Understanding The Robustness In Vision Transformers

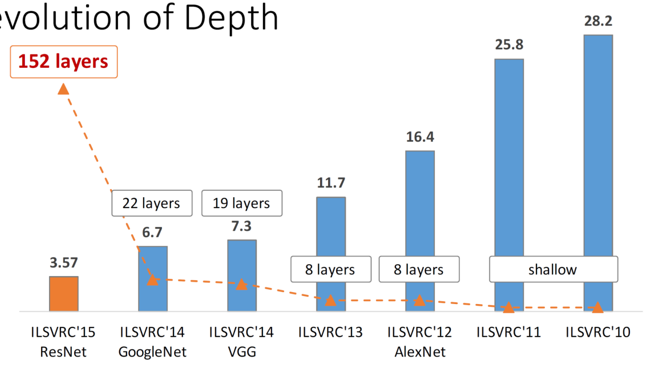
Daquan Zhou, Zhiding Yu, Enze Xie

Chaowei Xiao, Anima Anandkumar, Jiashi Feng and Jose M. Alvarez

Advances in Visual Recognition

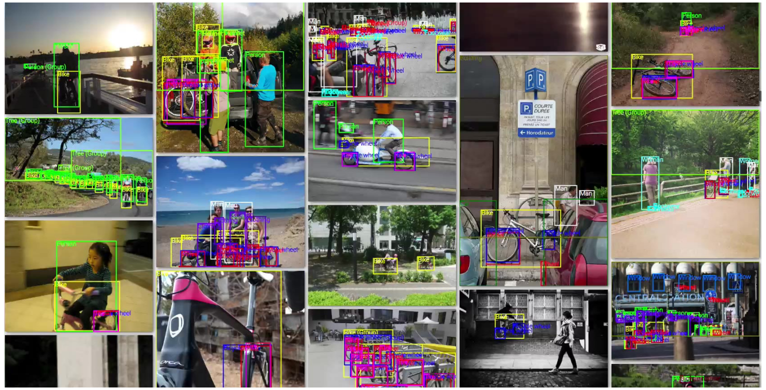
Larger Models

Revolution of Depth



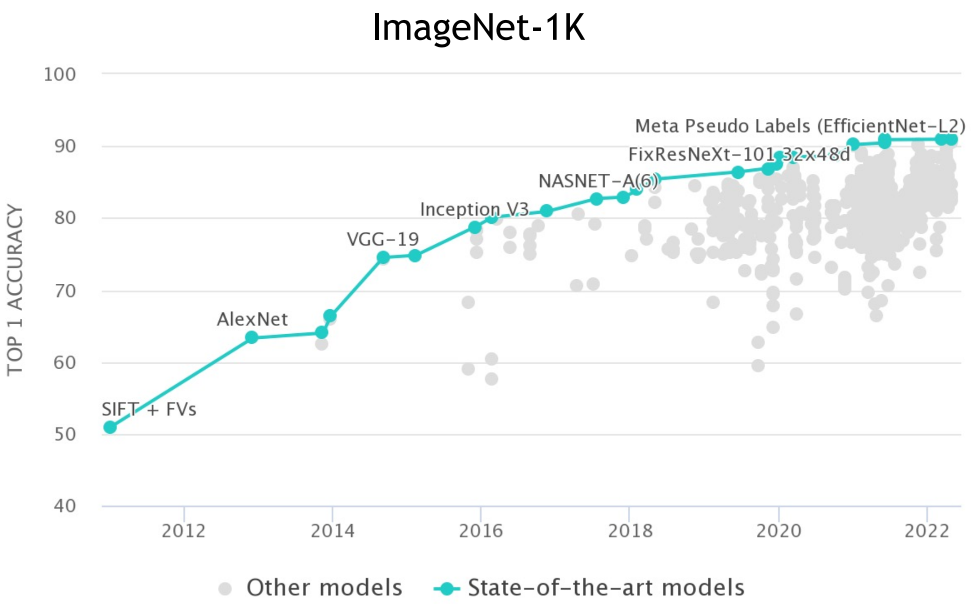
ImageNet Classification top-5 error (%)

Bigger Data



Faster Computing

Standard Visual Recognition Is Getting Saturated

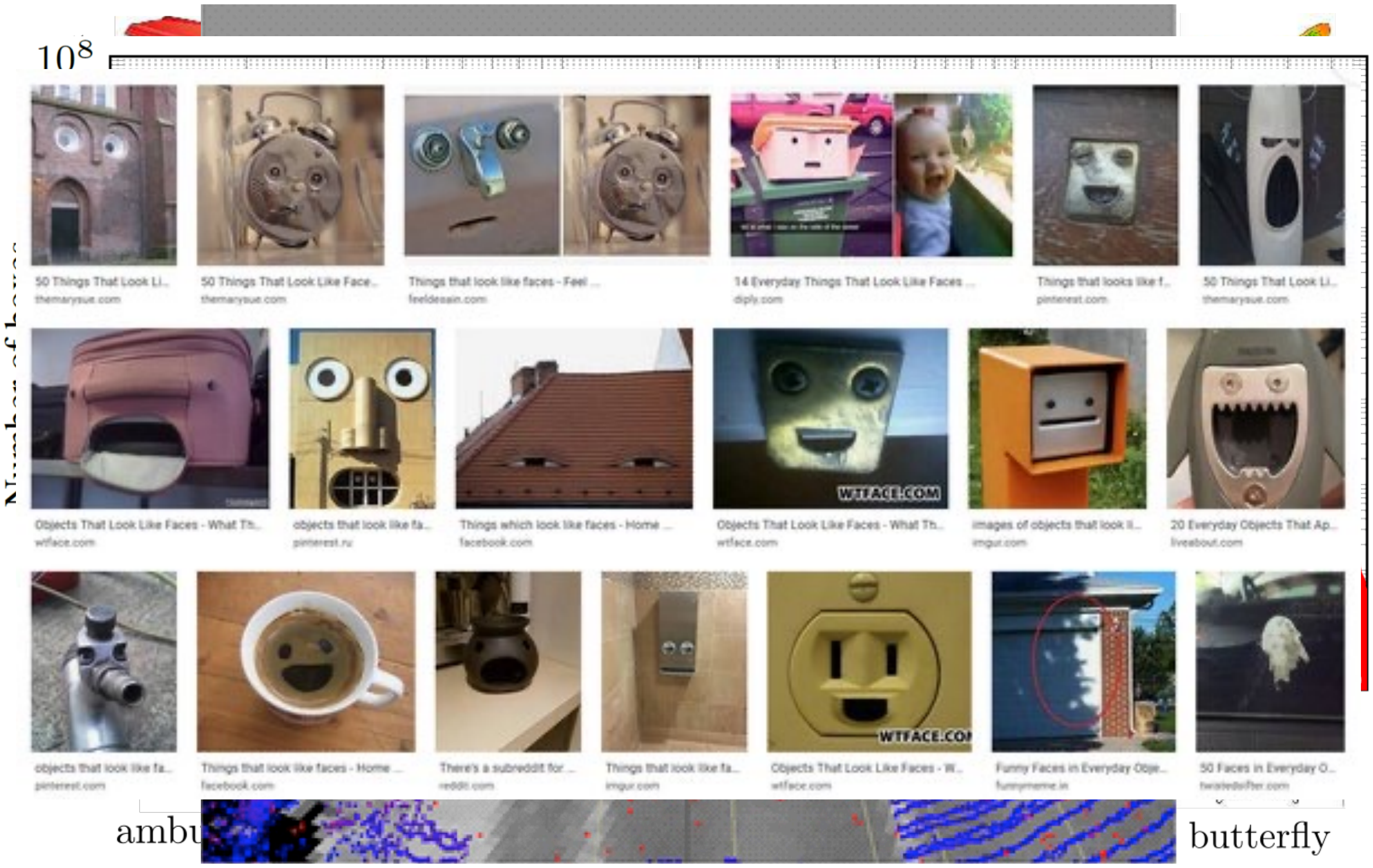


Top Performing Models

Rank	Model	Top 1 Accuracy	Top 5 Accuracy	Number of params
1	**CoCa** (finetuned)	91.00		2100M
2	**Model soups** (ViT-G/14)	90.94		1843M
3	**CoAtNet-7**	90.88%		2440M
8	**Meta Pseudo Labels** (EfficientNet-L2)	90.2%	98.8%	480M

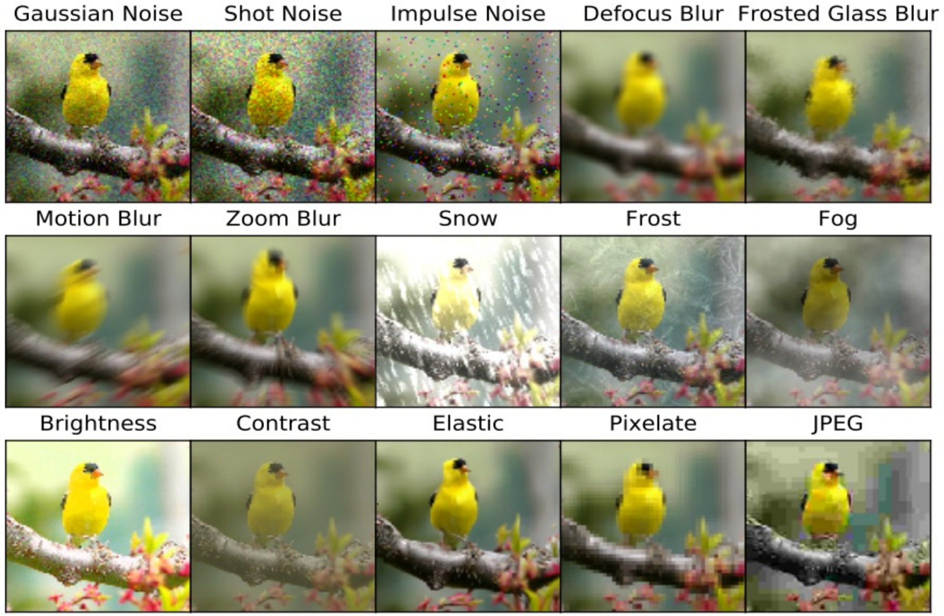
Challenge - Real World Data Are Imperfect

- **Domain shift**
- **Data noise**
- **Imbalanced distribution**
- Can contain *Occlusions*
- Can be *Cluttered*
- Can be *Ambiguous*
- Can be *Deceiving*
- …



More Challenging Scenarios

Corrupted ImageNet (ImageNet-C)

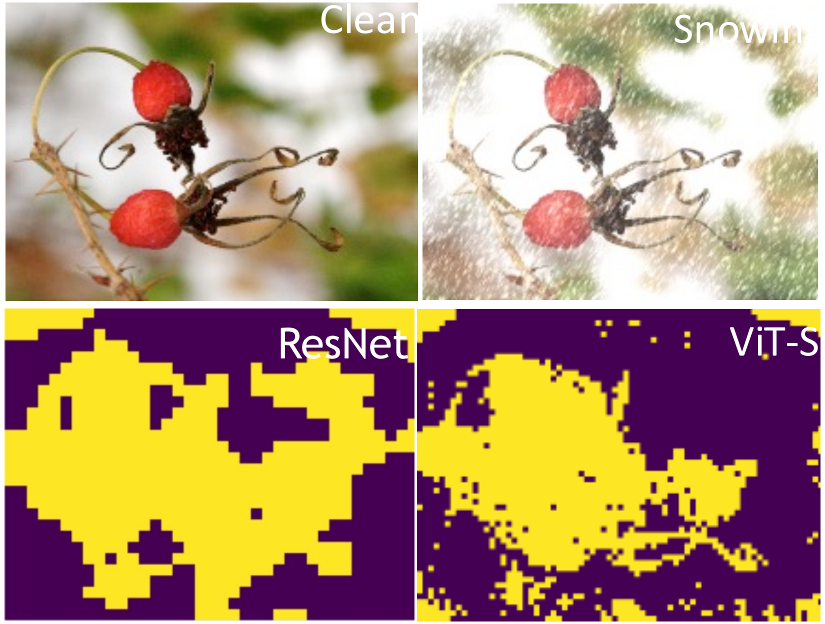


COCO-C/Cityscapes-C

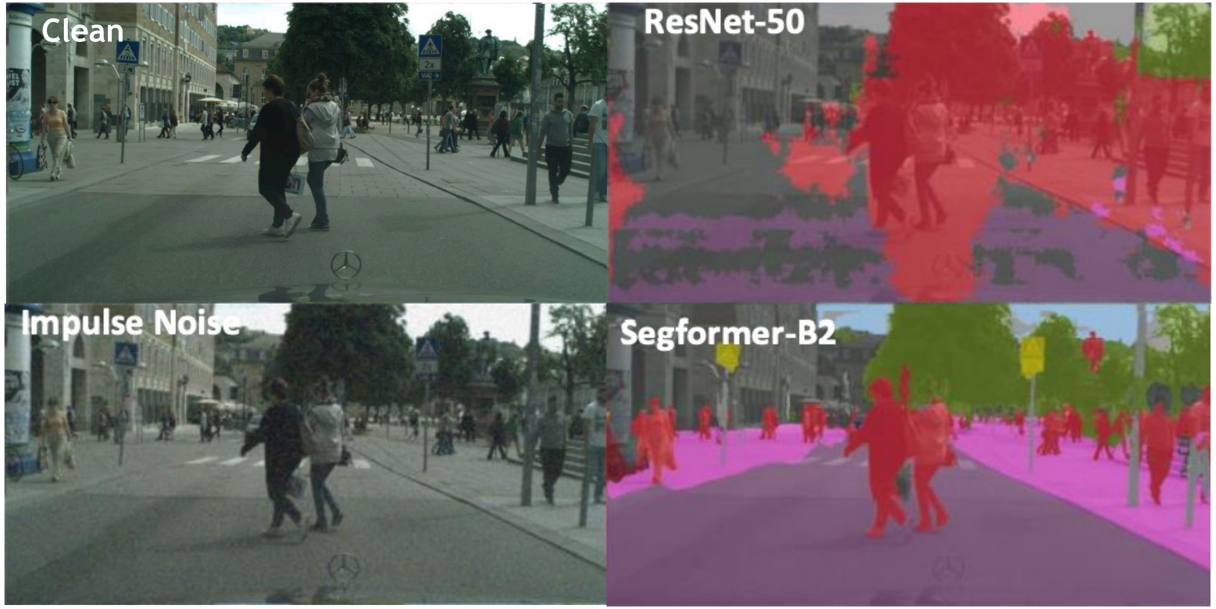
Hendrycks et al., Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, ICLR19

How Well Do Current DNNs Perform?

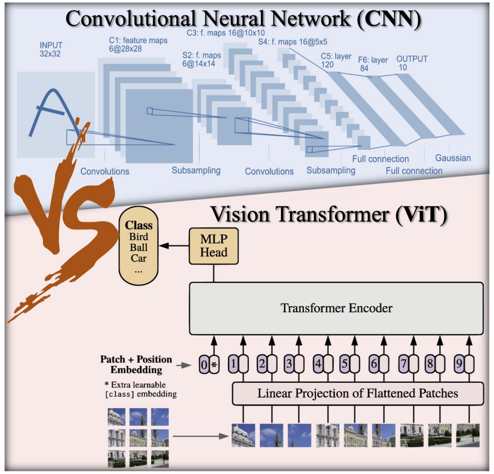
Image Classification



Semantic Segmentation



ViTs Are Robust Learners



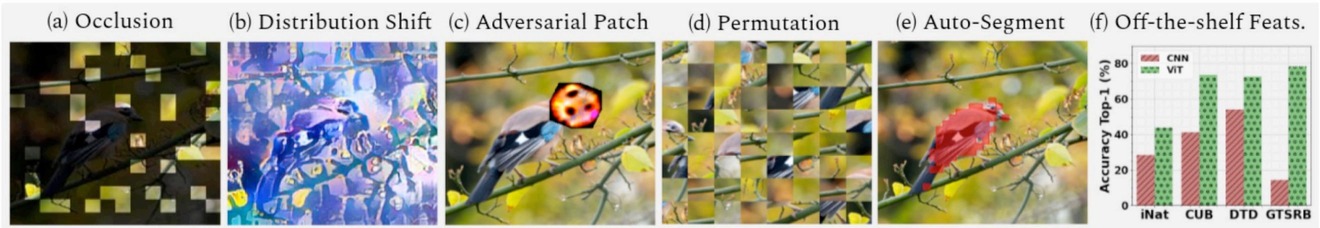
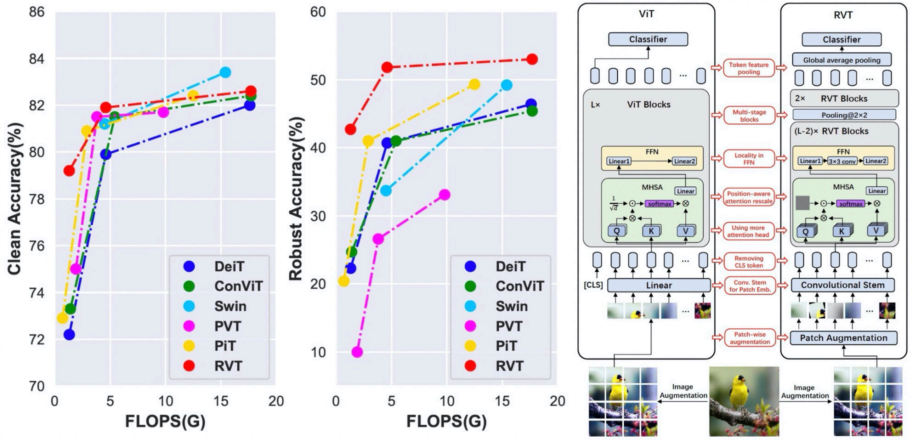


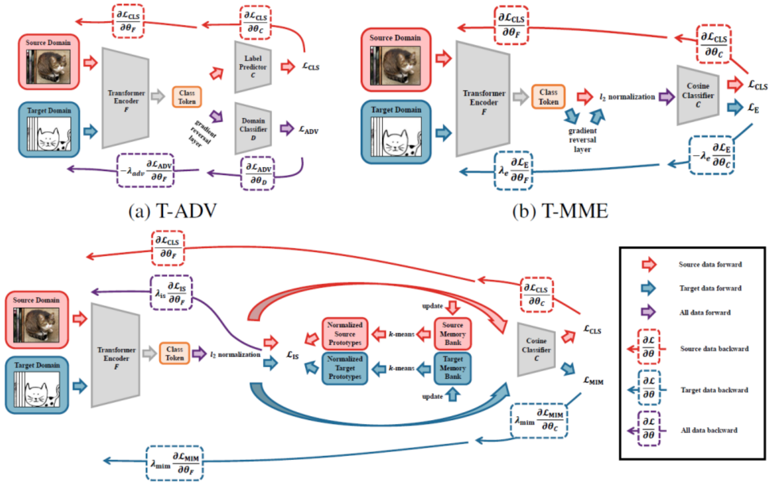
Figure 1: We show intriguing properties of ViT including impressive robustness to (a) severe occlusions, (b) distributional shifts (*e.g.*, stylization to remove texture cues), (c) adversarial perturbations, and (d) patch permutations. Furthermore, our ViT models trained to focus on shape cues can segment foregrounds without any pixel-level supervision (e). Finally, off-the-shelf features from ViT models generalize better than CNNs (f).

Naseer et al., Intriguing Properties of Vision Transformers, NeurIPS21

Bai et al., Are Transformers More Robust Than CNNs? NeurIPS21



Mao et al., RVT: Towards Robust Vision Transformer, CVPR22

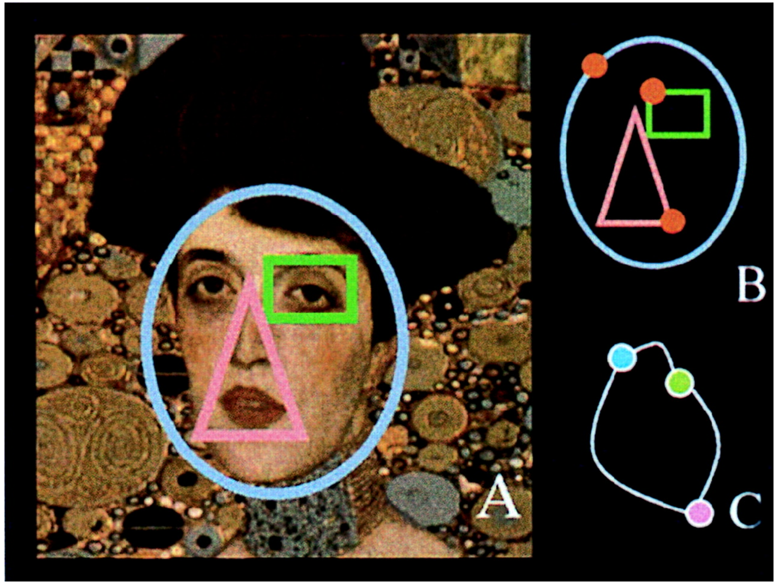


Zhang et al., Delving Deep into the Generalization of Vision Transformers under Distribution Shifts, CVPR22

Delving Deeper into ViT's Robustness

Visual Grouping and Information Bottleneck

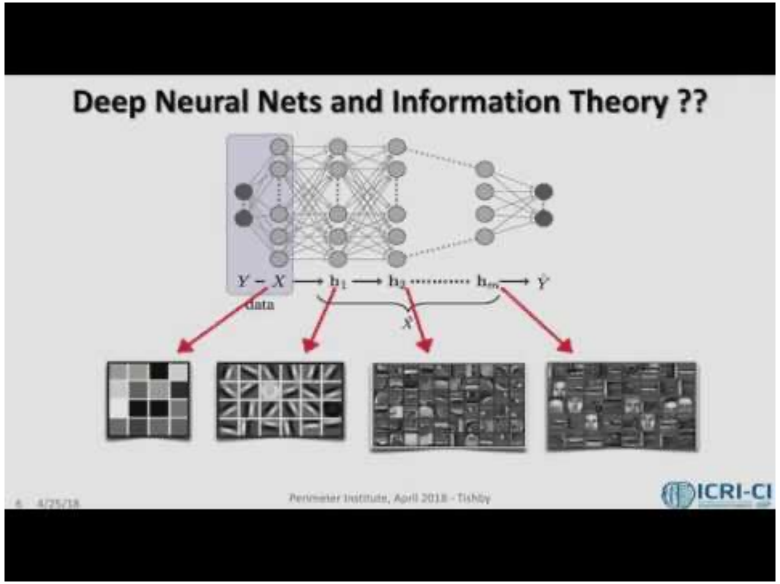
Visual Grouping



"I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees."

——Max Wertheimer

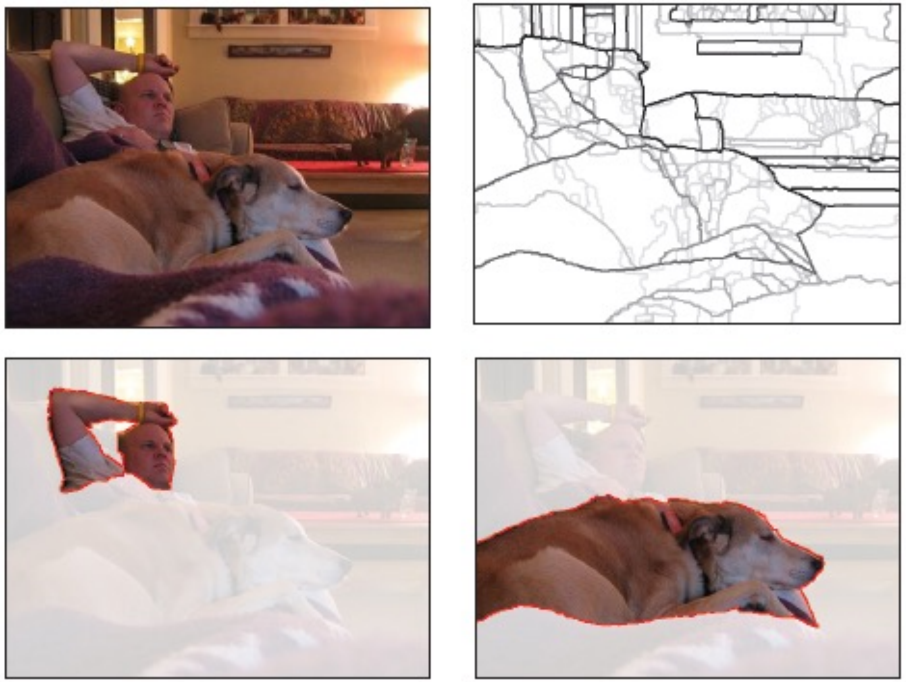
Information Bottleneck (IB)



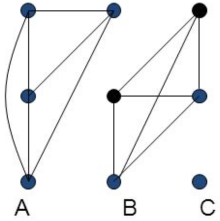
"Information bottlenecks are extremely interesting. I have to listen to it ten thousand times to really understand it. It's hard to hear such original ideas today. Maybe it's the key to the puzzle."

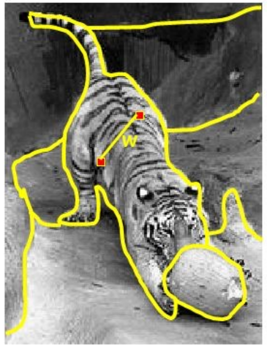
——Geoffrey Hinton

Visual Grouping



Segmentation by Graph Cuts

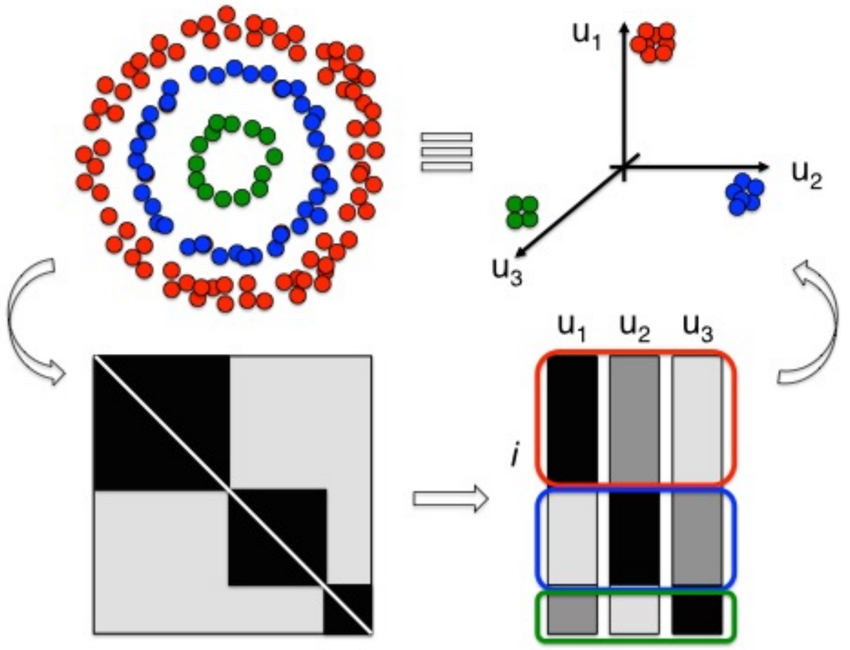




- Break Graph into Segments
  - Delete links that cross between segments
  - Easiest to break links that have low cost (low similarity)
    - similar pixels should be in the same segments
    - dissimilar pixels should be in different segments

Source: Seitz

Spectral Clustering vs. Self-Attention



$$D^{-1/2} A D^{-1/2} \boldsymbol{v} = \lambda \boldsymbol{v}$$

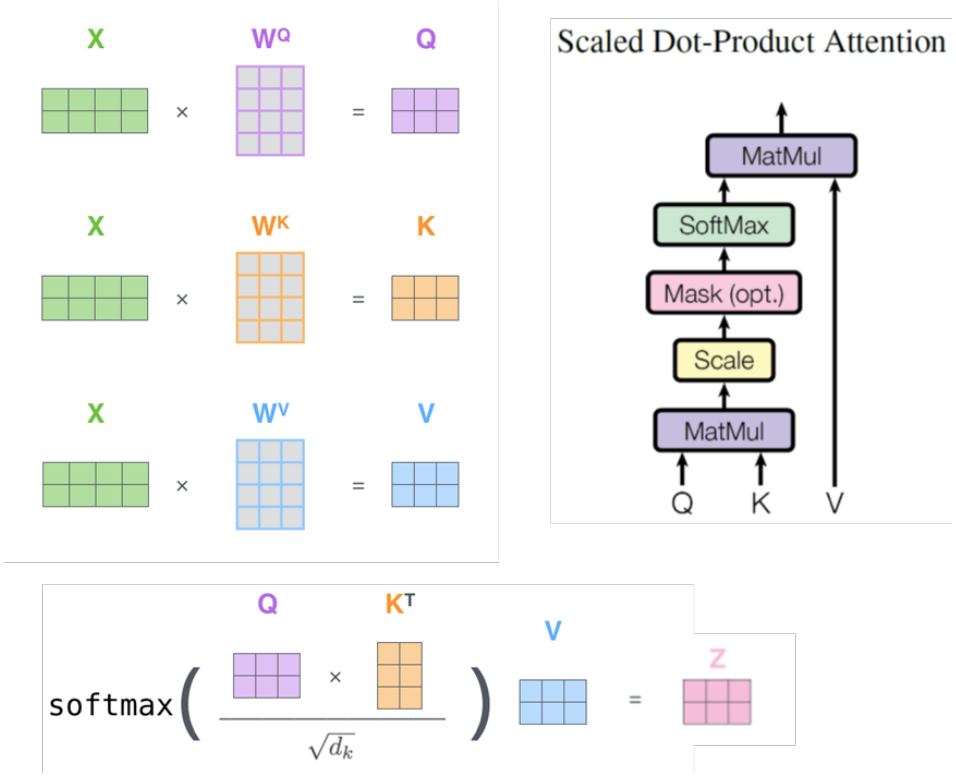
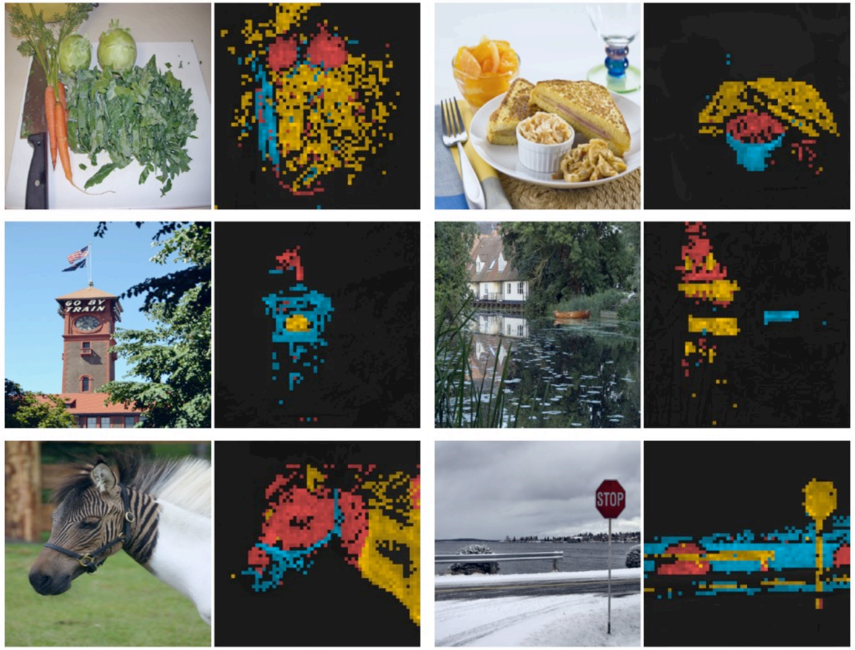
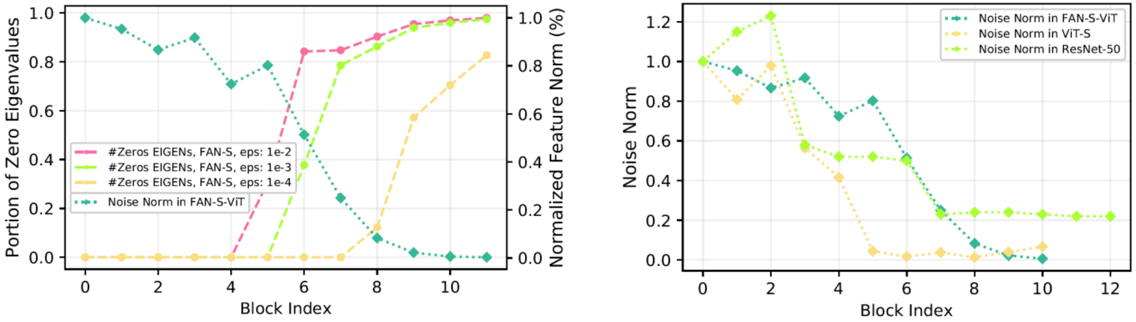


Image Credit: Spectral Clustering for Molecular Emission Segmentation.

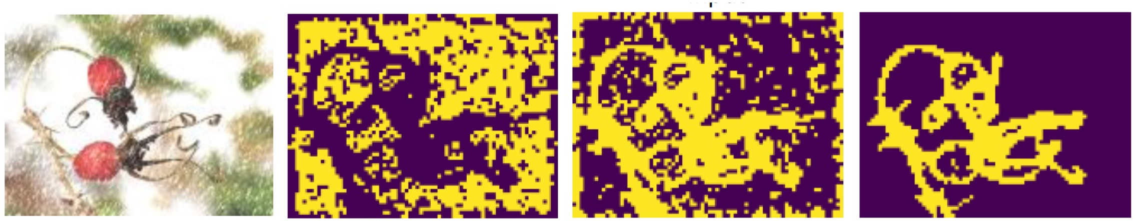
Image Credit: Jay Alammar, The Illustrated Transformer.

Emerging Properties in ViTs





Correlation between grouping and robustness over network blocks



Input BLock 6 Block8 Block9

Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV21

The Trinity among Visual Grouping, IB and Robust Generalization

Given a distribution $X \sim \mathcal{N}(X', \epsilon)$ with X being the observed noisy input and X' the target clean code, IB seeks a mapping $f(Z|X)$ such that Z contains the relevant information in X for predicting X'. This goal is formulated as the following information-theoretic optimization problem:

$$f^*_{\mathrm{IB}}(Z|X) = \arg\min_{f(Z|X)} I(X,Z) - I(Z,X'), \tag{3}$$

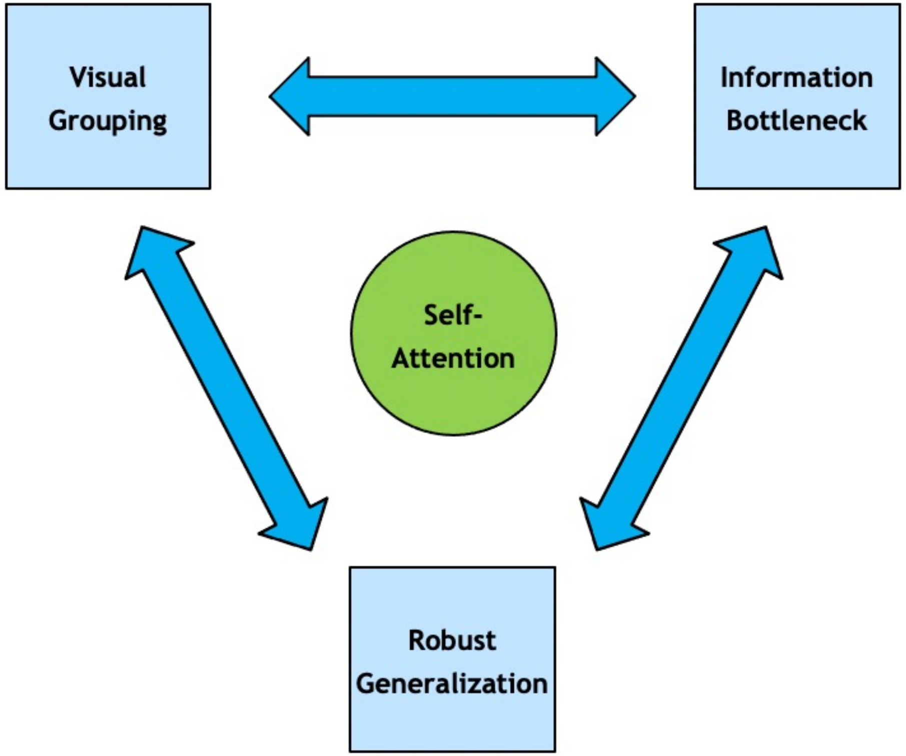
Proposition 2.1. *Under mild assumptions, the iterative step to optimize the objective in Eqn. (3) can be written as:*

$$\mathbf{z}_c = \sum_{i=1}^{n} \frac{\log[n_c/n]}{n \det \Sigma} \frac{\exp\left[\frac{\mu_c^\top \Sigma^{-1} \mathbf{x}_i}{1/2}\right]}{\sum_{c=1}^{n} \exp\left[\frac{\mu_c^\top \Sigma^{-1} \mathbf{x}_i}{1/2}\right]} \mathbf{x}_i, \tag{4}$$

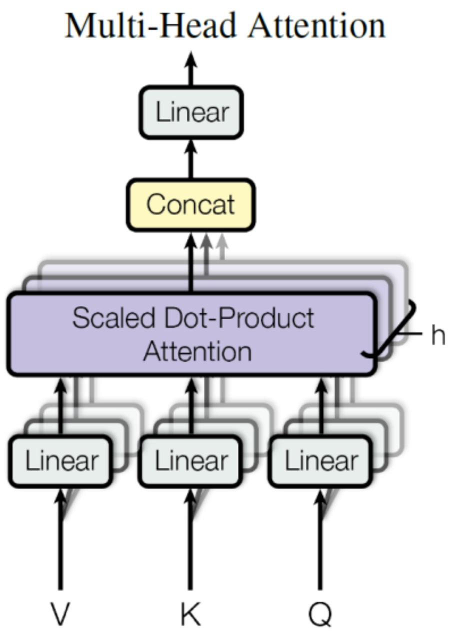
or in matrix form:

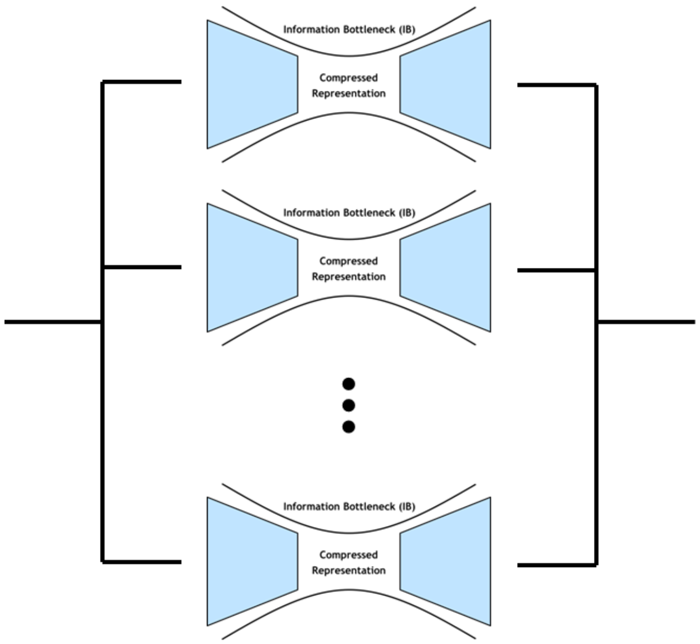
$$Z = \mathrm{Softmax}(Q^\top K/d)V^\top, \tag{5}$$

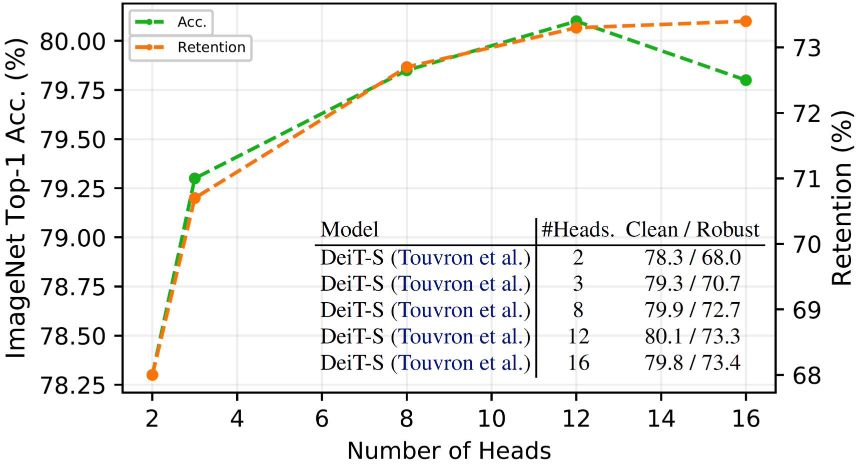
with $V = [\mathbf{x}_1, \dots, \mathbf{x}_N]\frac{\log[n_c/n]}{n \det \Sigma}$, $K = [\mu_1, \dots, \mu_N] = W_K X$, $Q = \Sigma^{-1}[\mathbf{x}_1, \dots, \mathbf{x}_N]$ *and* $d = 1/2$. *Here* n_c, Σ *and* W_K *are learnable variables.*



MSHA as Mixture of IBs

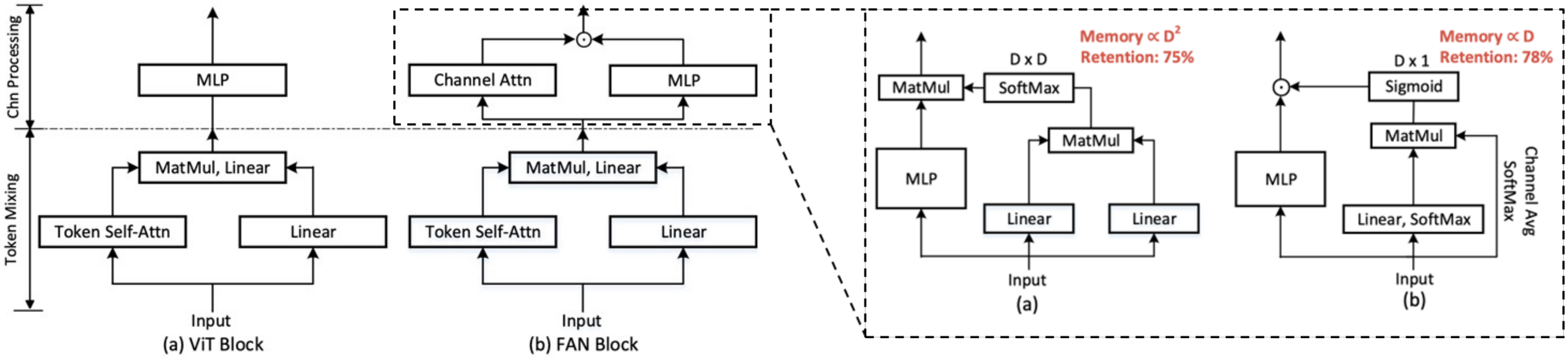




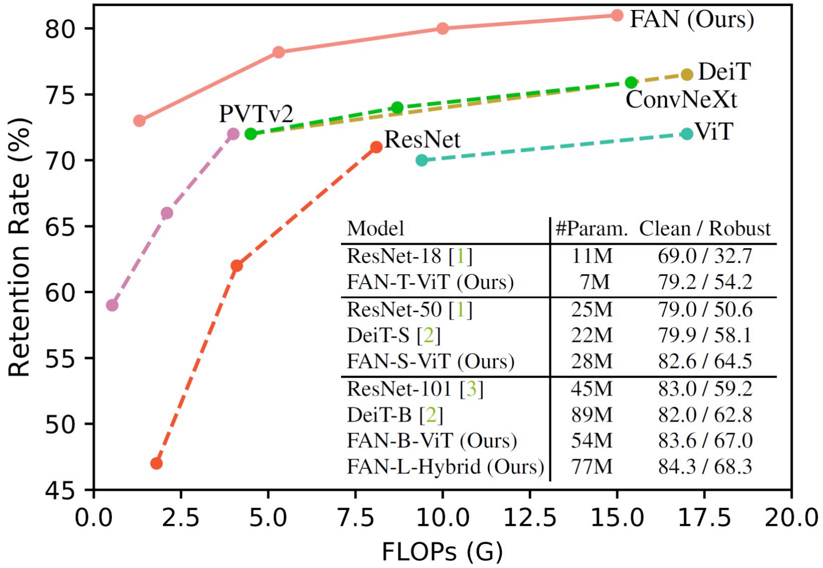


Fully Attentional Network

- Further deploy the attention mechanism reinforce the clustering phenomenon
- Fore-ground objects are better captured
- Directly apply SA along the channel dimension has two drawbacks
  1) Large computational overhead
  2) Low parameter efficiency



Main Results – Image Classification



Corrupted input | ResNet-50 | FAN-S (ours)

Model	Params (M)	Clean	IN-A	IN-R	IN-C
ImageNet-1K Pre-trained					
XCiT-S24 (El-Nouby et al.)	47.7	82.6	27.8	45.5	49.4
RVT-B* (Mao et al.)	91.8	82.6	28.5	48.7	46.8
Swin-B (Liu et al.)	87.8	83.4	35.8	64.2	54.4
ConvNeXt-B (Liu et al.)	88.6	83.8	36.7	51.3	46.8
FAN	76.8	84.3	41.8	53.2	43.0
ImageNet-22K Pre-trained					
ConvNeXt-B‡ (Liu et al.)	88.6	86.8	62.3	64.9	43.1
FAN	76.8	86.5	60.7	64.3	**35.8**
FAN‡	76.8	**87.1**	**74.5**	**71.1**	36.0

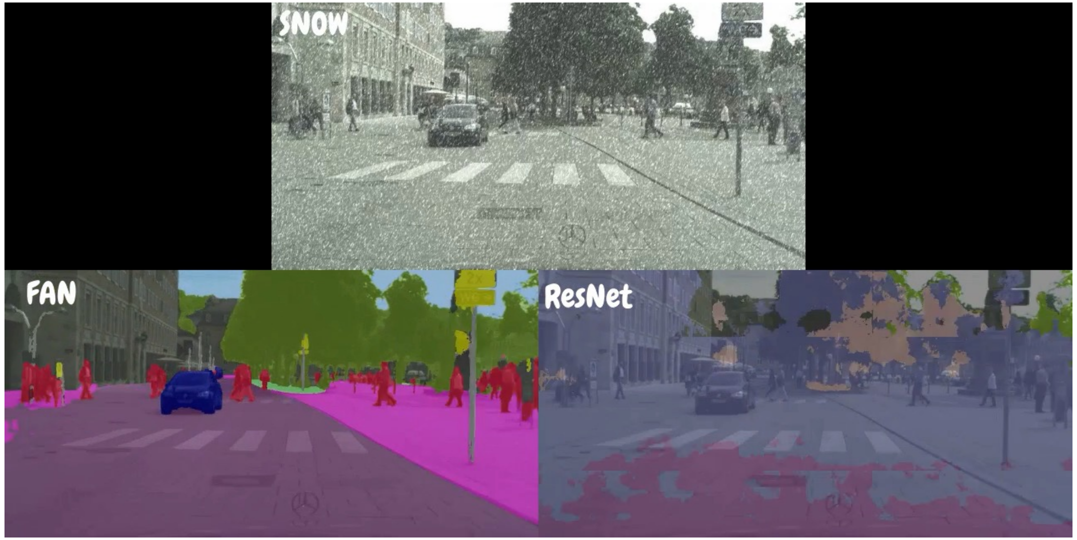
Main Results - Downstream Tasks

(a) **Main results on semantic segmentation.** 'R-' and 'X-' refer to DeepLabv3+, ResNet and Xception. The mIoUs of DeepLabv3+ framework are reported from [31]. FAN shows significantly stronger clean accuracy and robustness than other models.

Model	Encoder Size	City	City-C	Retention
DeepLabv3+ (R50)	25.4M	76.6	36.8	48.0%
DeepLabv3+ (R101)	47.9M	77.1	39.4	51.1%
ICNet [32]	-	65.9	28.0	42.5%
FCN-8s [33]	50.1M	66.7	27.4	41.1%
ResNet-38 [34]	-	77.5	32.6	42.1%
ConvNeXt-T [14]	29.0M	79.0	54.4	68.9%
SETR [35]	22.1M	76.0	55.3	72.8%
Swin-T [24]	28.4M	78.1	47.3	60.6%
SegFormer-B0 [10]	3.4M	76.2	48.8	64.0%
SegFormer-B1 [10]	13.1M	78.4	52.7	67.2%
SegFormer-B2 [10]	24.2M	81.0	59.6	73.6%
SegFormer-B5 [10]	81.4M	82.4	65.8	79.9%
FAN-T-Hybrid (Ours)	7.4M	81.2	57.1	70.3%
FAN-S-Hybrid (Ours)	26.3M	81.5	66.4	81.5%
FAN-B-Hybrid (Ours)	50.4M	82.2	66.9	81.5%
FAN-L-Hybrid (Ours)	76.8M	82.3	68.7	**83.5%**

(b) **Main results on object detection.** FAN shows stronger clean accuracy and robustness than other models. '†' denotes the accuracy pre-trained on ImageNet-22K.

Model	Encoder Size	COCO	COCO-C	Retention
Mask R-CNN				
ResNet-50 [1]	25.4M	39.9	21.3	53.3%
DeiT-S [2]	22.1M	40.0	26.9	67.3%
Swin-T [24]	28.0M	46.0	29.3	63.7%
ConvNeXt-T [24]		46.2		
FAN-T-Hybrid	7.0M	45.8	29.7	64.8%
FAN-S-Hybrid	26.3M	49.1	35.5	72.3%
Cascade R-CNN				
Swin-T		50.4		
ConvNeXt-T		50.4		
FAN-S-Hybrid	26.3M	53.3	38.7	72.6%
Swin-B		51.9		
ConvNeXt-B		52.7		
FAN-L-Hybrid	76.8M	54.1	40.6	75.0%
Swin-B†		53.0		
ConvNeXt-B†		54.0		
FAN-L-Hybrid†	76.8M	55.1	42.0	**76.2%**



International Conference On Machine Learning

Code Available

https://github.com/NVlabs/FAN