Top-Down Visual Attention from Analysis by Synthesis

Baifeng Shi, Trevor Darrell, Xin Wang

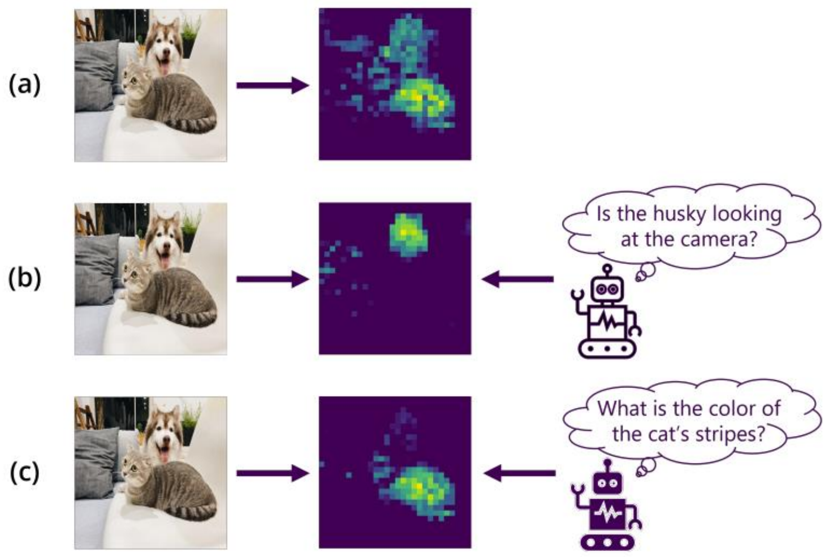
TUE-AM-201

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH
Microsoft Research

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

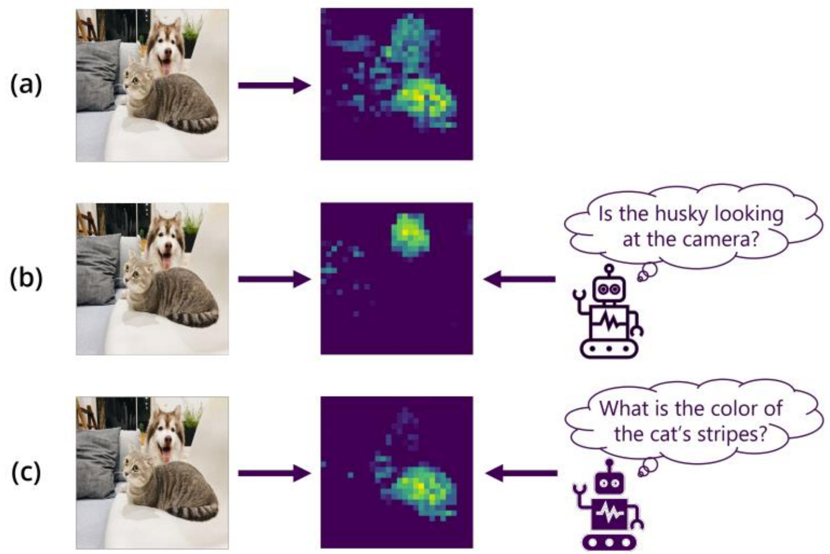
Bottom-Up vs. Top-Down Attention

- Bottom-up attention is stimulus-driven
  - Example: self-attention
  - Normally highlights all the salient objects
- Top-down attention is goal-directed
  - Only highlight the object relevant to the current task.
  - Helps extract representation that is adaptive to different tasks.



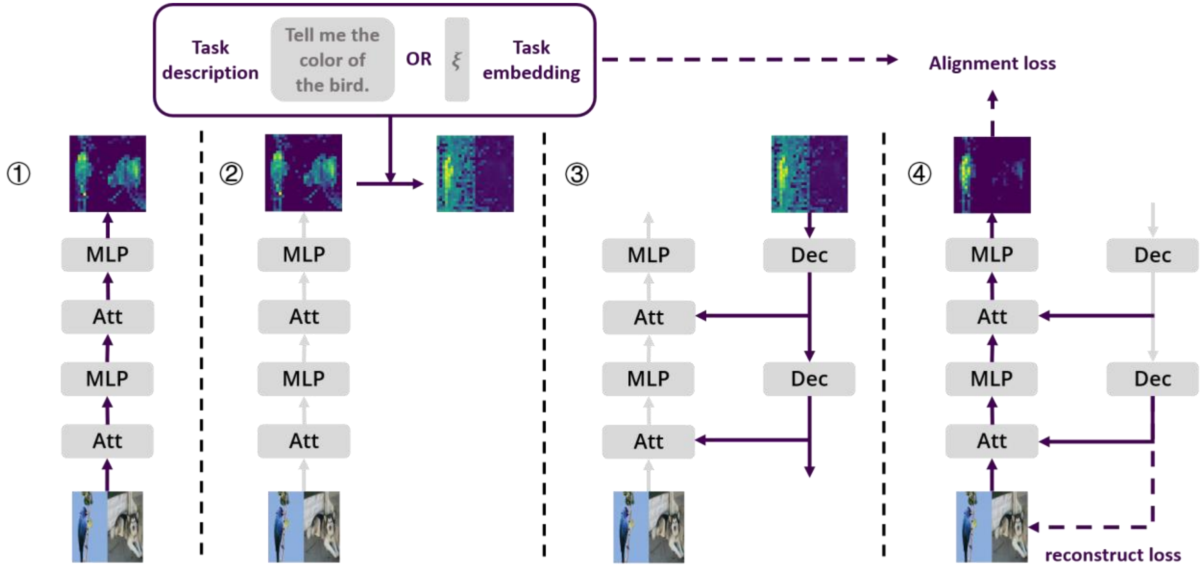
Motivation

- Top-down attention lacks a principled design.
- Current TD-Att algorithms are not compatible to self-attention/transformers



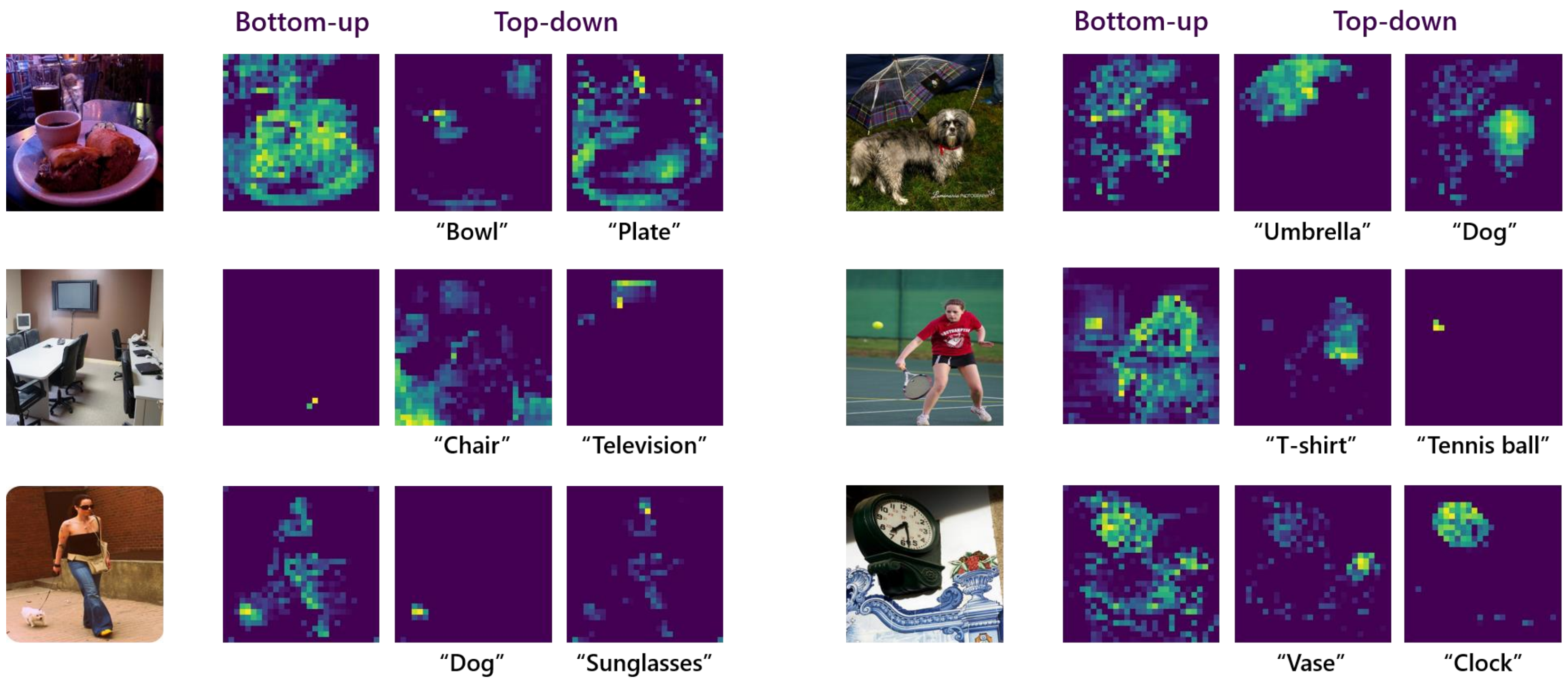
AbSViT: Analysis-by-Synthesis Vision Transformer

- We design AbSViT, a ViT-based model that learns top-down attention.
  - inspired by the classical idea of “Analysis by Synthesis”.



AbSViT: Analysis-by-Synthesis Vision Transformer

- AbSViT is able to adjust its attention to different objects given different instructions or tasks.



AbSViT: Analysis-by-Synthesis Vision Transformer

- AbSViT can improve performance on
  - Vision-Language tasks, such as VQA and zero-shot image retrieval
  - Vision-only tasks, such as image classification and semantic segmentation.

Model	VQAv2		Flickr-Zero-Shot		
	test-dev	test-std	IR@1	IR@5	IR@10
BEiT-B-16 [4]	68.45	-	32.24	-	-
CLIP-B-32 [46]	69.69	-	49.86	-	-
ViT-B	67.89	67.92	42.40	77.18	86.82
- PerceiverIO	67.87	67.93	42.52	76.92	86.73
- Feedback	67.99	68.13	42.04	77.38	86.90
- MaskAtt	67.53	67.51	41.89	76.53	86.78
- AbSViT	**68.72**	**68.78**	**45.28**	**77.98**	**87.52**

Model	PASCAL VOC	Cityscapes	ADE20K
ResNet-101 [12]	77.1	**78.7**	42.9
ViT-B	80.1	75.3	45.2
AbSViT-B	**81.3** (+1.2)	76.8 (+1.5)	**47.2** (+2.0)

Model	P/F	Clean	IN-C (↓)	IN-A	IN-SK	IN-R
PiT-B [27]	74/12.5	82.4	48.2	33.9	43.7	32.3
PVT-L [62]	61/9.8	81.7	59.8	26.6	42.7	30.2
Swin-B [38]	88/15.4	83.4	54.4	35.8	46.6	32.4
ConvNext-B [39]	89/15.4	83.8	46.8	36.7	51.3	38.2
ViT-B [18]	87/17.2	80.8	49.3	25.2	**43.3**	31.6
- AbS	99/38.9	**81.0**	**48.3**	**28.2**	42.9	31.7
RVT-B [42]	86/17.7	80.9	52.1	26.6	**39.6**	26.1
- AbS	100/39.5	80.9	**51.7**	**28.5**	39.3	26.0
FAN-B [75]	54/10.4	83.5	45.0	33.2	51.4	39.3
- AbS	62/21.8	**83.7**	**44.1**	**38.4**	**52.0**	**39.8**

Top-Down Attention from Analysis by Synthesis (AbS)

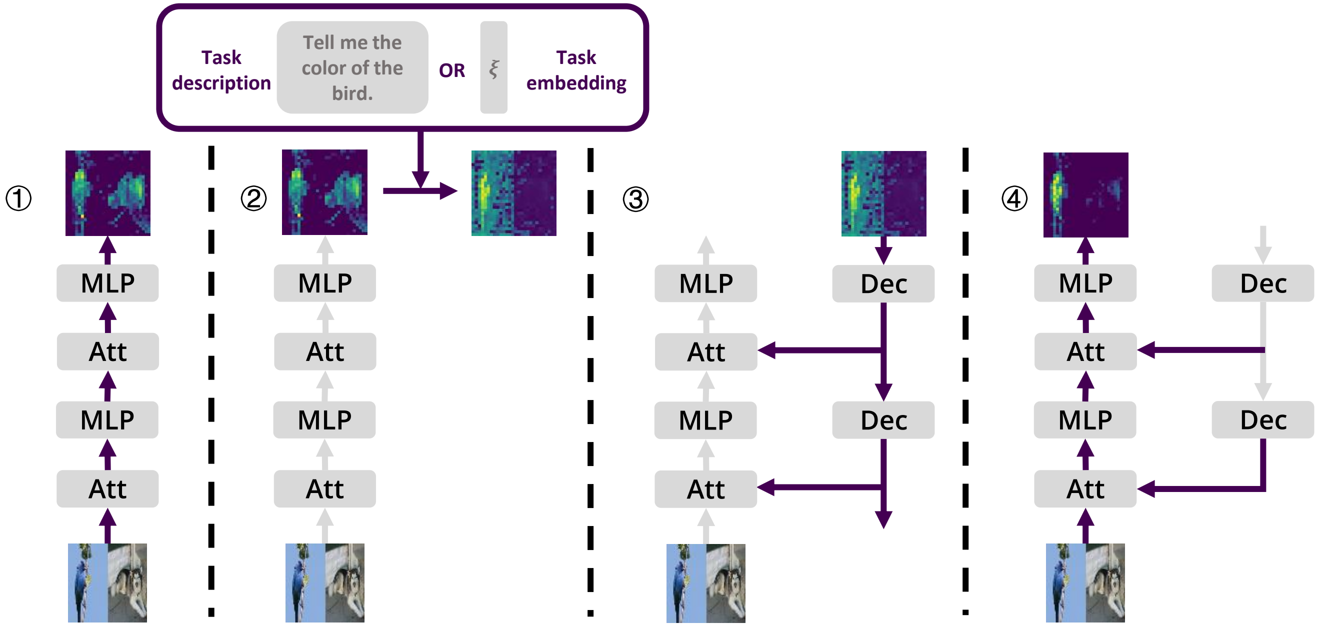
- AbS: vision as Bayesian inference: $\mathbf{z}^* = \arg\max_{\mathbf{z}} p(\mathbf{h}|\mathbf{z})p(\mathbf{z}).$
    - Our perception of the world is affected by a prior.
- We can formulate top-down attention as Bayesian inference:
    - Intuition: **High-level task is a Bayesian prior**, which affects the latent representation (and attention)
    - See math details in the paper
- => If the model learns Bayesian Inference, it learns top-down attention.

Analysis-by-Synthesis Vision Transformer (AbSViT)

- We propose AbSViT, a ViT-based model with top-down attention.
- It learns top-down attention by approximating Bayesian Inference.

AbSViT: Architecture

- AbSViT processes an image with **four steps**:
- ① **First feedforward.** The input image is passed through the feedforward backbone which is a regular ViT. Normally this step will highlight all the salient objects in the image.
- ② **Feature selection.** The output tokens from the first feedforward are reweighted based on their relevance to the task description or the task embedding. This step coarsely selects the tokens that are relevant to the task.
- ③ **Feedback.** The reweighted tokens are sent back through the feedback path as the top-down inputs to the intermediate attention layers.
- ④ **Second Feedforward.** We run the feedforward path again, but this time each attention layer receives an additional top-down input. The top-down input is added on the bottom-up signal for value matrix, and in this way highlights the task-relevant objects.

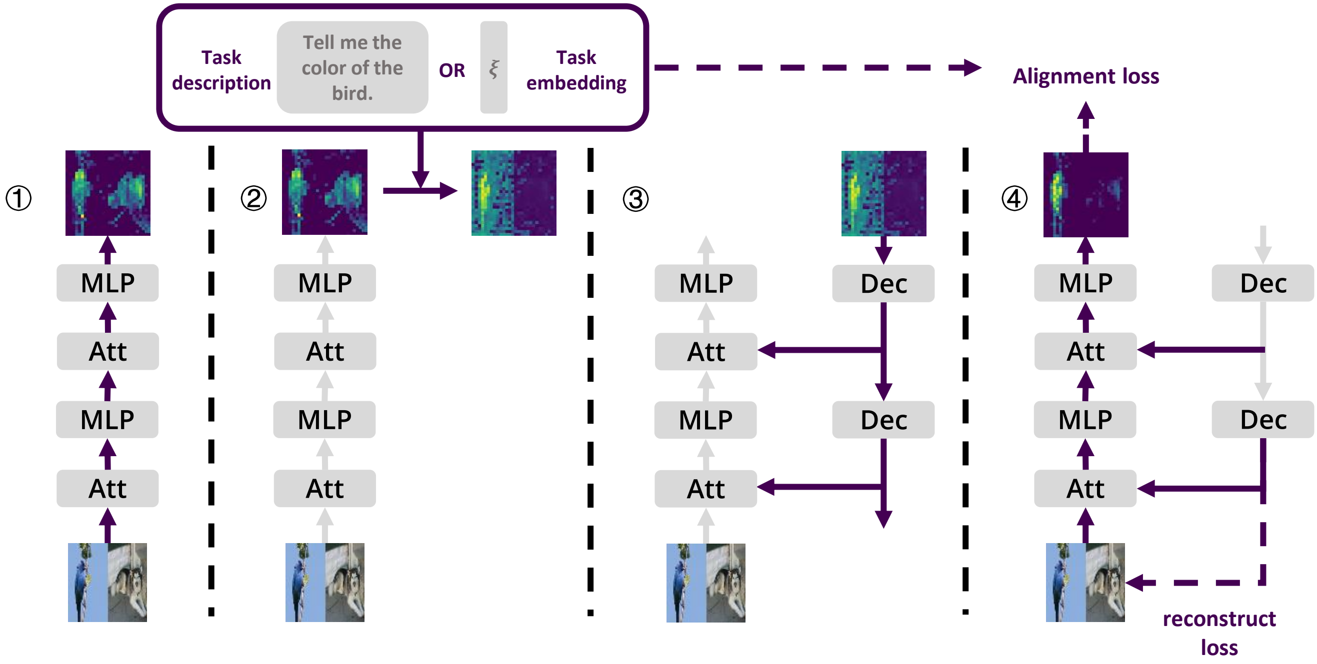


AbSViT: Learning Objective

- AbSViT is trained on two losses to variationally approximate Bayesian Inference (**similar to VAE**):

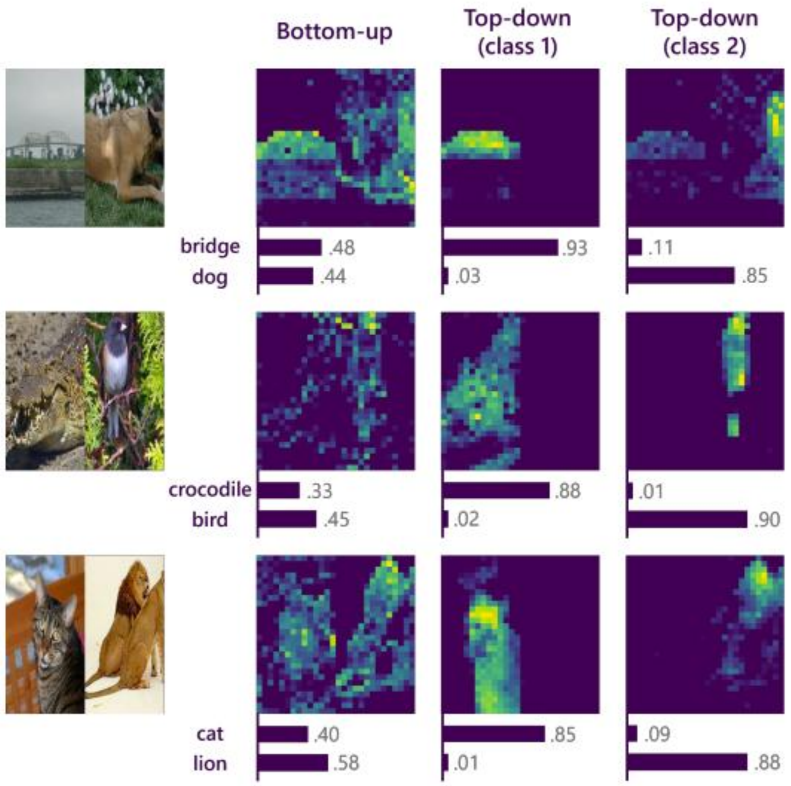
1) **Alignment loss**: make sure the output is aligned with the prior (task description or task embedding)

2) **Reconstruction loss**: encourage the feedback path to reconstruct the input image from the output tokens.



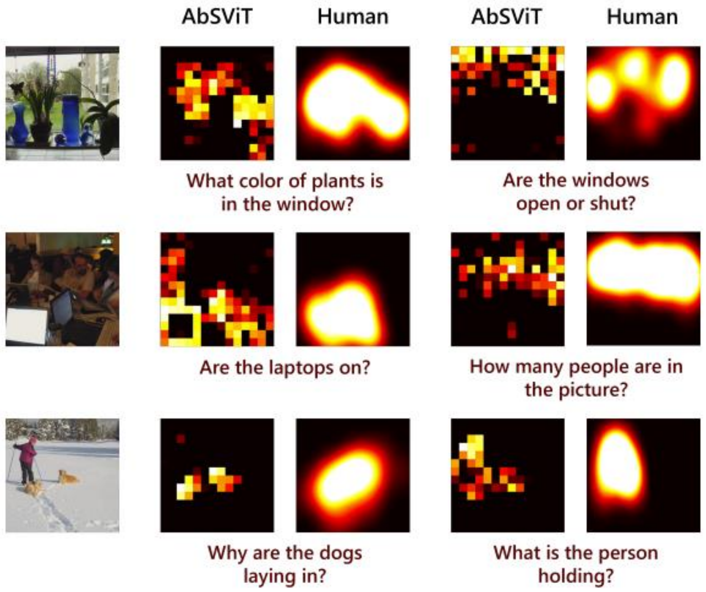
Top-Down Attention of AbSViT

- We show that AbSViT is able to adjust its attention with different priors.



Results on V&L Tasks

- We train Vision-Language tasks such as VQA and zero-shot image retrieval.
    - We use language to guide top-down attention.
- We show that AbSViT is able to adjust its attention based on the language input.
- AbSViT improves the performance on both tasks.



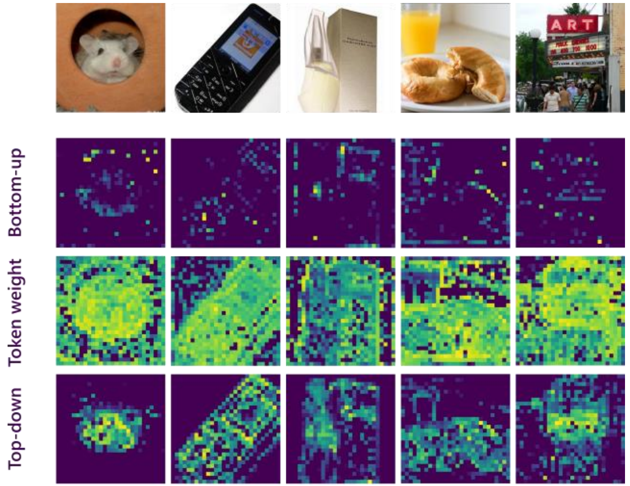
Model	VQAv2		Flickr-Zero-Shot		
	test-dev	test-std	IR@1	IR@5	IR@10
BEiT-B-16 [4]	68.45	-	32.24	-	-
CLIP-B-32 [46]	69.69	-	49.86	-	-
ViT-B	67.89	67.92	42.40	77.18	86.82
- PerceiverIO	67.87	67.93	42.52	76.92	86.73
- Feedback	67.99	68.13	42.04	77.38	86.90
- MaskAtt	67.53	67.51	41.89	76.53	86.78
- AbSViT	**68.72**	**68.78**	**45.28**	**77.98**	**87.52**

Result on Image Classification and Semantic Segmentation

- We train vision-only tasks such image classification and semantic segmentation.
  - We use a learnable task-embedding to guide top-down attention
- We show that AbSViT has a cleaned attention on the foreground objects.
- AbSViT improves the performance on both tasks, and also improves the robustness against noisy or OOD images.

Model	PASCAL VOC	Cityscapes	ADE20K
ResNet-101 [12]	77.1	**78.7**	42.9
ViT-B	80.1	75.3	45.2
AbSViT-B	**81.3** (+1.2)	76.8 (+1.5)	**47.2** (+2.0)

Model	P/F	Clean	IN-C (↓)	IN-A	IN-SK	IN-R
PiT-B [27]	74/12.5	82.4	48.2	33.9	43.7	32.3
PVT-L [62]	61/9.8	81.7	59.8	26.6	42.7	30.2
Swin-B [38]	88/15.4	83.4	54.4	35.8	46.6	32.4
ConvNext-B [39]	89/15.4	83.8	46.8	36.7	51.3	38.2
ViT-B [18]	87/17.2	80.8	49.3	25.2	**43.3**	31.6
- AbS	99/38.9	**81.0**	**48.3**	**28.2**	42.9	31.7
RVT-B [42]	86/17.7	80.9	52.1	26.6	**39.6**	26.1
- AbS	100/39.5	80.9	**51.7**	**28.5**	39.3	26.0
FAN-B [75]	54/10.4	83.5	45.0	33.2	51.4	39.3
- AbS	62/21.8	**83.7**	**44.1**	**38.4**	**52.0**	**39.8**



Thank you!

Webpage

GitHub

ArXiv