Intriguing Properties of Vision Transformers

Muzammal Naseer Kanchana Ranasinghe Salman Khan Munawar Hayat
Fahad Shahbaz Khan Ming-Hsuan Yang

Kishore Alajangi James Dwyer Michael Fielder Joseph Rivera

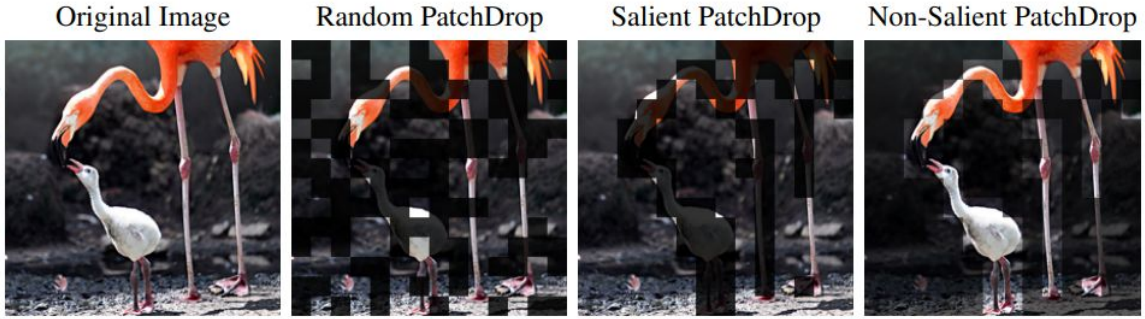
Introduction

- Compare Vision Transformers to CNN
  - ViT
  - DeiT
  - T2T
- Experiments
  - Image occlusions - robust
  - Modeling shape - performance comparable to humans
  - Positional Encoding and Preservation of Global Image Context - positional encoding not as important
  - Adversarial and Natural Perturbations - robust in most cases
  - Effective Off-the-Shelf Tokens - generalizable for downstream tasks

Are Vision Transformers Robust to Occlusions?

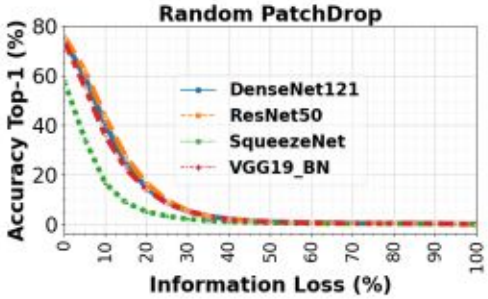
Are Vision Transformers Robust to Occlusions?

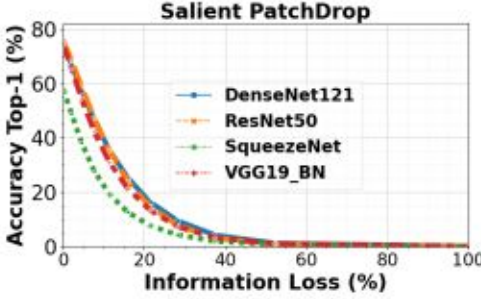
- Occlusion Modeling - PatchDrop
  - Random PatchDrop - randomly selected subset of patches is dropped
  - Salient PatchDrop
    - DINO used to select foreground/important patches
    - Patches above a threshold are dropped
  - Non-Salient PatchDrop
    - Patches below a threshold are dropped

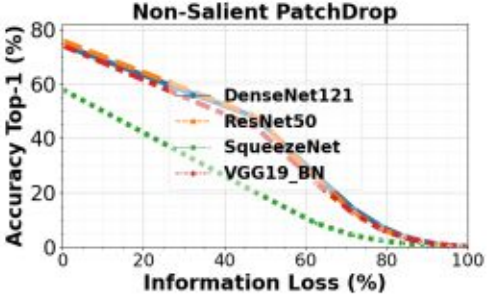


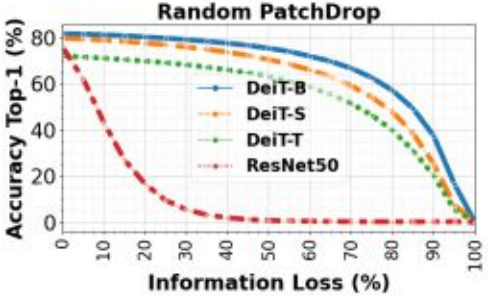
Are Vision Transformers Robust to Occlusions?

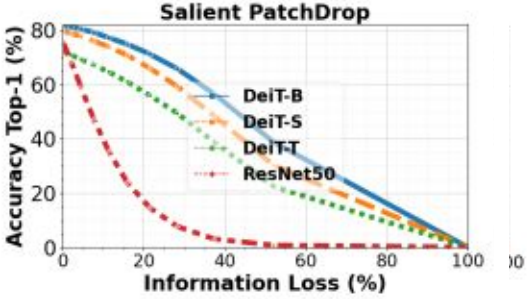
- All models pretrained on ImageNet
- Use ImageNet validation set (50k images)
- Random PatchDrop - mean accuracy of 5 runs
- Non-salient/salient - accuracy of 1 run

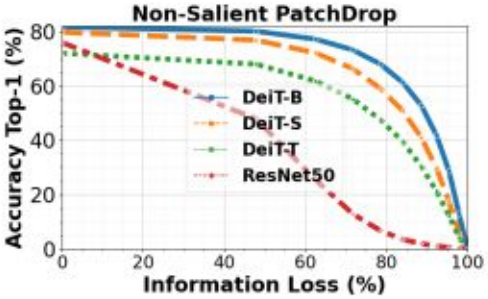

Random PatchDrop
Accuracy Top-1 (%)
Information Loss (%)
DenseNet121
ResNet50
SqueezeNet
VGG19_BN


Salient PatchDrop
Accuracy Top-1 (%)
Information Loss (%)
DenseNet121
ResNet50
SqueezeNet
VGG19_BN

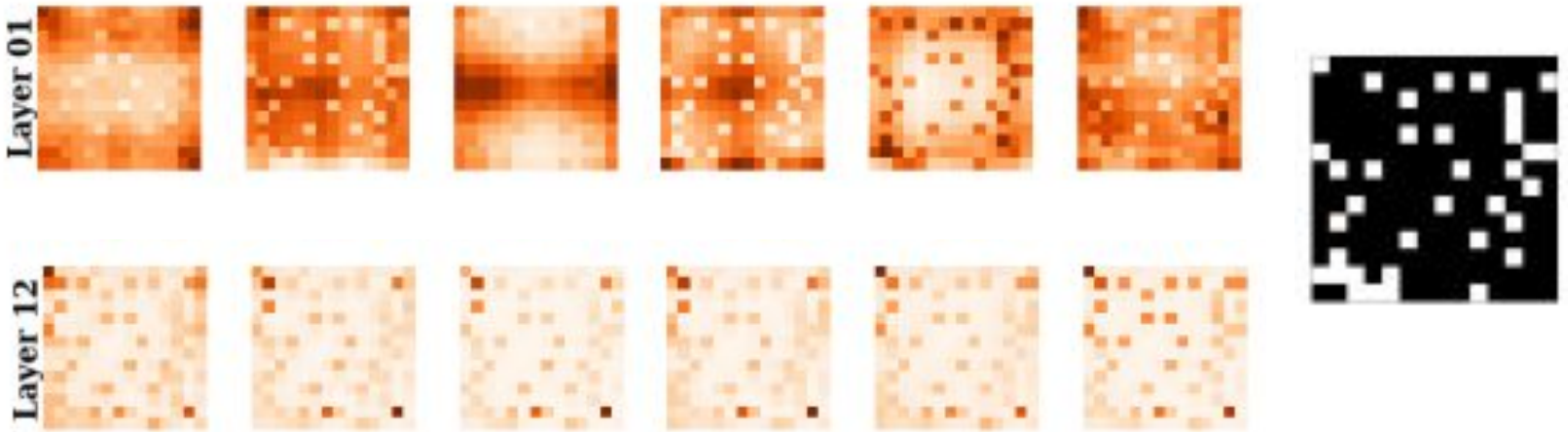

Non-Salient PatchDrop
Accuracy Top-1 (%)
Information Loss (%)
DenseNet121
ResNet50
SqueezeNet
VGG19_BN


Random PatchDrop
Accuracy Top-1 (%)
Information Loss (%)
DeiT-B
DeiT-S
DeiT-T
ResNet50


Salient PatchDrop
Accuracy Top-1 (%)
Information Loss (%)
DeiT-B
DeiT-S
DeiT-T
ResNet50


Non-Salient PatchDrop
Accuracy Top-1 (%)
Information Loss (%)
DeiT-B
DeiT-S
DeiT-T
ResNet50

Are Vision Transformers Robust to Occlusions?

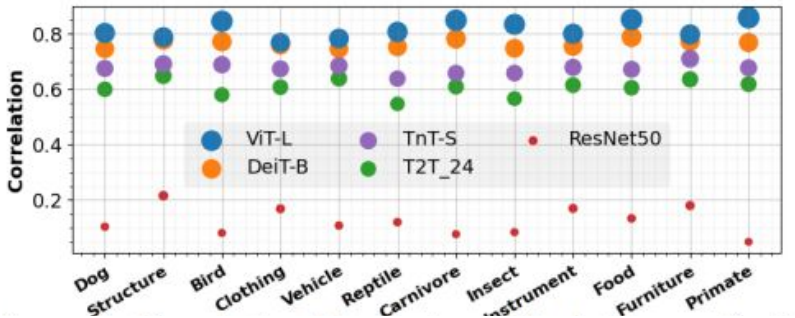


Are Vision Transformers Robust to Occlusions?

$$corr(u, v) = \frac{\sum_i \hat{u}_i \hat{v}_i}{n}$$

$$\hat{u}_i = \frac{u_i - E[u_i]}{\sigma(u_i)}$$

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



Shape vs Texture: Can Transformer Model Both Characteristics?

Shape vs Texture

- ImageNet converted to Stylized ImageNet (SIN)
- Data augmentations not used



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



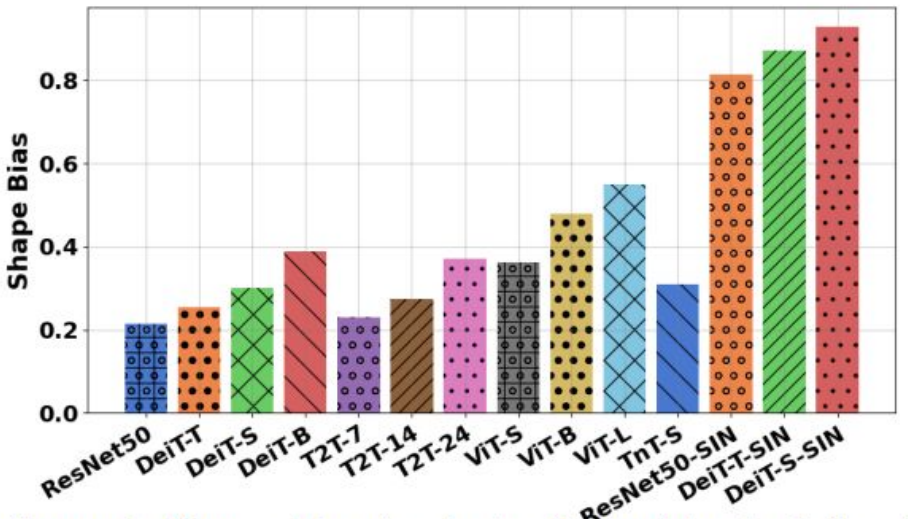
(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

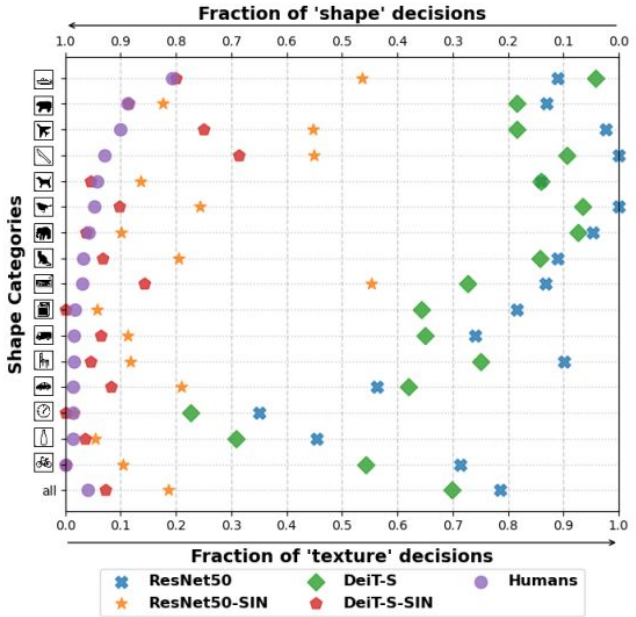
Shape vs Texture





Shape vs Texture

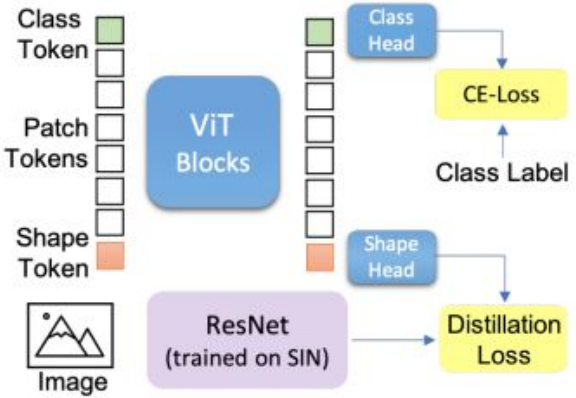




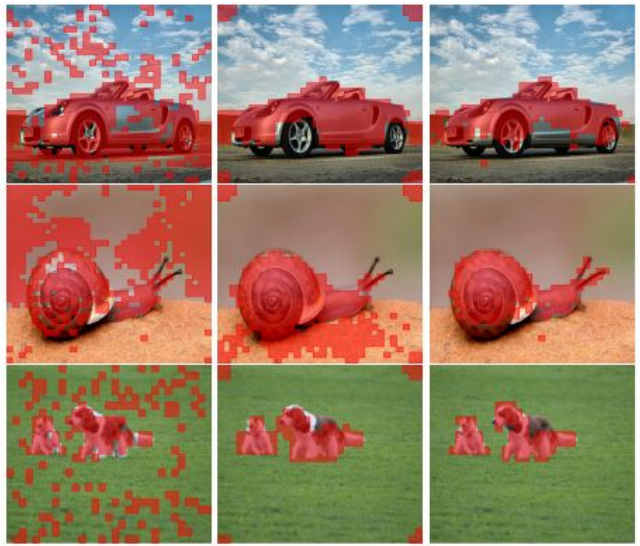
Shape vs Texture

- Transformers biased towards shape
- Shape Distillation
  - ResNet50-SIN → teacher

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



Shape vs Texture

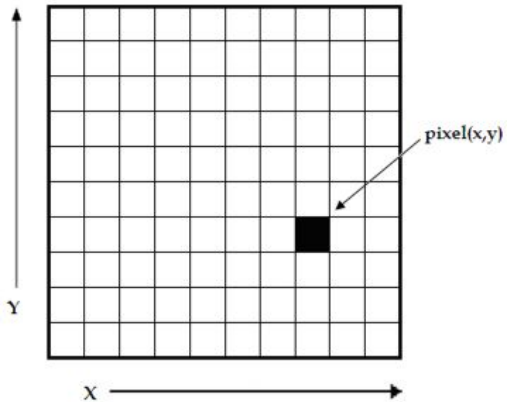


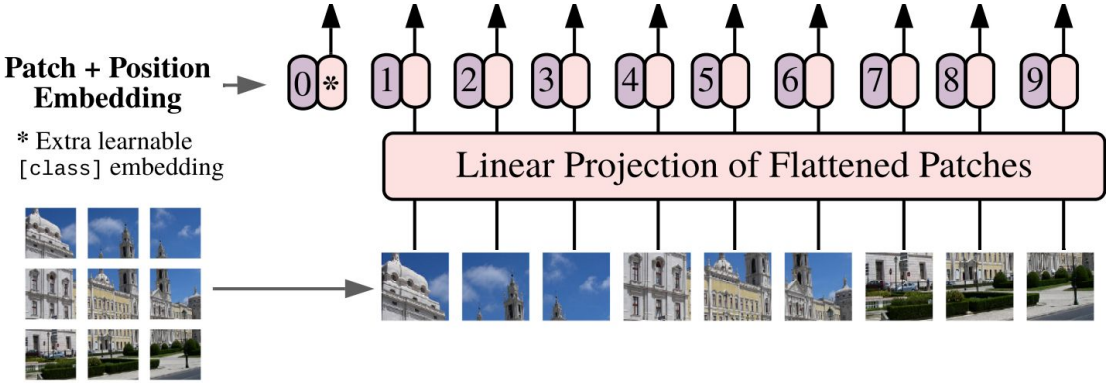
DeiT-S DeiT-S-SIN DeiT-S-SIN (Distilled)

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

Does Positional Encoding Preserve the Global Image Context?

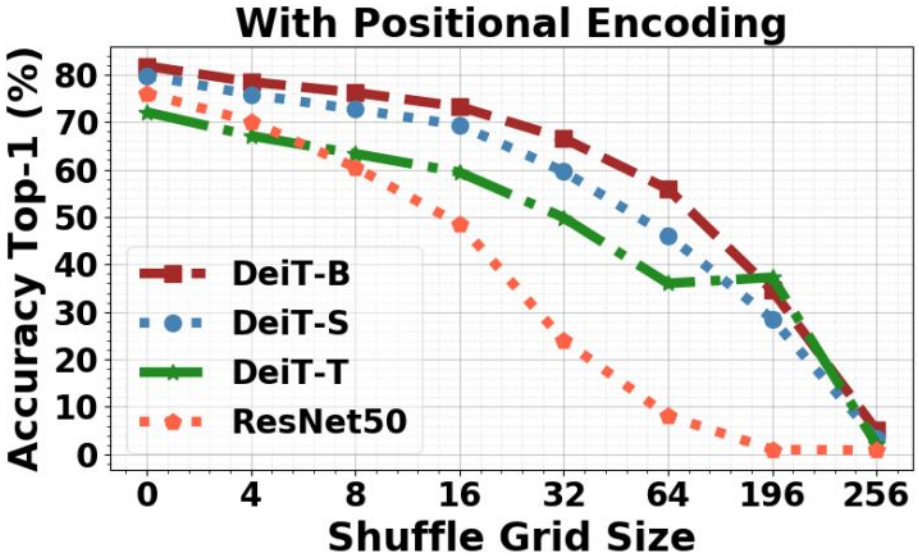
Background

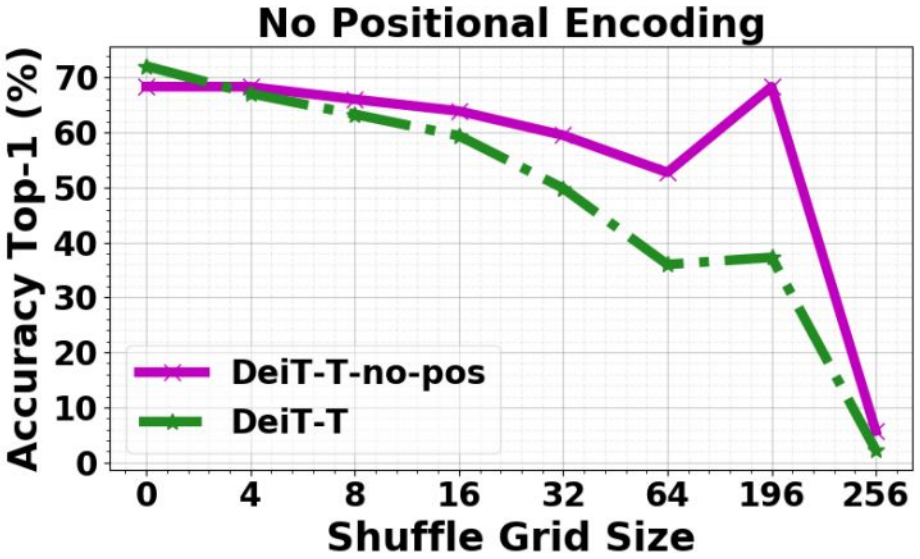

pixel(x,y)
Y
X

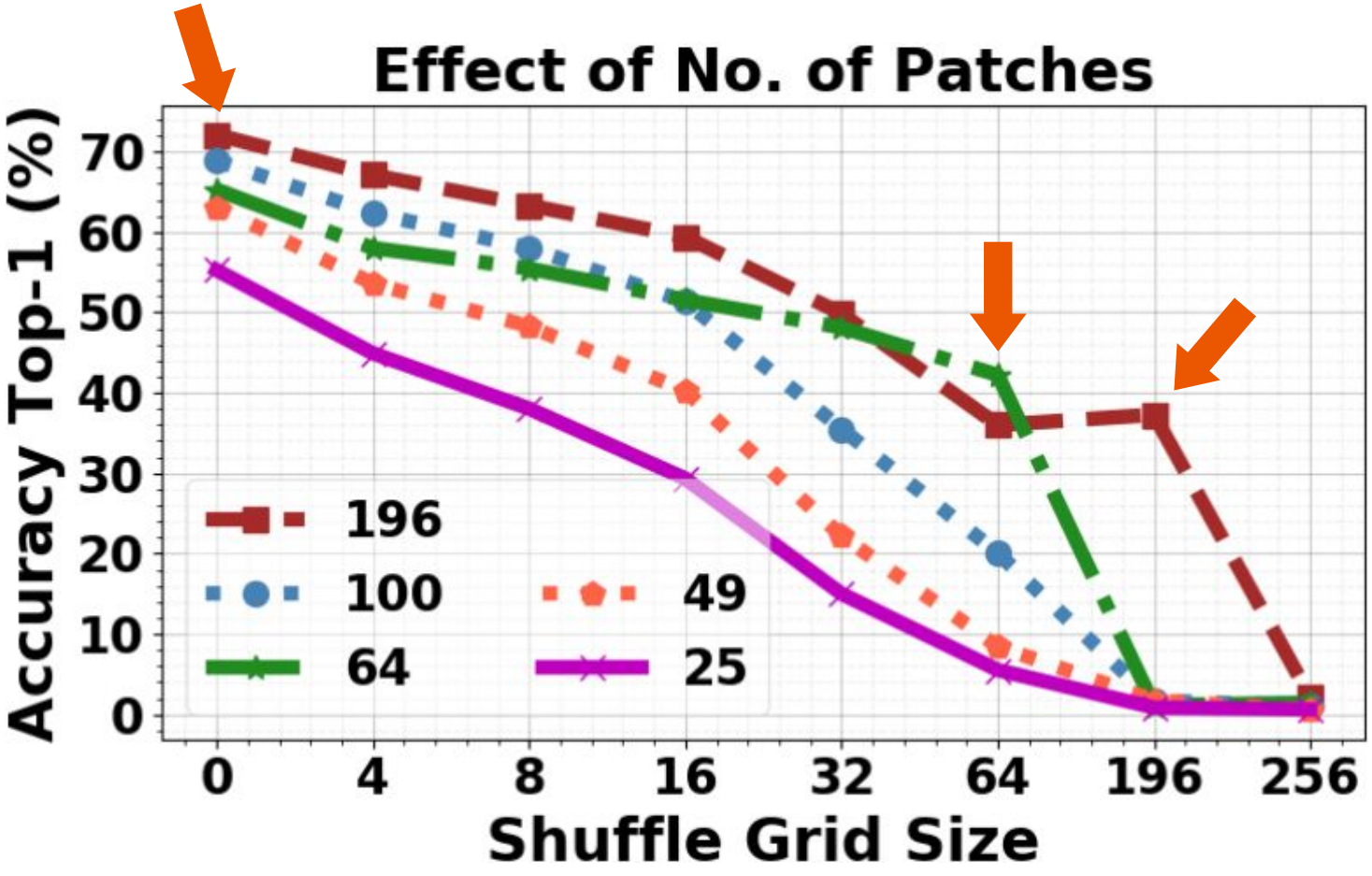

Patch + Position Embedding
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches



Original | 2 x 2 Grid | 4 x 4 Grid | 8 x 8 Grid | 14 x 14 Grid


With Positional Encoding
Accuracy Top-1 (%)
80
70
60
50
40
30
20
10
0
DeiT-B
DeiT-S
DeiT-T
ResNet50
0
4
8
16
32
64
196
256
Shuffle Grid Size


No Positional Encoding
Accuracy Top-1 (%)
70
60
50
40
30
20
10
0
DeiT-T-no-pos
DeiT-T
0
4
8
16
32
64
196
256
Shuffle Grid Size


Effect of No. of Patches
Accuracy Top-1 (%)
70
60
50
40
30
20
10
0
196
100
49
64
25
0
4
8
16
32
64
196
256
Shuffle Grid Size

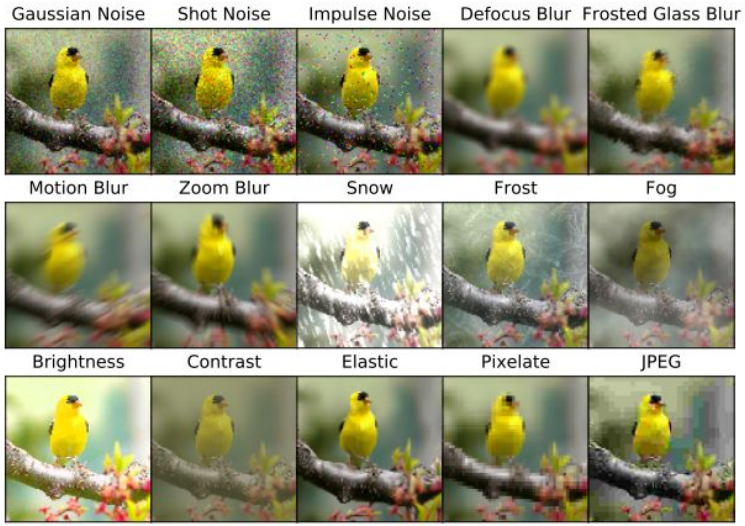
Conclusions

- ViT's show high permutation invariance to patch positions
- Pos. encodings for ViT do not preserve global image context
- Dependence on positional encodings to perform well under occlusions is also incorrect
- VITs robustness to occlusion a result of dynamic receptive field & patch size

Robustness of ViT to Adversarial and Natural Perturbations

Robustness to Natural Perturbations

- Are ViTs robust to common corruptions, like fog, rain, snow, and noise?
- ImageNet-C (Corruptions) is used as a baseline
- 15 common augmentation methods (seen right)
  - Noise, Blur, compression, etc.
- 5 levels of aggression in augmentation



ImageNet-C

Scoring Robustness to Natural Perturbations

$E^f_{s,c}$ Compute test error of a model **f** on corruption method **C** with severity **S e.g. ViT, C=Snow, S=5**

$$\mathrm{CE}^f_c = \left(\sum_{s=1}^{5} E^f_{s,c}\right) \Big/ \left(\sum_{s=1}^{5} E^{\mathrm{AlexNet}}_{s,c}\right)$$ Compute Corruption Error (**CE**) from all severities normalized

$\mathrm{CE}^f_{\text{Gaussian Noise}}, \mathrm{CE}^f_{\text{Shot Noise}}, \ldots, \mathrm{CE}^f_{\text{JPEG}}$ Average scores across all corruptions to get mean (**mCE**)

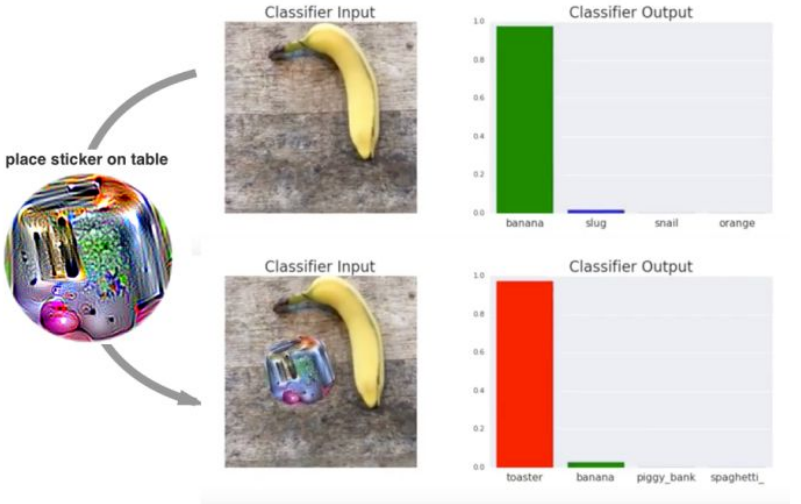
Robustness to Natural Perturbations

- **mCE** is a relative metric (lower is better)
- Most ViTs with data augmentations performed better than ResNet50 (Augmix)
- ViT-SIN models are more susceptible to natural distribution shifts, even compared to ResNet50
    - This supports a shape-bias vs robustness trade-off

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 Patch Attacks

- Adversarial Patches are taught to confidently confuse a classifier.
- Real-World Use Cases:
  - Highly dangerous to autonomous vehicles when road signs are used to understand speed limits, stop signs, etc.
  - Trick facial recognition software



Training an Adversarial Patch

Attack Patch

Input

A( ,  , location, rotation, scale,...) =

Dog: 99%

Cat: 1%

Dog Misclassified As Cat

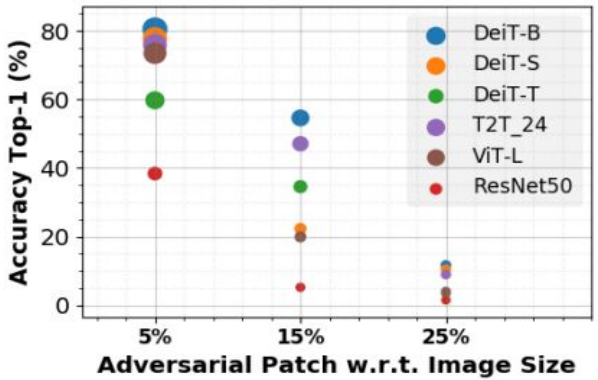


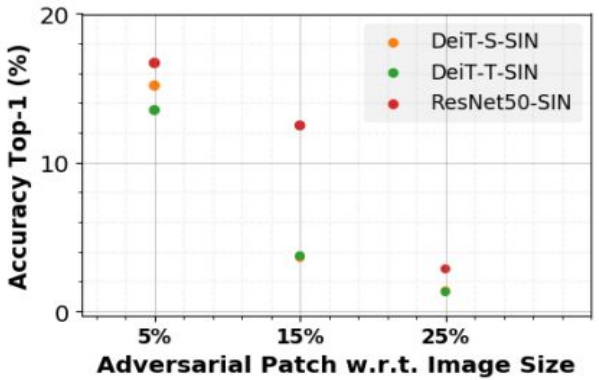
Dog: 1%

Cat: 99%

Robustness to Adversarial Attacks

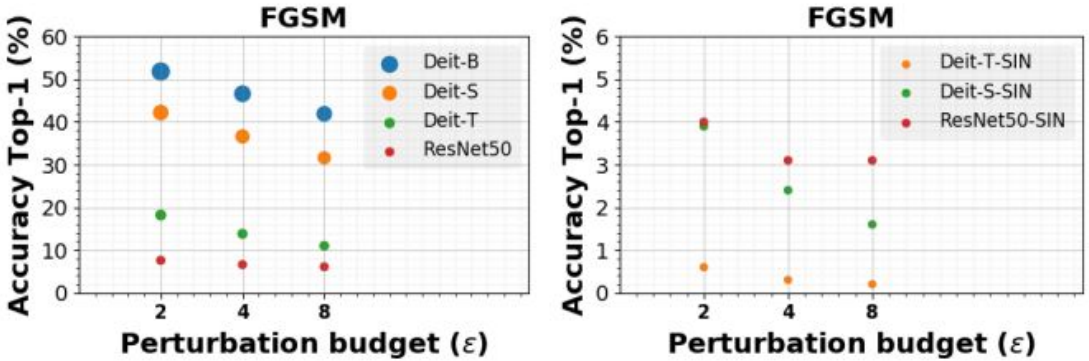
- Adversarial Patch Attack shown to be less effective on ViT models than the baseline ResNet50
  - The change in pixel values is unbounded for this attack
- ResNet models underperform when trained on ImageNet
- ViTs-SIN underperform ResNet50-SIN however, supporting shape-bias vs robustness trade-off



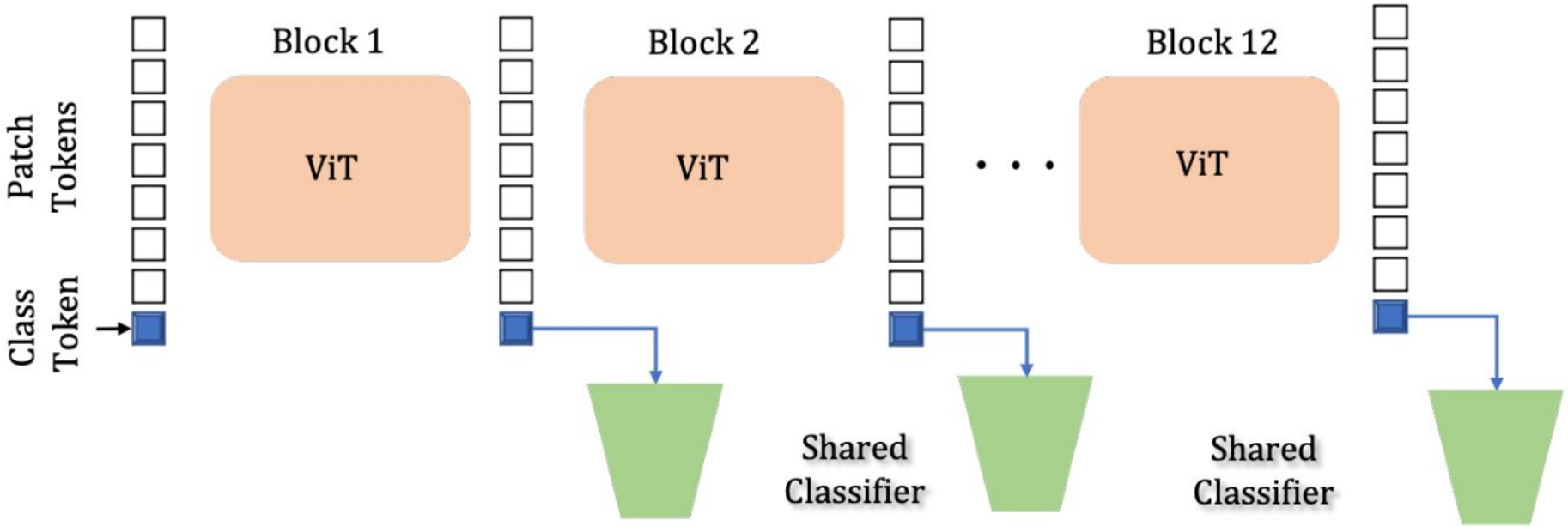


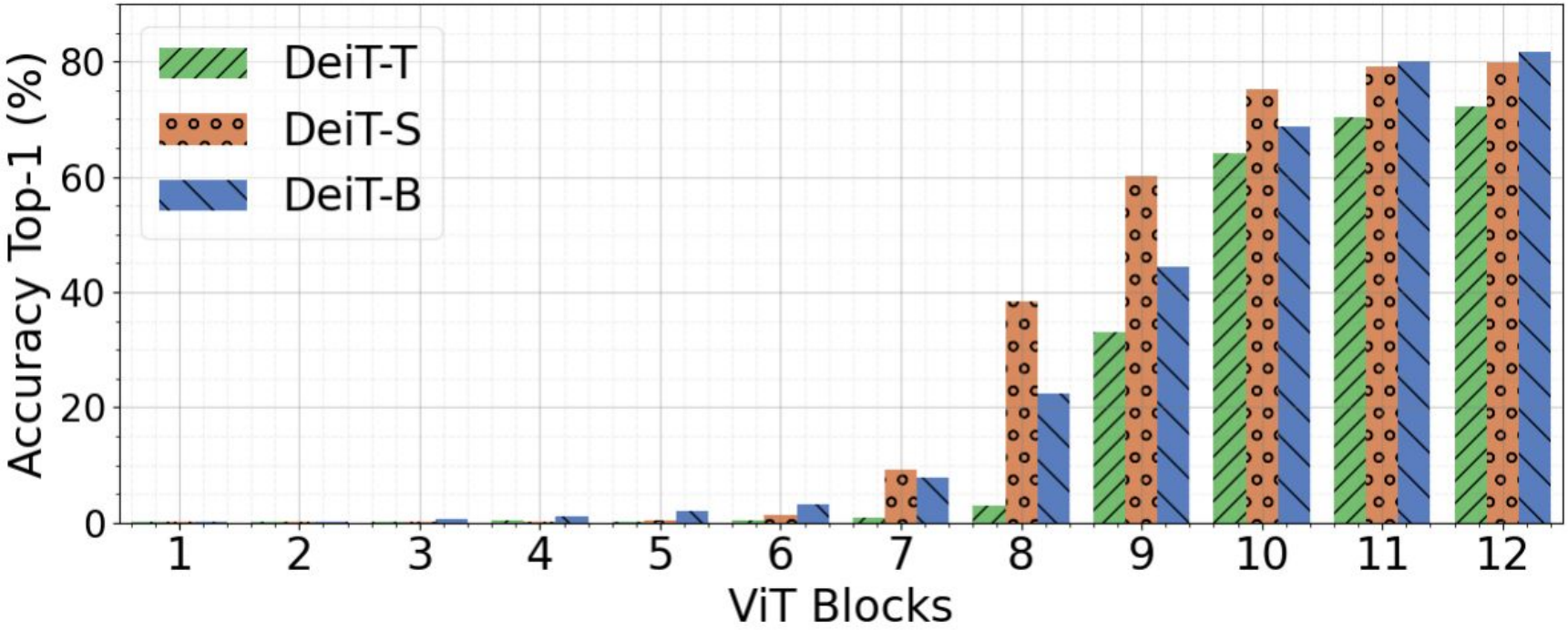
Robustness to Adversarial Attacks - FGSM

- The change in pixel values is bounded for these attacks, and white-box (access to model gradients)
- Fast Gradient Sign Method (FGSM) uses the model to generate an image that *maximizes* loss with respect to the true class
- Epsilon is used to budget how much each pixel can be changed



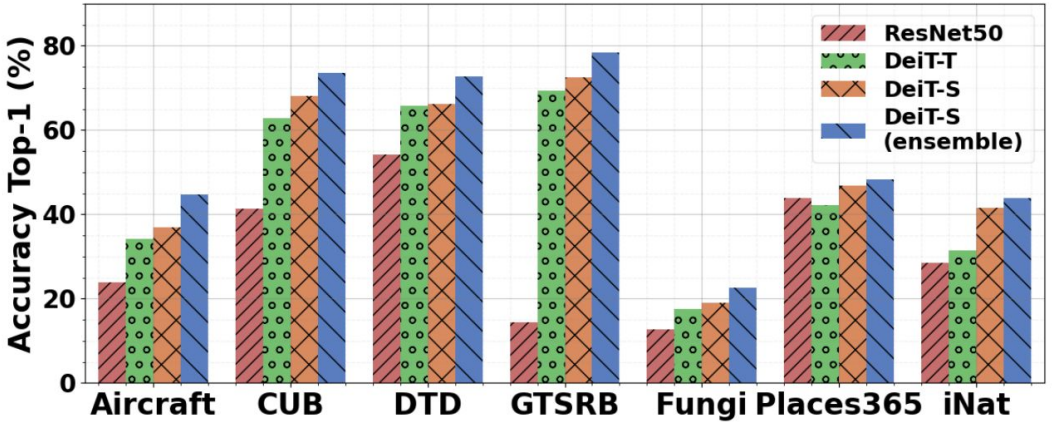
Effective Off-the-shelf Tokens for Vision Transformer

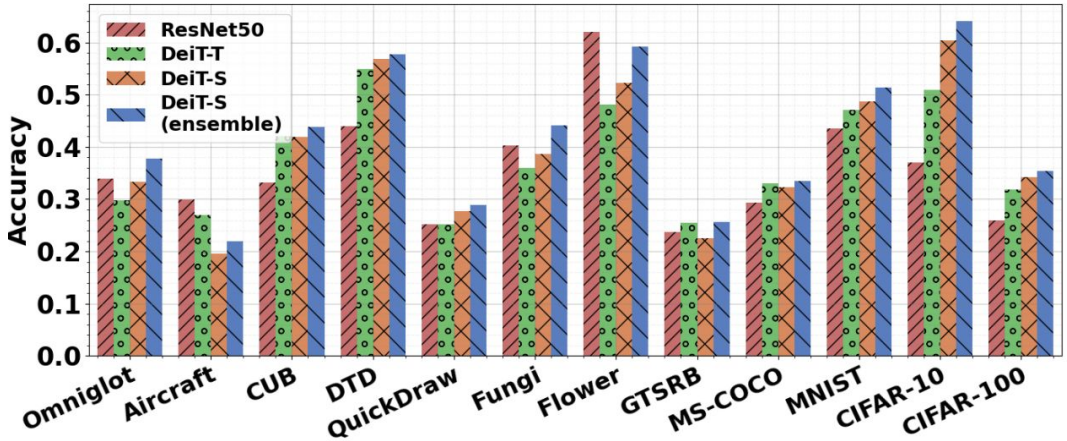

Patch Tokens
Class Token
Block 1
ViT
Block 2
ViT
Block 12
ViT
Shared Classifier
Shared Classifier


DeiT-T
DeiT-S
DeiT-B
Accuracy Top-1 (%)
0
20
40
60
80
1
2
3
4
5
6
7
8
9
10
11
12
ViT Blocks



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


Accuracy Top-1 (%)
0
20
40
60
80
ResNet50
DeiT-T
DeiT-S
DeiT-S (ensemble)
Aircraft
CUB
DTD
GTSRB
Fungi
Places365
iNat


Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
ResNet50
DeiT-T
DeiT-S
DeiT-S (ensemble)
Omniglot
Aircraft
CUB
DTD
QuickDraw
Fungi
Flower
GTSRB
MS-COCO
MNIST
CIFAR-10
CIFAR-100

Conclusion

- Vision Transformers perform well
  - Occlusion
  - Distributional shifts and patch permutations
  - Automatic segmentation
  - Adversarial patches and common corruptions
  - Transferability to downstream tasks

References

- M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. Shahbaz, M. Yang (2021). “Intriguing properties of vision transformers”. arXiv:2105.10497 [cs.CV].
- R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. Wichmann, W. Brendel. “Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness”. arXiv preprint arXiv:1811.12231, 2018.