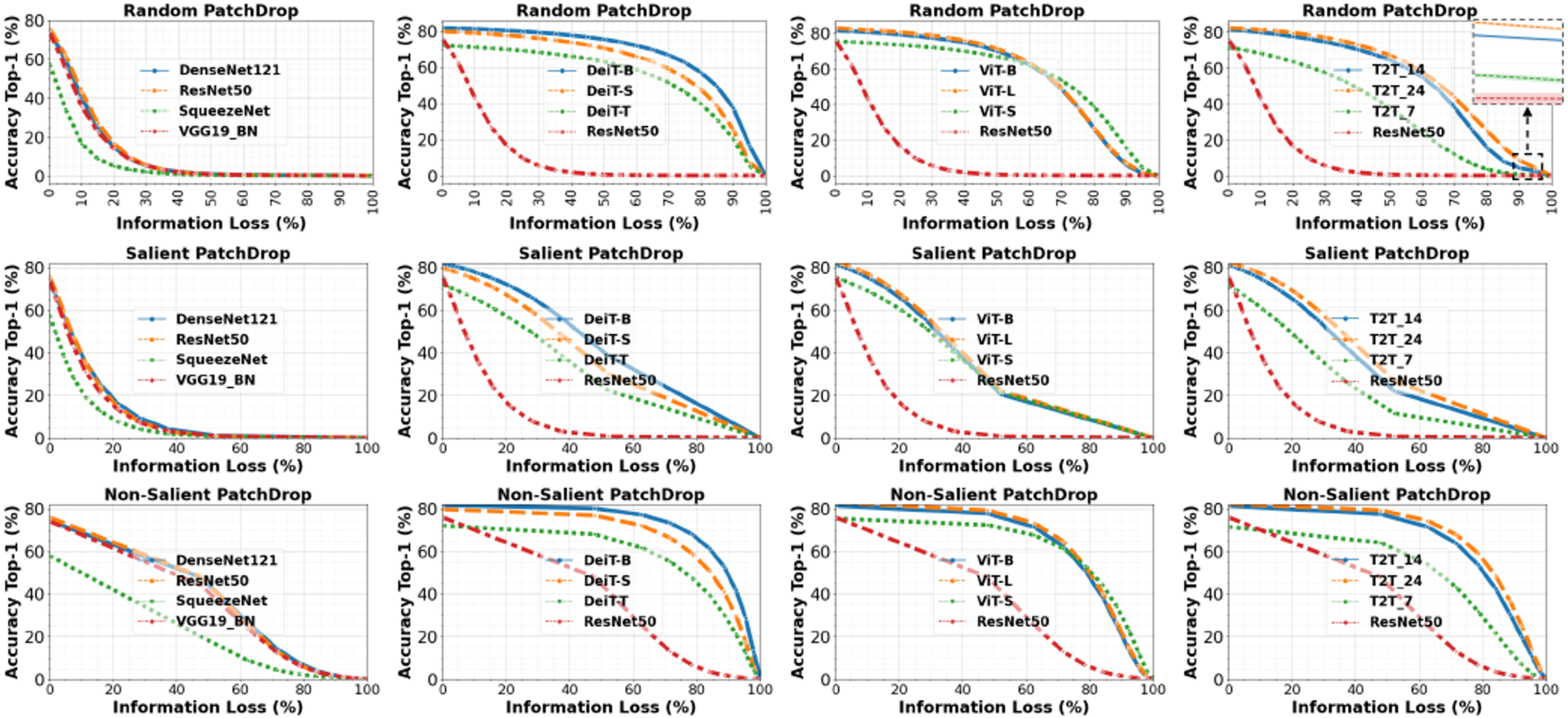
1. Random PatchDrop: Randomly drop patches
2. Salient (foreground) PatchDrop: Drop patches with most salient information
3. Non-salient (background) PatchDrop: Drop patches with least salient information



Example: An image of size 224*224*3 is split into 196 patches, each of size 16*16*3. As an example, dropping 100 such patches from the input is equivalent to losing 51% of the image content.

Are ViTs Robust to Occlusions?

Evaluations on ImageNet val. Set (50k images)

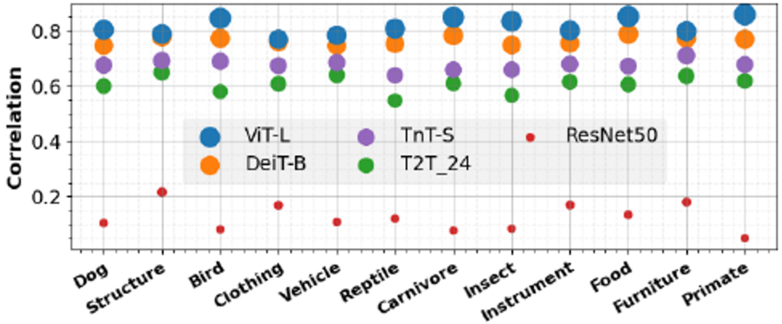


ViT's Features are Robust to Information Loss

Correlation b/w features
- occluded vs non-occluded images

ResNet: features before logit layer
ViT: Class Token of last block

Model	Correlation Coefficient: Random PatchDrop		
	25% Dropped	50% Dropped	75% Dropped
ResNet50	0.32±0.16	0.13±0.11	0.07±0.09
TnT-S	0.83±0.08	0.67±0.12	0.46±0.17
ViT-L	**0.92±0.06**	**0.81±0.13**	0.50±0.21
Deit-B	0.90±0.06	0.77±0.10	**0.56±0.15**
T2T-24	0.80±0.10	0.60±0.15	0.31±0.17



- Visualize the attention maps
- Initial layers attend to all areas
- Deeper layers focus more on leftover (non-occluded) regions

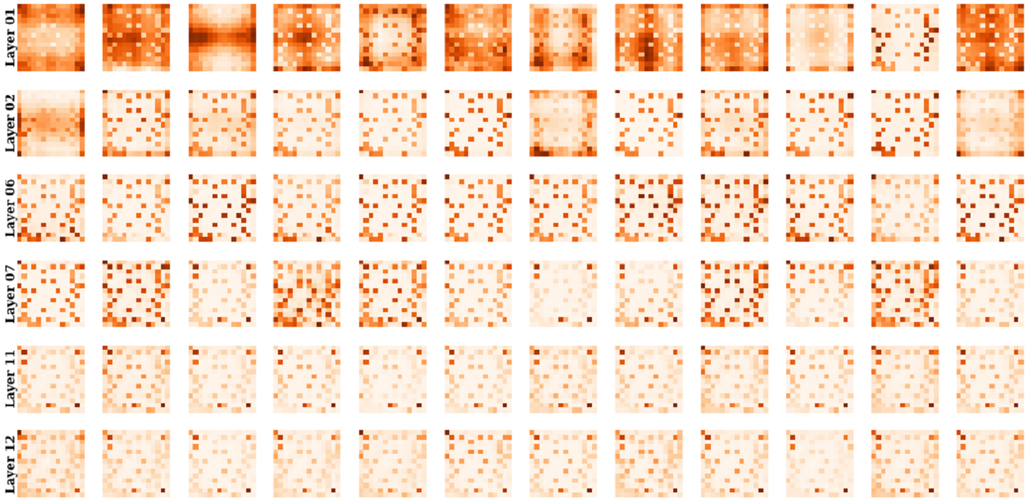


Figure 4: Attention maps (averaged over the entire ImageNet val. set) relevant to each head in multiple layers of an ImageNet pre-trained DeiT-B model. All images are occluded (RandomPatchDrop) with the same mask (bottom right). Observe how later layers clearly attend to non-occluded regions of images to make a decision, an evidence of the model's highly dynamic receptive field.

Shape vs. Texture: Can Transformer Model Both?

- CNNs are biased towards texture than shape; while humans are more biased towards shapes

(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." ICLR'19

Shape vs. Texture: Can Transformer Model Both?

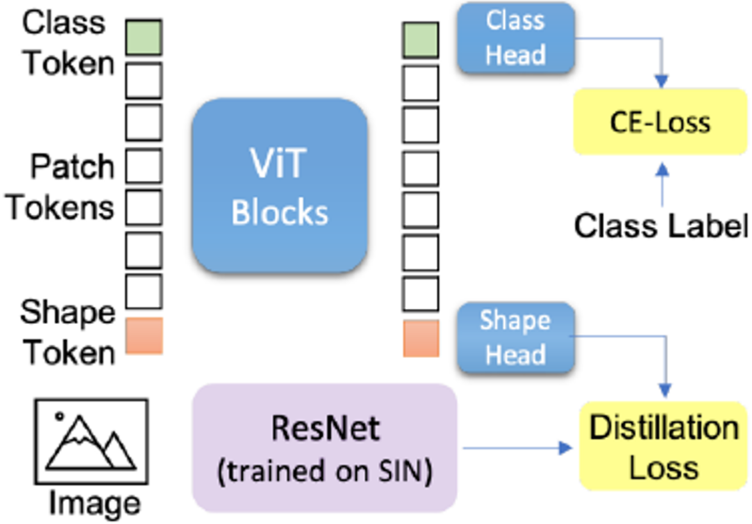
Training without local texture - Stylized ImageNet (SIN)

- Trained ViTs and ResNets on SIN
  - No heavy augmentations (mixup)

Stylized ImageNet (SIN): Textures are highly distorted

Knowledge Distillation from a shape model

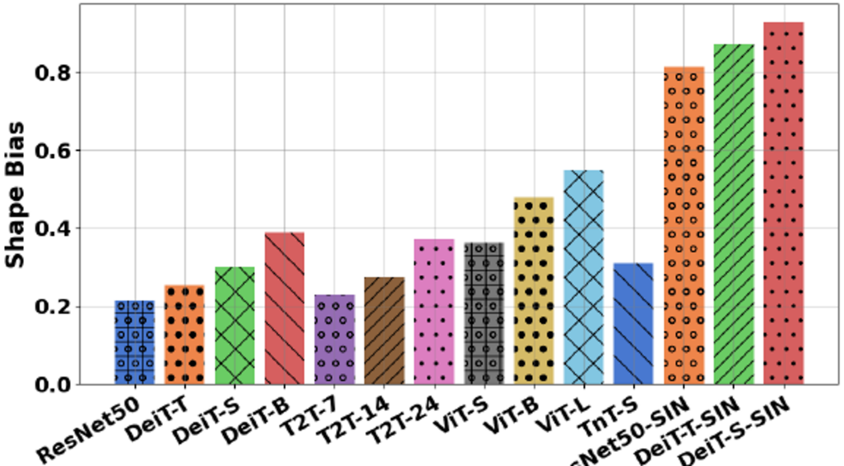
- Additional Shape Token to distill knowledge from ResNet50-SIN

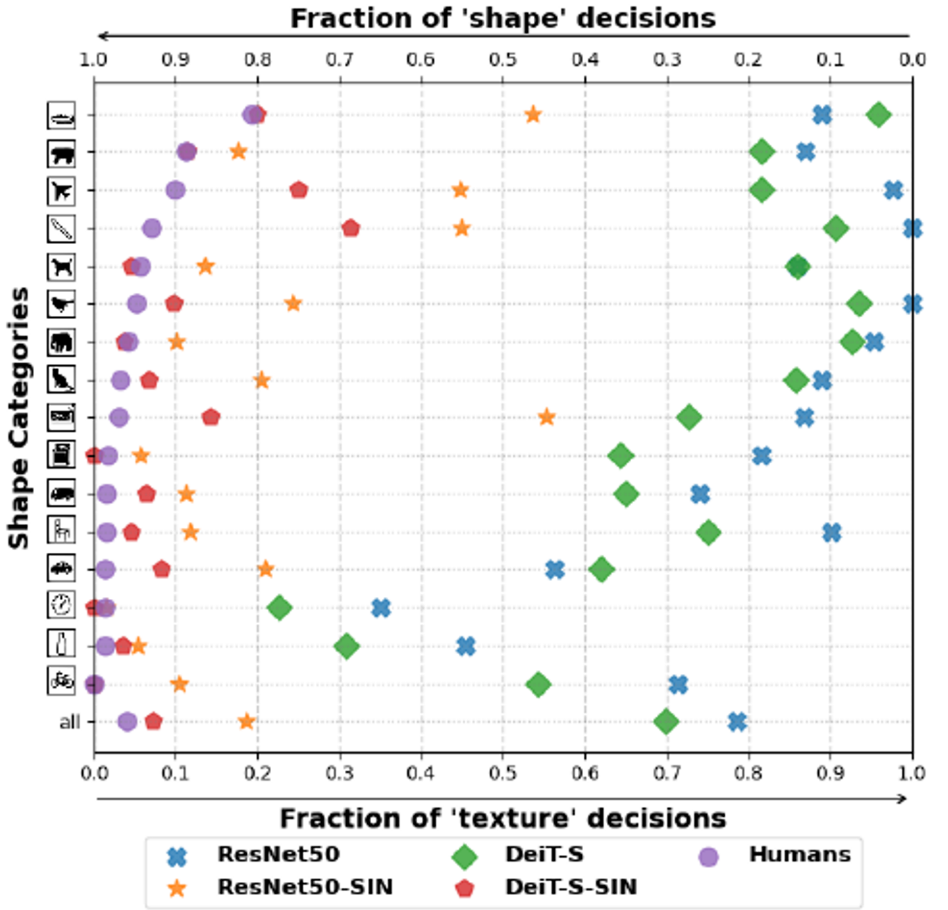


Shape Bias Analysis

Fraction of decisions based on either shape or texture
- ViTs have shape bias comparable to Humans

Class-mean shape bias.
- ViTs better than CNNs
- Training on SIN increases shape bias

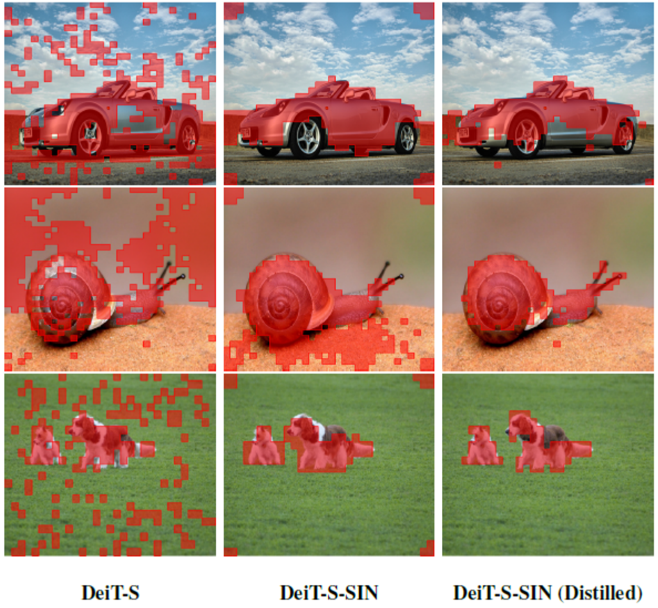




Shape-biased ViT -- Automated Segmentation

- ViTs concentrate on the foreground & ignore the background once trained with distorted texture
- Automated Segmentation without pixel-level supervision
- Jaccard similarity between ground truth and masks generated from the attention maps of ViT models
  - PASCAL-VOC12 validation set.
- DINO - A similar behaviour is observed

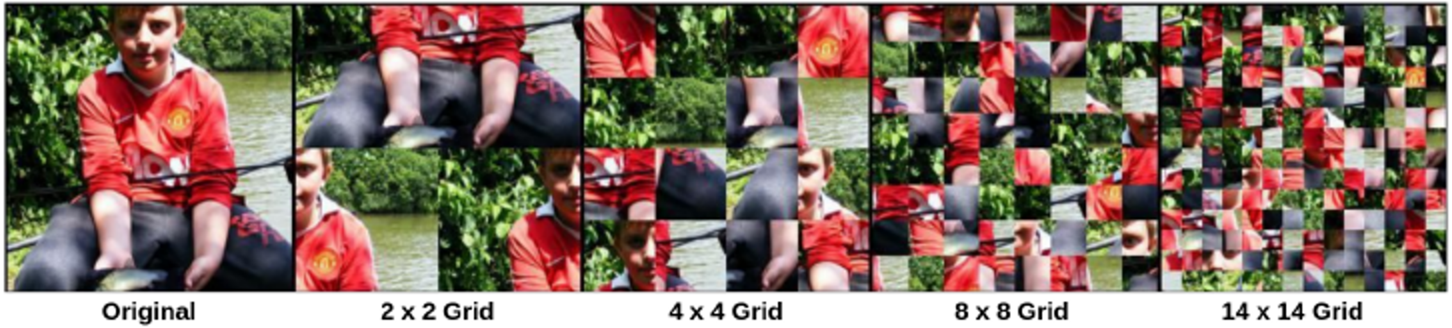
Model	Distilled	Token Type	Jaccard Index
DeiT-T-Random	✗	cls	19.6
DeiT-T	✗	cls	32.2
DeiT-T-SIN	✗	cls	29.4
DeiT-T-SIN	✓	cls	40.0
DeiT-T-SIN	✓	shape	42.2
DeiT-S-Random	✗	cls	22.0
DeiT-S	✗	cls	29.2
DeiT-S-SIN	✗	cls	37.5
DeiT-S-SIN	✓	cls	42.0
DeiT-S-SIN	✓	shape	42.4



Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." *arXiv preprint arXiv:2104.14294* (2021).

Does Positional Encoding Preserve the Global Image Context?

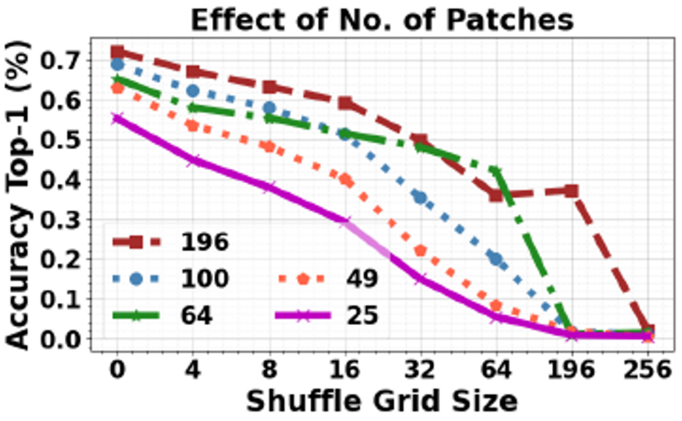
- Self-attention is invariant to sequence order
  - ViTs use Positional Encoding for spatial context
- Do ViTs excel under occlusions because of positional encoding?
  - Effect of position encoding towards injecting structure is limited

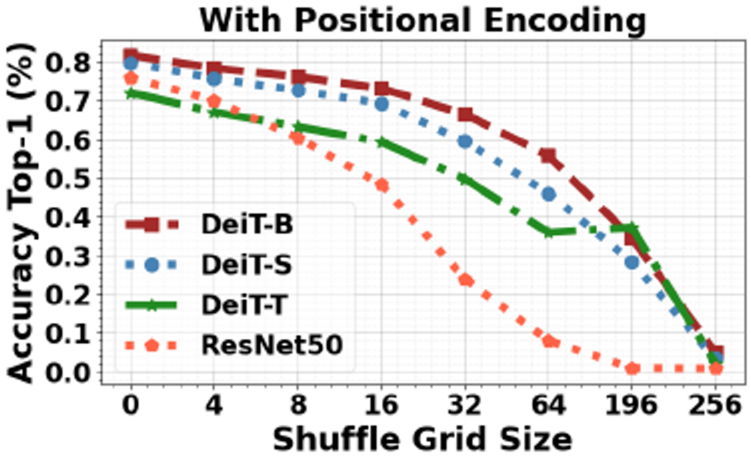


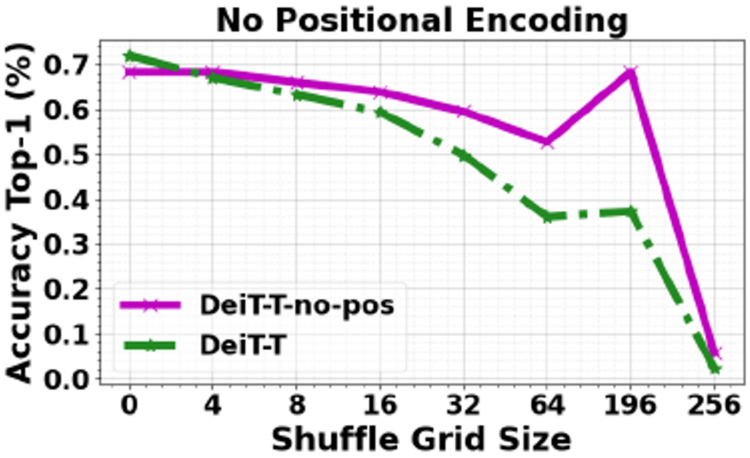
Shuffle Patches i.e., Randomly permute them - Destroy the spatial structure

ViTs - Context - Position Encoding

- After Shuffling, ViTs better retains accuracy than CNN
- Positional Embedding is not absolutely crucial to recover global context
- w/o encoding, ViT achieves better permute invariance
- More patches help: accuracy + less-sensitive to shuffling







Robustness to Natural Perturbations

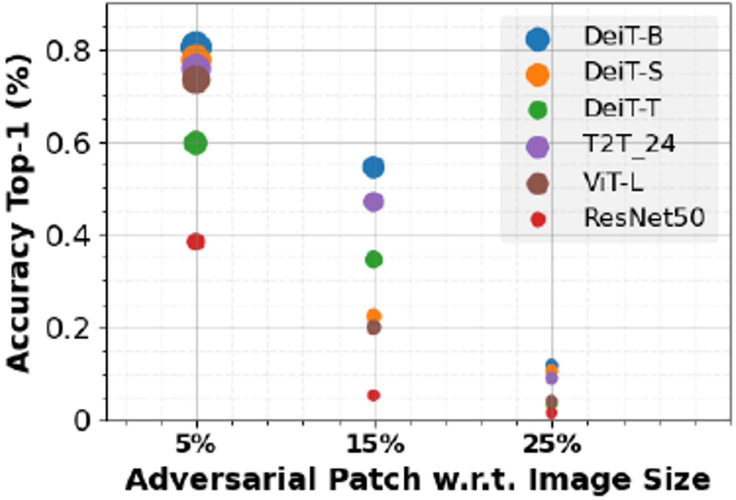
Mean corruption error on synthetic common corruptions (e.g., rain, fog, snow and noise). Lower the better.

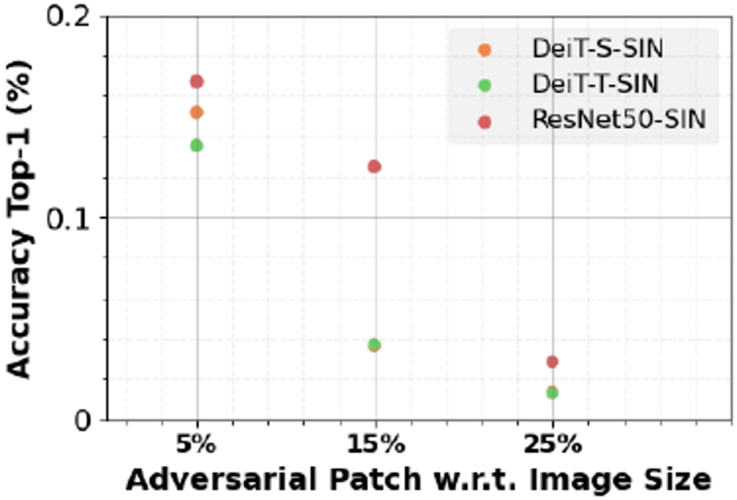
- ViTs show better robustness against natural perturbations than CNNs
- Training on SIN to achieve higher shape bias makes both CNNs and ViTs vulnerable to perturbations.
- Data Augmentation helps for both CNNs and ViTs. Augmix: ResNet50 trained with augmentations

Trained with Augmentations						Trained without Augmentation			
DeiT-B	DeiT-S	DeiT-T	T2T-24	TnT-S	Augmix	ResNet50	ResNet50-SIN	DeiT-T-SIN	DeiT-S-SIN
48.5	54.6	71.1	49.1	53.1	65.3	76.7	77.3	94.4	84.0

Robustness to Adversarial Perturbations

- Robustness against adversarial patch attack (untargeted, universal patch in white-box setting) [A]
- ViTs exhibit better adversarial robustness
- ImageNet trained models are more robust than SIN, shape-bias vs robustness tradeoff [B]





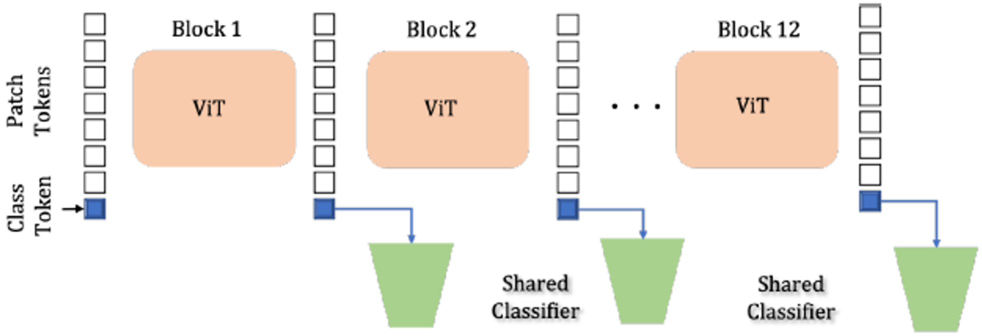
[A] Brown, Tom B., et al. "Adversarial patch." *arXiv preprint arXiv:1712.09665* (2017).
[B] Mummadi etal, “Does enhanced shape bias improve neural network robustness to common Corruptions?” ICLR’21

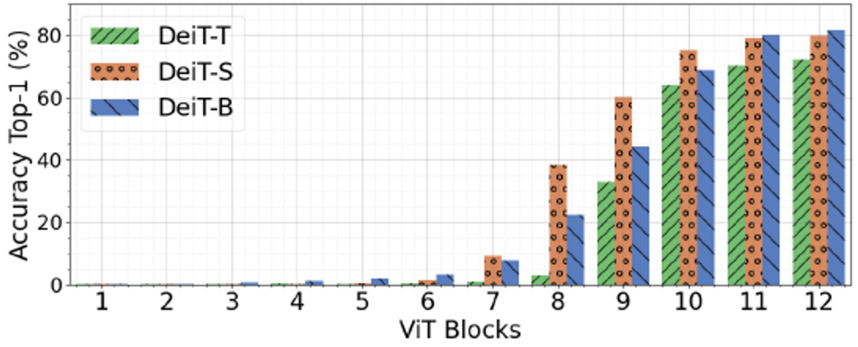
Effective Off-the-shelf Tokens for Vision Transformer

ImageNet pretrained ViT transferred to CUB

- Linear classifier on class token (or combination)
- Class tokens generated by the deeper blocks are more discriminative for classification
- **Can we design an effective ensemble of blocks?**
- Class token vs patch-token
    - Comparable performance, compute overhead

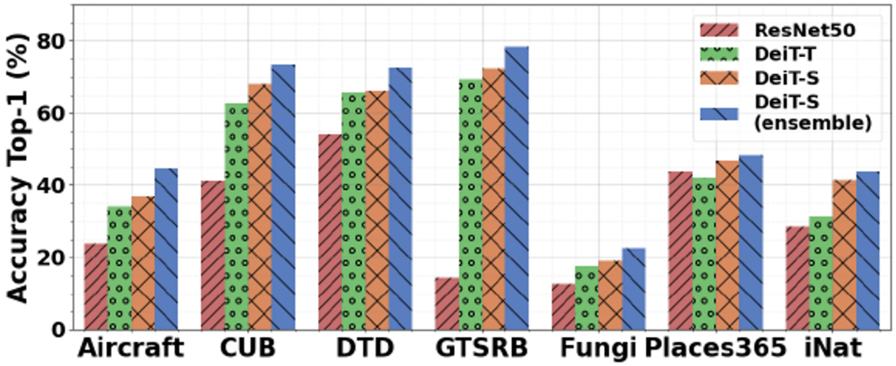
Blocks	Class Token	Patch Tokens	Top-1 (%)
Only 12^{th} (last block)	✓	✗	68.16
	✓	✓	70.66
From 1^{st} to 12^{th}	✓	✗	72.90
	✓	✓	73.16
From 9^{th} to 12^{th}	✓	✗	**73.58**
	✓	✓	73.37



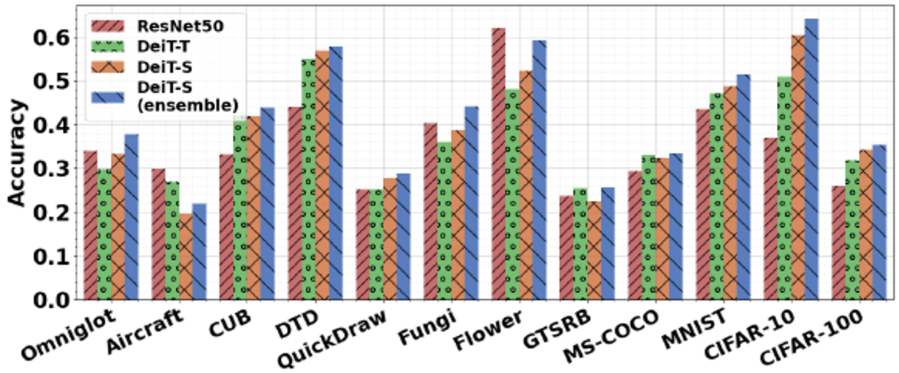


Off-the-shelf Features - CNNs vs ViTs

- **Visual Classification:** Diverse datasets for fine-grained recognition, texture classification, traffic sign recognition, specie classification and scene recognition. Classes ranging from 43 to 1394
  - ViTs consistently perform better than CNNs
- **Few-Shot Learning:** Meta-Dataset: dataset of datasets (made up of 10 datasets).
  - Transfer better across domains e.g., QuickDraw



Visual Classification



Few Shot Learning

Conclusions

ViTs show better robustness against

- Occlusions - Information Loss
- Permutations - Broken Spatial Structure
- Adversarial+Natural Perturbations

ViTs have highly dynamic and flexible receptive field

ViTs can incorporate complimentary info. e.g., texture + shape

ViTs can exhibit shape bias, comparable to humans

ViTs features generalize well across different domains/distributions

Thanks!!!