

Bottom-Up and Top-Down Attention in Deep Vision Models

Baifeng Shi



Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2024-49

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-49.html>

May 6, 2024

Copyright © 2024, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Bottom-Up and Top-Down Attention in Deep Vision Models

by Baifeng Shi

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences,
University of California at Berkeley, in partial satisfaction of the requirements for the
degree of **Master of Science, Plan II.**

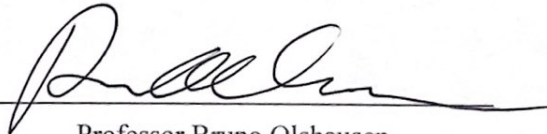
Approval for the Report and Comprehensive Examination:

Committee


Professor Trevor Darrell
Research Advisor

05 / 03 / 2024

(Date)



Professor Bruno Olshausen
Second Reader

05 / 03 / 2024

(Date)

Abstract

Bottom-Up and Top-Down Attention in Deep Vision Models

by

Baifeng Shi

in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Trevor Darrell, Chair

Visual attention helps achieve robust perception under noise, corruption, and distribution shifts in human vision, which are areas where modern neural networks still fall short. In the first chapter, we present VARS, Visual Attention from Recurrent Sparse reconstruction, a new attention formulation built on two prominent features of the human visual attention mechanism: *recurrency* and *sparsity*. Related features are grouped together via recurrent connections between neurons, with salient objects emerging via sparse regularization. VARS adopts an attractor network with recurrent connections that converges toward a stable pattern over time. Network layers are represented as ordinary differential equations (ODEs), formulating attention as a recurrent attractor network that equivalently optimizes the sparse reconstruction of input using a dictionary of “templates” encoding underlying patterns of data. We show that self-attention is a special case of VARS with a single-step optimization and no sparsity constraint. VARS can be readily used as a replacement for self-attention in popular vision transformers, consistently improving their robustness across various benchmarks.

While VARS is stimulus-driven and highlights all the salient objects in an image, intelligent agents like humans often guide their attention based on the high-level task at hand, focusing only on task-related objects. This ability of task-guided top-down attention provides task-adaptive representation and helps the model generalize to various tasks. In the second chapter, we consider top-down attention from a classic Analysis-by-Synthesis (AbS) perspective of vision. VARS indicates a functional equivalence between visual attention and sparse reconstruction; we show that an AbS visual system that optimizes a similar sparse reconstruction objective modulated by a goal-directed top-down signal naturally simulates top-down attention. We further propose Analysis-by-Synthesis Vision Transformer (AbSViT), which is a top-down modulated ViT model that variationally approximates AbS, and achieves controllable top-down attention. For real-world applications, AbSViT consistently improves over baselines on Vision-Language tasks such as VQA and zero-shot retrieval where language guides the top-down attention. AbSViT can also serve as a general backbone, improving

performance on classification, semantic segmentation, and model robustness.

Contents

Contents	i
1 Visual Attention Emerges from Recurrent Sparse Reconstruction	1
1.1 Introduction	1
1.2 Related Work	2
1.3 VARS Formulation	3
1.4 Experiments	9
1.5 Conclusion	15
2 Top-Down Visual Attention from Analysis by Synthesis	16
2.1 Introduction	16
2.2 Related Work	18
2.3 Preliminaries: Attention as Sparse Reconstruction	19
2.4 Top-Down Attention from AbS	20
2.5 Analysis-by-Synthesis Vision Transformer	21
2.6 Experiments	25
2.7 Limitations and Future Work	32
2.8 Conclusion	33
Bibliography	34
A Appendix	44
A.1 Additional Qualitative Results	44
A.2 Details on Derivation of the Sparse Reconstruction Problem	51
A.3 Derivation of Eq. (10)	51
A.4 Additional Results on Natural Images	52
A.5 Additional Implementation Details	52

Chapter 1

Visual Attention Emerges from Recurrent Sparse Reconstruction

1.1 Introduction

One of the hallmarks of human visual perception is its robustness under severe noise, corruption, and distribution shifts [6, 7]. Although having surpassed human performance on ImageNet [44], convolutional neural networks (CNNs) are still far behind the human visual systems on robustness [28, 36] – CNNs are vulnerable under random image corruption [47], adversarial perturbation [116], and distribution shifts [125, 27].

Vision transformers [30] have been reported to be more robust to image corruption and distribution shifts than CNNs under certain conditions [89, 96]. One hypothesis is that the self-attention module, a key component of vision transformers, helps improve robustness [96], achieving state-of-the-art performance on a variety of robustness benchmarks [81]. Although the robustness of vision transformers still seems to be far behind human vision [49, 28], recent work suggests that attention is a key to achieving (perhaps human-level) robustness in computer vision.

The cognitive science literature has also suggested a close relationship between attention mechanisms and robustness in human vision [60, 133].¹ For example, visual attention in human vision has been shown to selectively amplify certain patterns in the input signal and repress others that are not desired or meaningful, leading to robust recognition under challenging conditions such as occlusion [118], clutter [87, 124], and severe corruptions [134].

These findings motivate us to improve robustness of neural networks by designing attention inspired by the human visual attention. However, despite the existing computational models of human visual attention [147], their concrete instantiation in DNNs is still missing, and the connection between human visual attention and existing attention designs is also vague [114].

¹Here we limit our focus on bottom-up attention, *i.e.*, attention that fully depends on input and is not modulated by the high-level task. See [147] for a review on different types of attention in the human visual system.

In this work, we introduce VARS—Visual Attention from Recurrent Sparse reconstruction—a new attention formulation inspired by the *recurrency* and *sparsity* commonly observed in the human visual system. Human visual attention contains the process of grouping and selecting salient features while repressing irrelevant signals [26]. One of its neural foundations is the recurrent connections between neurons in the same layer, as opposed to the feed-forward connections from lower to higher layers [115, 38, 9, 71, 65]. By iteratively connecting neurons, salient features with strong correlations are grouped and amplified [106, 91].

Sparsity, on the other hand, also plays an important role in visual attention, where distracting information is muted by the sparsity constraint and only the most salient parts of the input remains [13, 129]. Sparsity is also a natural outcome of recurrent connections [108], and the two together can be used to formulate visual attention.

Built upon this observation, VARS shows that visual attention naturally emerges from a recurrent sparse reconstruction of input signals in deep neural networks. We start from an ordinary differential equation (ODE) description of neural networks and adopt an attractor network [41, 151, 144] to describe the recurrently connected neurons that arrive at an equilibrium state over time. We then reformulate the computation model into an encoder-decoder style module and show that by adding inhibitory recurrent connections between the encoding neurons, the ODE is equivalent to optimizing the sparse reconstruction of the input using a learned dictionary of “templates” encoding underlying data patterns. In practice, the feed-forward pathway of our attention module only involves optimizing the sparse reconstruction, which can be efficiently solved by the iterative shrinkage-thresholding algorithm [40].

We present multiple variants of VARS by instantiating the learned dictionary in the sparse reconstruction as a **static** (input-independent), **dynamic** (input-dependent), or **static+dynamic** (combination of the two) set of templates. We show that the existing self-attention design [121], widely adopted in vision transformers, is a special case of VARS with a dynamic dictionary but only using a single step update of the ODE without sparsity constraints. VARS extends self-attention and exhibits higher robustness in practice.

We evaluate VARS on five large-scale robustness benchmarks of naturally corrupted, adversarially perturbed and out-of-distribution images on ImageNet, where VARS consistently outperforms previous methods. We also assess the quality of attention maps on human eye fixation and image segmentation datasets, and show that VARS produces higher quality attention maps than self-attention.

1.2 Related Work

Recurrency in vision. Recurrent connections are as ubiquitous as feed-forward connections in the human visual system [35]. Various phenomena in human vision are credited to recurrency, such as visual grouping and pattern completion [107, 91], robust recognition and segmentation under clutter [122], and even perceptual illusions [85].

In deep learning, Nayebi *et al.* [90] find that convolutional recurrent models can better capture neural dynamics in the primate visual system. Other work has designed recurrent modules to help networks attend to salient features such as contours [73], to group and segregate an object from its context [61, 21], or to conduct Bayesian inference of corrupted images [56]. Zoran *et al.* [150] combine an LSTM [53] and self-attention to improve adversarial robustness. However, the LSTM is only used to generate the queries of self-attention and does not affect the attention mechanism itself. Instead, we show visual attention emerges from recurrency and can be formulated as a recurrent contractor network, which is equivalent to optimizing sparse reconstruction of the input signals.

Sparsity in vision. It has long been hypothesized that the primary visual cortex (V1) encodes incoming stimuli in a sparse manner [94]. Olshausen and Field [93] show that localized and oriented filters resembling the simple cells in the visual cortex can spontaneously emerge through dictionary learning via sparse coding. Some work has extended this hypothesis to the prestriate cortex such as V2 [66]. Rozell *et al.* [108] propose Locally Competitive Algorithms (LCA) as a biologically-plausible neural mechanism for computing sparse representations in the visual cortex based on neural recurrency.

Recently there are studies on designing sparse neural networks, but they mostly focus on reducing the computational complexity via sparse weights or connections [54, 74, 42]. Other work has exploited sparse data structure in neural networks [132, 34]. In this work, we draw a connection between sparsity and recurrency as well as visual attention and build a new attention formulation on top of it to improve the robustness of deep neural networks.

Visual attention and robustness. Attention widely exists in human brains [110] and is adopted in machine learning models [121]. A number of studies focus on object-based and bottom-up attention. Various computational models have been proposed for visual attention [64, 57, 147], where neural recurrency help highlight salient features.

In deep learning, the attention mechanism is widely used to process language or visual data [121, 30]. The recently proposed self-attention based vision transformers [30] can scale to large datasets better than conventional CNNs and achieve better robustness under distribution shifts [81, 89, 96].

In a complementary line of work to address model robustness, researchers have shown that model robustness can be improved through data augmentation [37, 104], designing more robust model architectures [29] and training strategies [80, 128]. Some other work [79, 112] improves model robustness from a bio-inspired view. In this work, we propose a new attention design, (partially) inspired by human vision systems, which can be used as a replacement for self-attention in vision transformers to further improve model robustness.

1.3 VARS Formulation

In this section, we first formulate the neural recurrency based on an ODE description of neural dynamics (Section 1.3) and show its equivalence to optimizing the sparse reconstruction of

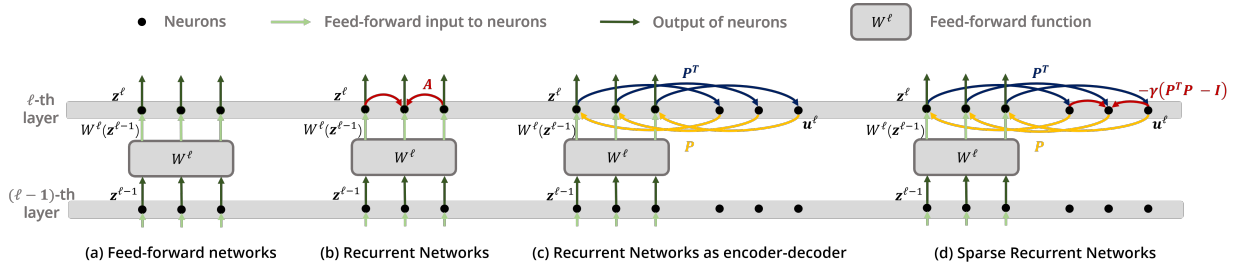


Figure 1.1: **Illustrations on neural dynamics.** (a) *Feed-forward networks.* The output $\mathbf{z}^{\ell-1}$ from the $(\ell - 1)$ -th layer’s neurons is processed by the feed-forward function W^ℓ into $W^\ell(\mathbf{z}^{\ell-1})$, used as the input to the ℓ -th layer’s neurons. The ℓ -th layer’s neuron output \mathbf{z}^ℓ is identical to the input. (b) *Recurrent networks.* When the ℓ -th layer’s neurons are recurrently connected, the output \mathbf{z}^ℓ is wired back by weight matrix \mathbf{A} serving as an additional input to the neurons. The final output is the steady state $\mathbf{z}^{\ell*} = \mathbf{A}\mathbf{z}^{\ell*} + W^\ell(\mathbf{z}^{\ell-1})$. (c) *Recurrent networks as encoder-decoder.* The neurons with recurrent connections in (b) have the same steady state as an encoder-decoder structure, where the auxiliary layer encodes \mathbf{z}^ℓ by \mathbf{P}^T and its output is decoded by \mathbf{P} and sent back. (d) *Sparse recurrent networks.* We adopt a sparse structure in the encoding by adding inhibitive recurrent connections $-\gamma(\mathbf{P}^T\mathbf{P} - \mathbf{I})$ between \mathbf{u}^ℓ .

the inputs (Section 1.3). Then we draw a connection between visual attention and recurrent sparse reconstruction and describe the design of VARS (Section 1.3). We find self-attention is a special case of VARS (Section 1.3) and give different model instantiations in Section 1.3.

Neural Recurrency

ODE descriptions of neural dynamics. We start with an ODE description of feed-forward neural networks [23]. Let $\mathbf{z}^\ell \in \mathbb{R}^d$ denote the output of the ℓ -th layer’s neurons in a neural network.² In a feed-forward neural network, the output of the ℓ -th layer is

$$\mathbf{z}^\ell = W^\ell(\mathbf{z}^{\ell-1}), \quad (1.1)$$

where $W^\ell(\cdot)$ is a feed-forward function (*e.g.*, a convolutional or fully connected operator). This equation can be viewed as an *equilibrium state* ($\frac{d\mathbf{z}^\ell}{dt} = 0$) of the following differential equation:

$$\frac{d\mathbf{z}^\ell}{dt} = -\mathbf{z}^\ell + W^\ell(\mathbf{z}^{\ell-1}), \quad (1.2)$$

which defines the neural dynamics of ℓ -th layer’s neurons. This can be seen as the (simplified) dynamics of biological neurons: \mathbf{z}^ℓ is the membrane potential which is charged by the

²Although \mathbf{z}^ℓ may have shape like $h \times w \times c$, in our formulation we always use the vectorized version of the input, *i.e.*, $d = hwc$.

feed-forward input $W^\ell(\mathbf{z}^{\ell-1})$ and discharged by the self leakage $-\mathbf{z}^\ell$ [23]. Note that in the feed-forward case, the output \mathbf{z}^ℓ of the neurons is identical to the input $W^\ell(\mathbf{z}^{\ell-1})$. Figure 1.1(a) provides an illustration.

Horizontal recurrent connections. In the feed-forward case, the input $W^\ell(\mathbf{z}^{\ell-1})$ to the ℓ -th layer solely depends on the output $\mathbf{z}^{\ell-1}$ from the previous layer. However, when the neurons in the ℓ -th layer are recurrently connected (illustrated in Figure 1.1(b)), the input also depends on the ℓ -th layer's output \mathbf{z}^ℓ itself, which we denote by $\widetilde{W}^\ell(\mathbf{z}^{\ell-1}, \mathbf{z}^\ell)$. Therefore, the output of a recurrently connected layer is the equilibrium state of an updated differential equation

$$\frac{d\mathbf{z}^\ell}{dt} = -\mathbf{z}^\ell + \widetilde{W}^\ell(\mathbf{z}^{\ell-1}, \mathbf{z}^\ell). \quad (1.3)$$

In this case, the equilibrium $\mathbf{z}^{\ell*} = \widetilde{W}^\ell(\mathbf{z}^{\ell-1}, \mathbf{z}^{\ell*})$ ³ typically does not have a closed-form solution, and in practice the differential equation is often solved by rolling out each step of updates as in recurrent neural networks (RNNs) [53] or by using root-finding techniques [4].

Following the previous work [147], we decompose $\widetilde{W}^\ell(\mathbf{z}^{\ell-1}, \mathbf{z}^\ell)$ into a feed-forward input $W^\ell(\mathbf{z}^{\ell-1})$ and an additional recurrent input $\mathbf{A}^\ell \mathbf{z}^\ell$ and rewrite Equation 1.3 as

$$\frac{d\mathbf{z}^\ell}{dt} = -\mathbf{z}^\ell + \mathbf{A}^\ell \mathbf{z}^\ell + \mathbf{x}^\ell \quad (1.4)$$

where $\mathbf{x}^\ell = W^\ell(\mathbf{z}^{\ell-1})$ is the feed-forward input of the neurons in the ℓ -th layer and $\mathbf{A}^\ell \in \mathbb{R}^{d \times d}$, often assumed symmetric and positive semi-definite [55, 17], is the weight of horizontal recurrent connections of neurons in the ℓ -th layer (Figure 1.1(b)). Note that Equation 1.4 is also a special case of the continuous attractor neural network [131], which is used as a computational model for various neural behaviors such as visual attention [147]. In what follows, for simplicity and without loss of generality, we only focus on a single-layer scenario and omit the superscript ℓ .

Recurrency Entails Sparse Reconstruction

We have defined the neural recurrency in ODEs and now we build its connection to the optimization of sparse reconstruction to understand the functionality of recurrency.

We notice that Equation 1.4 can also be viewed as an encoder-decoder structure. Since $\{\mathbf{A} \mid \mathbf{A} \in \mathcal{S}_+^d\} = \{\mathbf{P}\mathbf{P}^T \mid \mathbf{P} \in \mathbb{R}^{d \times d'}, d' \geq d\}$, we can reparameterize \mathbf{A} as $\mathbf{P}\mathbf{P}^T$, and turn Equation 1.4 into

$$\frac{d\mathbf{z}}{dt} = -\mathbf{z} + \mathbf{P}\mathbf{P}^T \mathbf{z} + \mathbf{x}, \quad (1.5)$$

which has the same steady state solution as

$$\begin{cases} \frac{d\mathbf{z}}{dt} = -\mathbf{z} + \mathbf{P}\mathbf{u} + \mathbf{x}, \\ \frac{d\mathbf{u}}{dt} = -\mathbf{u} + \mathbf{P}^T \mathbf{z}. \end{cases} \quad (1.6)$$

$$\frac{d\mathbf{u}}{dt} = -\mathbf{u} + \mathbf{P}^T \mathbf{z}. \quad (1.7)$$

³We use asterisks to denote equilibrium states.

Here, we introduce an auxiliary layer $\mathbf{u} \in \mathbb{R}^{d'}$ that receives input $\mathbf{P}^T \mathbf{z}$ and converges to $\mathbf{u}^* = \mathbf{P}^T \mathbf{z}^*$. Meanwhile \mathbf{z} converges to $\mathbf{z}^* = \mathbf{P} \mathbf{u}^* + \mathbf{x}$. We can view the steady state solution as an equilibrium between an encoder and a decoder, where the column vectors of \mathbf{P} serve as atoms (or “templates”) of a dictionary, \mathbf{u}^* is the encoding of \mathbf{z}^* through the template matching $\mathbf{P}^T \mathbf{z}^*$, and $\mathbf{z}^* = \mathbf{P} \mathbf{u}^* + \mathbf{x}$ is the decoding of \mathbf{u}^* plus a residual connection \mathbf{x} . (Figure 1.1(c)).

Next, we show that this encoder-decoder structure naturally connects with the hypothesis of sparse coding in V1, which states that the encoding of visual signals should be sparse [94]. Moreover, sparsity naturally emerges as we build the inhibitive recurrent connections between the encoding neurons (\mathbf{u} in our case) [108]. Specifically, the encoding is sparse when there exists mutual inhibition between neurons that encode the similar feature and mutual excitation between neurons that encode different features. To show this, we follow [108] and add recurrent connections between \mathbf{u} , modeled by the weight matrix $-\gamma(\mathbf{P}^T \mathbf{P} - \mathbf{I})$.⁴ In addition, we add hyperparameters α and β to control the strength of self-leakage, and the element-wise activation functions $g(\cdot)$ to gate the output from the neurons [23]. As a result, we update the dynamics of \mathbf{z} and \mathbf{u} as (see also Figure 1.1(d)):

$$\begin{cases} \frac{d\mathbf{z}}{dt} = -\alpha \mathbf{z} + \mathbf{P} g(\mathbf{u}) + \mathbf{x}, & (1.8) \\ \frac{d\mathbf{u}}{dt} = -\beta \mathbf{u} - \gamma(\mathbf{P}^T \mathbf{P} - \mathbf{I}) g(\mathbf{u}) + \mathbf{P}^T \mathbf{z}. & (1.9) \end{cases}$$

By taking $\alpha = 1$ and $\beta = \gamma = 2$ and choosing g as an element-wise thresholding function $g(\mathbf{u}_i) = \text{sgn}(\mathbf{u}_i) \cdot (|\mathbf{u}_i| - \frac{\lambda}{2})_+$, where $\text{sgn}(\cdot)$ is the sign function, $(\cdot)_+$ is ReLU, and λ controls the sparse constraint (see Appendix for more details), Equation 1.8-1.9 have the equilibrium state as

$$\begin{cases} \tilde{\mathbf{u}}^* = \arg \min_{\tilde{\mathbf{u}} \in \mathbb{R}^{d'}} \frac{1}{2} \|\mathbf{P} \tilde{\mathbf{u}} - \mathbf{x}\|_2^2 + \lambda \|\tilde{\mathbf{u}}\|_1, & (1.10) \\ \mathbf{z}^* = \mathbf{P} \tilde{\mathbf{u}}^* + \mathbf{x}, & (1.11) \end{cases}$$

where $\tilde{\mathbf{u}} = g(\mathbf{u})$ is the gated output of the encoding neurons. We can see that by adding the recurrent connections in \mathbf{u} and the activation functions $g(\cdot)$, the output of ℓ -th layer’s neurons is not only a simple “copy-paste” of the feed-forward input \mathbf{x} (Equation 1.1) but also with a sparse reconstruction of the input $\mathbf{P} \tilde{\mathbf{u}}^*$. This formulation also indicates that solving the dynamics of a sparse recurrent network is equivalent to sparse reconstruction of the input signal.

VARS: Attention from Sparse Reconstruction

The core design of VARS is based on the observation that visual attention is achieved through (i) *grouping* different features and different locations into separate objects and (ii) *selecting* the most salient objects and suppressing distracting or noisy ones [26]. Note that the dynamics of sparse encoding \mathbf{u} (Equation 1.9) contains a similar process, where the features in \mathbf{z} are

⁴Each \mathbf{u}_μ encodes \mathbf{z} by matching it with the feature template \mathbf{P}^μ , the μ -th column of \mathbf{P} . The inhibition strength between \mathbf{u}_μ and \mathbf{u}_ν is $\gamma(\mathbf{P}^{\mu T} \mathbf{P}^\nu - 1)$, which is higher when \mathbf{u}_μ and \mathbf{u}_ν encodes similar features.

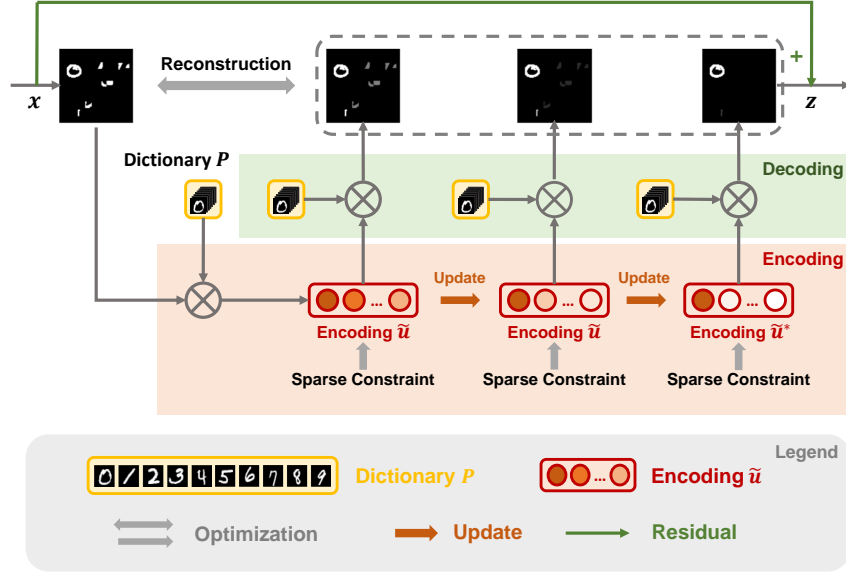


Figure 1.2: **Overview of VARS.** First, we initialize $\tilde{\mathbf{u}}$ as $\mathbf{P}^T \mathbf{x}$, the encoding of the input. Then, for each iteration, we update $\tilde{\mathbf{u}}$ to minimize the reconstruction error between \mathbf{x} and the decoded $\mathbf{P}\tilde{\mathbf{u}}$, as well as the sparsity constraint. After multiple steps, the converged $\tilde{\mathbf{u}}^*$ is decoded and output together with a residual term.

grouped by each template \mathbf{P}^μ through $(\mathbf{P}^\mu)^T \mathbf{z}$ and fed into the encoding \mathbf{u}_μ ,⁵ meanwhile the recurrent term $-\gamma(\mathbf{P}^T \mathbf{P} - \mathbf{I})g(\mathbf{u})$ imposes a sparse structure on \mathbf{u} so that only the \mathbf{u}_μ that encodes the most salient template objects will survive.

Here we introduce VARS, a module that achieves visual attention via sparse reconstruction following the formulation in Equation 1.10-1.11, *i.e.*, VARS takes a feed-forward input \mathbf{x} and outputs the sparse reconstruction $\mathbf{P}\tilde{\mathbf{u}}^*$ of \mathbf{x} and a residual term. The VARS module can be plugged into neural networks to help attend features.

In practice, VARS optimizes the sparse reconstruction (Equation 1.10) iteratively, as illustrated in Figure 1.2. First, we initialize $\tilde{\mathbf{u}}$ as the encoding of input \mathbf{x} , *i.e.*, $\tilde{\mathbf{u}} \leftarrow \mathbf{P}^T \mathbf{x}$ (the red block in Figure 1.2). Then, for each iteration, we decode $\tilde{\mathbf{u}}$ into $\mathbf{P}\tilde{\mathbf{u}}$ (the green block in Figure 1.2) and update $\tilde{\mathbf{u}}$ by minimizing the reconstruction error $\frac{1}{2}\|\mathbf{P}\tilde{\mathbf{u}} - \mathbf{x}\|_2^2$ and the sparsity constraint $\lambda\|\tilde{\mathbf{u}}\|_1$. We adopt the update rule in the Learned Iterative Shrinkage Thresholding Algorithm (LISTA) [40], *i.e.*, each update is

$$\tilde{\mathbf{u}} \leftarrow \mathcal{S}_{\frac{w_1 \cdot \lambda}{w_2 \cdot L}} \left(\tilde{\mathbf{u}} - \frac{1}{w_2 \cdot L} (\Gamma^T \mathbf{P}^T \mathbf{P} \Gamma \tilde{\mathbf{u}} - \mathbf{P}^T \mathbf{x}) \right), \quad (1.12)$$

⁵We denote the μ -th column of \mathbf{P} by \mathbf{P}^μ . For example, in the binary case, if each template contains an object, *i.e.*, $\mathbf{P}_i^\mu = 1$ if the object occupies the location i , then $(\mathbf{P}^\mu)^T \mathbf{x} = \sum_{i \in \{i | \mathbf{P}_i^\mu = 1\}} \mathbf{x}_i$ is the collection of all features in locations that the object occupies.

where $\mathcal{S}_\lambda(x) = \text{sgn}(x) \cdot (|x| - \lambda)_+$ is an element-wise thresholding function, and L is the largest singular value of $\mathbf{P}^T \mathbf{P}$. $w_1, w_2 \in \mathbb{R}$ and $\mathbf{\Gamma} \in \mathbb{R}^{d' \times d'}$ are learned parameters so that after several iterations of Equation 1.12 the output is close to the sparse reconstruction. After multiple updates, we decode the converged $\tilde{\mathbf{u}}^*$ into $\mathbf{P}\tilde{\mathbf{u}}^*$ and output it together with a residual term.

Overall, VARS groups features of different objects in the input into different encoding neurons, and the most salient objects (*e.g.*, “0” in the input in Figure 1.2) are preserved while other distractors are suppressed by sparse reconstruction. VARS can be easily plugged into any neural network and is computationally efficient thanks to the LISTA algorithm with fast convergence (see Section 2.6).

Self-Attention as a Special Case of VARS

Here we show that self-attention can be viewed as a special case of VARS using a dynamic instantiation of the dictionary with single step approximation and no sparsity constraint.

Self-attention formulation. In self-attention, the input $\mathbf{X} \in \mathbb{R}^{N \times C}$ contains N tokens and C channels. We use the superscript μ for the channel index and the subscript i for the token index, *e.g.*, \mathbf{X}_i^μ is the μ -th channel in the i -th token. In each head, self-attention gives the output \mathbf{Z} by

$$\mathbf{Z}_i^\mu = \sum_j \mathbf{K}(\mathbf{X}, \mathbf{X})_{ij} \cdot \mathbf{X}_j^\mu + \mathbf{X}_i^\mu, \quad (1.13)$$

where $\mathbf{K}(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{N \times N}$ measures the similarity between tokens in \mathbf{X} , *i.e.*, $\mathbf{K}(\mathbf{X}, \mathbf{X})_{ij} = e^{(\mathbf{W}\mathbf{X}_i)^T (\mathbf{W}'\mathbf{X}_j)}$, with \mathbf{W} and \mathbf{W}' as query and key projections.⁶ Self-attention is compute intensive given its quadratic computational complexity. Performer [16], a recently proposed variant of vision transformers, approximates the similarity kernel with the inner product of the feature maps, *i.e.*, $\mathbf{K}(\mathbf{X}, \mathbf{X}) \approx \Phi(\mathbf{X})\Phi'(\mathbf{X})^T$, where $\Phi(\mathbf{X}), \Phi'(\mathbf{X}) \in \mathbb{R}^{N \times C'}$ are specific (random) feature maps of \mathbf{X} .

Connection between self-attention and VARS. Following the formulation in Performer, we can rewrite Equation 1.13 as

$$\mathbf{Z}^\mu = \Phi(\mathbf{X})\Phi(\mathbf{X})^T \mathbf{X}^\mu + \mathbf{X}^\mu. \quad (1.14)$$

Here we use a symmetric similarity kernel by setting $\mathbf{W} = \mathbf{W}'$, which means $\Phi = \Phi'$. This feed-forward computation is a single-step Euler update⁷ of the differential equation

$$\frac{d\mathbf{Z}^\mu}{dt} = -\mathbf{Z}^\mu + \Phi(\mathbf{X})\Phi(\mathbf{X})^T \mathbf{Z}^\mu + \mathbf{X}^\mu, \quad (1.15)$$

which has a similar form with the ODE description of a recurrent layer (Equation 1.5), except that Equation 1.15 uses $\Phi(\mathbf{X})$ as the dictionary which is dependent on the specific input \mathbf{X}

⁶Here we ignore the value projection (as in non-local blocks [127]) as well as the normalization term.

⁷with initialization $\mathbf{Z}^\mu = \mathbf{X}^\mu$ and step-size of 1.

while Equation 1.5 uses a static dictionary \mathbf{P} learned from the entire dataset. This shows self-attention is a variant of recurrent networks using a dynamic dictionary. See Appendix A.1 for the visualization of the dynamic dictionary.

However, compared to VARS (Equation 1.10-1.11), self-attention (Equation 1.15) does not have the inhibitive recurrent connections (Equation 1.9) to turn into sparse reconstruction, and updates the ODE (Equation 1.15) only via a single step. Therefore, we introduce VARS with a dynamic dictionary:

$$\begin{cases} \tilde{\mathbf{U}}^{\mu*} = \arg \min_{\tilde{\mathbf{U}}^{\mu}} \frac{1}{2} \|\Phi(\mathbf{X})\tilde{\mathbf{U}}^{\mu} - \mathbf{X}^{\mu}\|_2^2 + 2\lambda \|\tilde{\mathbf{U}}^{\mu}\|_1 & (1.16) \\ \mathbf{Z}^{\mu*} = \Phi(\mathbf{X})\tilde{\mathbf{U}}^{\mu*} + \mathbf{X}^{\mu}, & (1.17) \end{cases}$$

which optimizes the sparse reconstruction of each channel \mathbf{X}^{μ} in the input. We refer VARS with a static dictionary to as VARS-S (Equation 1.10-1.11) and VARS with an dynamic dictionary as VARS-D (Equation 1.16-1.17).

VARS Instantiations

Dictionary designs. So far, we introduced two instantiations of VARS: VARS with a static dictionary \mathbf{P} (VARS-S) (Equation 1.10-1.11) and VARS with a dynamic dictionary $\Phi(\mathbf{X})$ (VARS-D) (Equation 1.16-1.17). Both variants have their own merits as \mathbf{P} learns a general pattern of the dataset while $\Phi(\mathbf{X})$ captures the specialized information on a per input basis. Therefore, we consider a third instantiation of VARS, VARS-SD to combine both the static and dynamic dictionaries as $[\mathbf{P}; \Phi(\mathbf{X})]$, *i.e.*, using the union of atoms in \mathbf{P} and $\Phi(\mathbf{X})$. We test all three variants in our experiments.

Instantiation of \mathbf{P} . In theory, \mathbf{P} can be any real matrix of the specific shape. However, since each column of \mathbf{P} is a template which is a signal in the spatial feature domain, we may impose certain inductive biases such as translational symmetry when instantiating \mathbf{P} . To this end, we design \mathbf{P} by making its templates kernels with different translations, which means \mathbf{P}^T is a convolution layer, *i.e.*, $(\mathbf{P}^T \mathbf{x})_{\uparrow} = conv(\mathbf{x}_{\uparrow})$ where $(\cdot)_{\uparrow}$ unflattens a vector into a 2D signal. Then \mathbf{P} is a deconvolution layer with its kernel shared with the convolution layer, used in both `static` and `static+dynamic` dictionaries.

1.4 Experiments

We test VARS on five robustness benchmarks on the ImageNet dataset including naturally corrupted, out of distribution, and adversarial images (Section 1.4). We also evaluate our models on the following settings: domain generalization (Section 1.4), image segmentation, and human eye fixation predictions (Section 1.4). Finally, we analyze and ablate the design choices of VARS in Section 1.4.

Experimental Setup. We evaluate on multiple datasets (Table 1.1) and the models are pretrained on ImageNet-1K [25]. For baselines, we consider DeiT [120], a commonly-used

Table 1.1: **Dataset overview.** We evaluate VARS on five robustness benchmarks and perform evaluation in three additional settings.

	Dataset Name	Type
Robustness	ImageNet-C (IN-C) [47]	Natural corruption
	ImageNet-R (IN-R) [49]	Out of distribution
	ImageNet-SK (IN-SK) [125]	Out of distribution
	PGD [80]	Adversarial attack
	ImageNet-A (IN-A) [48]	Natural adv. example
Others	PACS [70]	Domain generalization
	PASCAL VOC [33]	Semantic segmentation
	MIT1003 [59]	Human eye fixation

vision transformer model, and RVT [81], the state-of-the-art vision transformer on various robustness benchmarks. For generality, we also test on GFNet [103] using a linear token-mixer instead of self-attention. We apply VARS in the baselines by replacing the global operators (self-attention or token mixer). For all the models, we adopt the convolutional patch-embedding [136] to facilitate training and also apply Performer approximation which VARS is also built on. We use * to denote the modified baselines. Results of both original and modified baselines are reported.

Evaluation on Robustness Benchmarks

We show the evaluation results on the robustness benchmarks in Table 1.2. First, we can see vision transformers are generally more robust than the CNN counterparts even with an order of magnitude smaller numbers of parameters and FLOPs. We can also observe that VARS consistently improves the baselines across different benchmarks. For example, compared to DeiT, **VARS-SD** reduces the error rate from 67% to 62.5% on IN-C and improves the accuracy from 34.5% to 40.2% on IN-R, from 21.7% to 27.5% on IN-SK, which are over 5 absolute points improvements. Similar results are observed with GFNet and RVT.

Moreover, when built on top of the RVT network design, **VARS-SD** outperforms or is on par with the previous methods across the five benchmarks. Note that VARS is built upon RVT*, the modified version of RVT (see Experimental Setup). As shown in Table 1.2, RVT* has weaker initial performance than the vanilla RVT model, mainly due to the Performer approximation of the self-attention.

In Figure 1.3, we visualize the attention maps of self-attention and **VARS-S**, **-D**, **-SD** under different image corruption scenarios. We see that for a clean image, all the attention designs can roughly locate salient regions around the main object (airplane). However, self-attention only highlights the center part of the object while **VARS-SD** and **VARS-D** capture the contour of the object more clearly. For corrupted images, we observe that attention maps from vanilla self-attention tend to be noisier than VARS. For example, with severe weather corruption

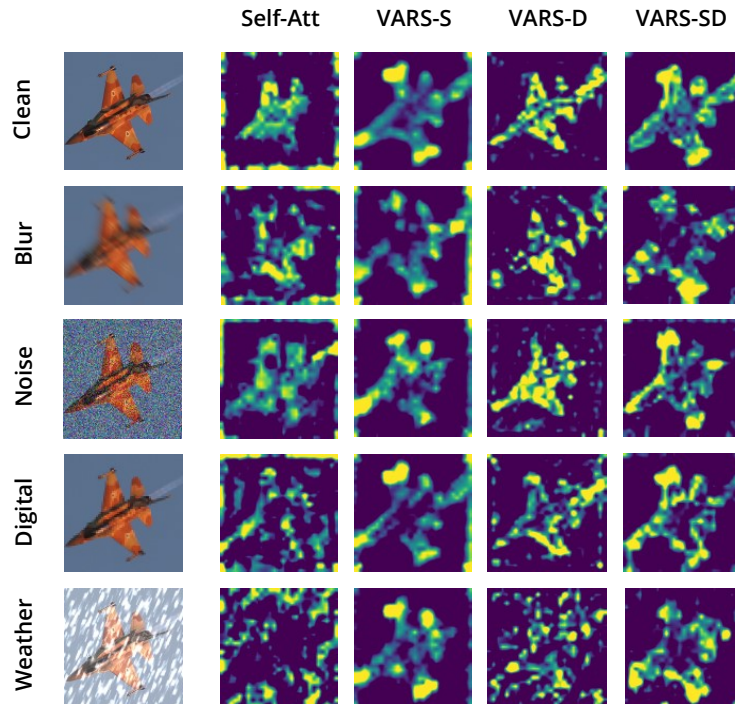


Figure 1.3: **Attention maps under image corruption.** We can see VARS consistently highlights the core parts of the object (airplane) while self-attention can miss them (*e.g.*, weather). The attention maps of VARS-SD are most stable and sharp.

(the last row), self-attention misses the main object and rather highlights the snow effect, while VARS-SD still captures the object and suppresses the noise. We also notice that VARS-S tend to produce a blurrier attention map compared to the ones with a dynamic dictionary, which might be due to the weaker expressivity of static dictionary compared to the dynamic dictionary.

Evaluation on Domain Generalization

Domain generalization is a related setting, which evaluates the models' generalization to unseen domains at test time. Here we finetune the ImageNet-pretrained models on three source domains in PACS [70] and test them on the left-out target domain.

Table 1.3 shows that VARS-SD outperforms the RVT baseline across all four target domains. Specifically, VARS-SD improves RVT* from 81.25% to 86.08% on the *Art* domain and from 94.19% to 96.47% on the *Photo* domain. These results indicate that our attention module is more robust than self-attention when generalizing to unseen domains.

Table 1.2: **Evaluation results on robustness benchmarks.** We find that VARS consistently improves over the self-attention counterparts (DeiT*, GFNet* and RVT*). VARS-SD outperforms or is on par with previous methods despite using a weaker initial model. The best performance of each vision transformer architecture is bold and the underlined values are the overall state-of-the-art performance.

	Model	GFLOPs	Params.(M)	Clean \uparrow	IN-C \downarrow	IN-R \uparrow	IN-SK \uparrow	PGD \uparrow	IN-A \uparrow
CNNs	RegNetY-4GF [101]	4.0	20.6	79.2	68.7	38.8	25.9	2.4	8.9
	ResNet50 [43]	4.1	25.6	79.0	65.5	42.5	31.5	12.5	5.9
	ResNeXt50-32x4d [139]	4.3	25.0	79.8	64.7	41.5	29.3	13.5	10.7
	InceptionV3 [117]	5.7	27.2	77.4	80.6	38.9	27.6	3.1	10.0
Transformers	PiT-Ti [50]	0.7	4.9	72.9	69.1	34.6	21.6	5.1	6.2
	ConViT [20]	1.4	5.7	73.3	68.4	35.2	22.4	7.5	8.9
	PVT [126]	1.9	13.2	75.0	79.6	33.9	21.5	0.5	7.9
	DeiT [120]	1.3	5.7	72.2	71.1	32.6	20.2	6.2	7.3
	GFNet [103]	1.3	7.5	74.6	65.9	40.4	27.0	7.6	6.3
	RVT [81]	1.3	8.6	78.4	58.2	<u>43.7</u>	30.0	11.7	13.3
	DeiT*	1.3	5.7	74.7	67.0	34.5	21.7	11.9	9.4
	w/ VARS-S	1.0	6.8	73.7	69.8	36.8	24.8	10.8	4.9
	w/ VARS-D	1.4	5.4	75.6	64.9	39.6	27.5	13.7	10.2
	w/ VARS-SD	1.4	5.8	76.5	62.5	40.2	27.5	13.4	11.5
	GFNet-Ti*	1.3	7.5	74.6	65.9	40.4	27.0	7.6	6.3
	w/ VARS-S	1.3	7.5	74.1	63.5	40.8	28.6	9.5	5.8
	w/ VARS-D	1.9	9.8	77.8	58.6	41.2	29.0	15.9	12.6
	w/ VARS-SD	1.9	10.4	78.2	<u>57.4</u>	41.0	29.5	<u>16.2</u>	13.0
	RVT*	1.3	8.6	77.6	60.4	41.7	28.7	11.1	11.1
	w/ VARS-S	1.0	9.2	76.8	61.8	43.2	30.1	7.6	9.1
w/ VARS-D	1.2	8.0	78.2	58.7	42.0	29.8	11.7	12.4	
w/ VARS-SD	1.5	9.2	78.4	58.3	42.5	<u>30.5</u>	11.4	<u>13.4</u>	

Table 1.3: **Evaluation of domain generalization on PACS.** Our VARS-SD outperforms the baseline RVT* and other variants.

Target	Photo	Sketch	Cartoon	Art
RVT*	94.19	81.73	79.78	81.25
VARS-S	93.89	82.62	80.16	81.49
VARS-D	96.29	80.40	80.33	84.77
VARS-SD	96.47	82.78	80.98	86.08

Evaluation on Segmentation and Eye Fixation

Attention as coarse image segmentation. Recently, self-supervised vision transformers [10] have been shown to produce attention maps that are similar to the semantic segmentation of foreground objects. Following [10], we evaluate RVT* with self-attention and VARS on the validation set of PASCAL VOC 2012 using the model trained on ImageNet-1K. To

Table 1.4: **Segmentation evaluation on PASCAL VOC using attention maps.** Our VARS-SD improves the mean IOU score of the baseline RVT* and is more selective (higher FN scores).

	RVT*	VAR-S	VAR-D	VAR-SD
mIoU \uparrow	39.92	43.33	42.03	44.15
FP \downarrow	49.41	23.11	25.23	29.28
FN \downarrow	3.95	12.08	11.77	8.76

Table 1.5: **Evaluation on human eye fixations.** Here our VAR-S achieves the highest score while all variants outperforms RVT*.

Metric	RVT*	VAR-S	VAR-D	VAR-SD
NSS	0.502	0.737	0.632	0.678

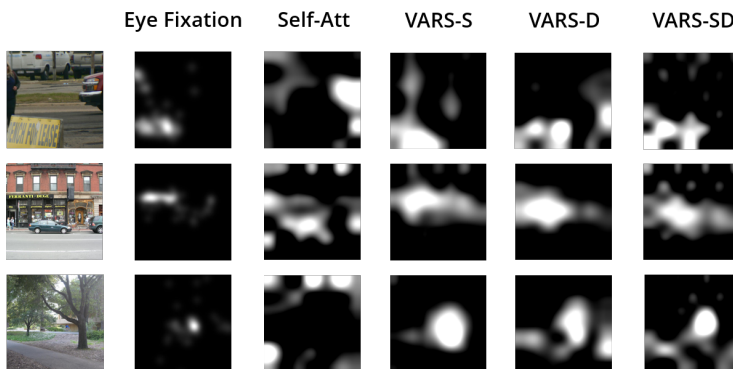


Figure 1.4: **Visualization on eye fixation.** VARS’ attention maps are more consistent with human eye fixation than self-attention’s.

obtain a segmentation map, we normalize an attention map from the global average of tokens to $[0, 1]$ and use a threshold 0.3 to distinguish foreground objects from the background (class agnostic). The main evaluation metric is mean IoU which evaluates the overlapping area between a predicted segmentation map and the ground truth. We also consider false positive (FP) and false negative (FN) rates as metrics.

Table 1.4 shows that all three variants of VARS achieve higher mean IoU compared to the self-attention counterpart RVT*, where VARS-SD improves the score from 39.92% to 44.15%. Also, the FP rate is substantially reduced by our attention framework, indicating that VARS can effectively filter out distracting information and preserve only the relevant information about the foreground objects. Another observation is that VARS has a higher FN rate, suggesting VARS is more selective than self-attention and emphasize more on the core parts of the objects.

Alignment with human eye fixations. Since human eye fixation is under the guidance

of bottom-up attention [147], here we investigate how close our attention maps are to the human eye fixation maps. Here we evaluate the ImageNet-pretrained RVT with self-attention and VARS on MIT1003 [59], containing 1K natural images with eye fixation maps collected from 15 human observers. We adopt the metric of normalized scanpath saliency (NSS) [97] that measures an average of normalized attention value at fixated positions.

Table 1.5 shows that RVT with VARS achieves higher NSS scores than RVT with self-attention (*i.e.*, RVT*), aligning better with the human eye fixations data. Figure 1.4 shows the attention maps captured from humans and generated by the models. We notice that VARS predicts regions that are more closely aligned with human attention, while self-attention tend to highlight irrelevant background regions.

Analysis and Ablation Study

Recurrent refinement of attention. VARS performs recurrent sparse reconstruction of the inputs in an iterative manner. In Figure 1.5, we visualize the attention maps VARS-S on ImageNet validation samples at different updating steps. We can see that VARS refines the attention maps through recurrent updates, *i.e.*, the attention maps become more focused on the core parts of the objects while suppressing the background and other distracting objects.

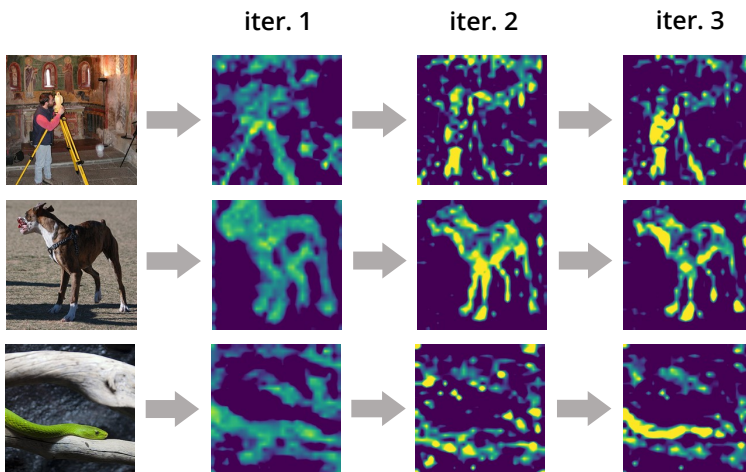


Figure 1.5: **Recurrent refinement of attention maps.** VARS refines the attention maps iteratively during the recurrent updates.

Number of recurrent updates. Figure 1.6 (left) shows the accuracy on ImageNet-C over different number of updates k . We find that the model has a similar performance between $k = 3$ and 5 with a drop of performance at $k = 1$ and 7. We choose $k = 3$ in our experiments for efficiency.

Strength of sparse constraints. Figure 1.6 (right) shows the accuracy over different λ values that determine the level of sparse regularization during the reconstruction of input. We observe that the curves are relatively flat which indicates VARS is not very sensitive to

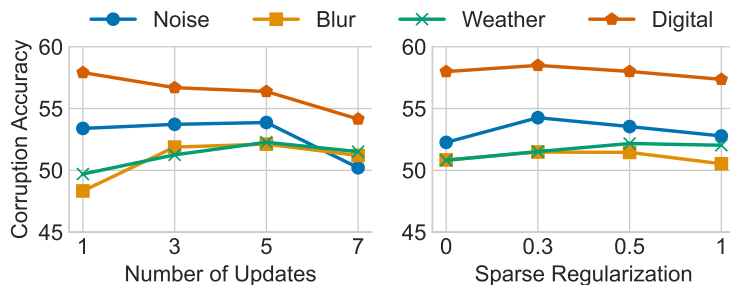


Figure 1.6: **Hyperparameter analysis.** We study the number of updates (left) and the level of sparse regularization (right) in sparse reconstruction. VARS performs similarly with 3 to 5 iteration steps and we choose 3 for better efficiency. VARS is not sensitive to the level of sparse regularization and we use 0.3 in the experiments.

the strength of the sparse regularization. We adopt $\lambda = 0.3$ in our experiments which has a slightly better performance than the other values.

1.5 Conclusion

We introduced a new attention formulation—Visual Attention from Recurrent Sparse reconstruction (VARS)—which takes inspiration from the robustness characteristics of human vision. We observed a connection among visual attention, recurrency, and sparsity and showed that contemporary attention models can be derived from recurrent sparse reconstruction of input signals. VARS adopts an ODE based formulation to describe neural dynamics; equilibrium states are solved by iteratively optimizing the sparse reconstruction of input. We showed that self-attention is a special case of VARS with approximate neural dynamics and no sparsity constraints. VARS is a general attention module that can be plugged into vision transformers, replacing the self-attention module, offering improved performance. We conducted extensive evaluation on five robustness benchmarks and three additional datasets of related settings to understand the properties of VARS. We found VARS increases model robustness with improved quality of attention maps across various datasets and settings.

Chapter 2

Top-Down Visual Attention from Analysis by Synthesis

2.1 Introduction

Human visual attention is often *task-guided*, *i.e.*, we tend to focus on different objects when processing different tasks [147, 11]. For example, when we answer different questions about one image, we only attend to the objects that are relevant to the question (Figure 2.1 (b-c)). This stands in contrast with the widely-used self-attention [30], which is completely *stimulus-driven*, *i.e.*, it highlights all the salient objects in the image without task-guided selection (Figure 2.1 (a)). While the stimulus-driven bottom-up attention has shown promising results in visual representation learning [10], current vision transformers still lack the ability of task-guided top-down attention, which provides task-adaptive representation and potentially improves task-specific performances [1, 142, 140]. Although some algorithms of top-down attention are proposed in the literature [95, 14, 1, 142, 140], they are incompatible with self-attention-based transformers and principled and unified designs are still missing.

Previous work [68, 15, 8, 102, 69] has studied the mechanism of top-down attention in human vision systems, hypothesizing top-down attention is a result of the human visual system performing Analysis by Synthesis (AbS). AbS [63, 146] is a classic idea that suggests the human visual perception depends on both the input image and a high-level prior about the latent cause of the image, and different priors can lead to different ways to perceive the same image (*e.g.*, visual illusion [67] and bistable perception [111]). This is formulated as Bayesian inference $\max_{\mathbf{z}} p(\mathbf{h}|\mathbf{z})p(\mathbf{z})$, where \mathbf{h} is the input image, and \mathbf{z} is the latent representation. It is hypothesized that the high-level goal can be formulated as a prior to direct the low-level recognition of different objects through AbS, achieving top-down attention. Still, existing works [145, 15, 86] are conceptual and hardly guide model designs in practice.

In this work, we present a novel perspective on how AbS entails top-down attention, followed by a new Analysis-by-Synthesis Vision Transformer (AbSViT) based on the findings. We start from previous work [113], which shows that visual attention (*e.g.*, self-attention)

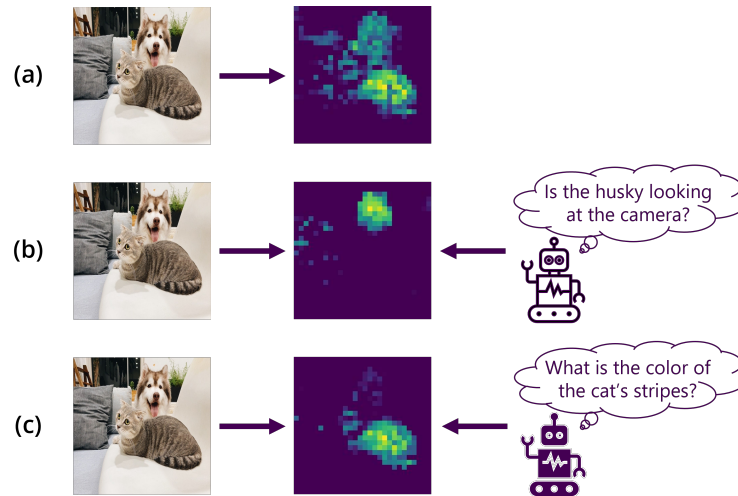


Figure 2.1: **Top-down vs. bottom-up attention.** (a) Bottom-up attention is stimulus-driven, *i.e.*, any salient objects (dog and cat) in the image may attract attention. (b-c) Top-down attention is task-guided. For example, when the task is to answer a question about a specific object, the attention will only center on that object and ignore the others. In this way, a more focused representation can be extracted for the current goal.

is functionally equivalent to sparse reconstruction which reconstructs the input using a dictionary containing templates of separate objects in the input. We show that AbS optimizes a similar *sparse reconstruction* objective modulated by a top-down signal. The top-down signal depends on the prior and acts as a preference on which object templates to choose to reconstruct the input. Therefore, only the objects consistent with the high-level prior are selected, equivalent to top-down attention.

Inspired by the connection, we propose AbSViT, a ViT [30] model with prior-conditioned top-down modulation trained to approximate AbS in a variational way. AbSViT contains a feedforward (encoding) and a feedback (decoding) pathway. The feedforward path is a regular ViT, and the feedback path contains linear decoders for each layer. Each inference starts with an initial feedforward run. The output tokens are manipulated by the prior and fed back through the decoders to each self-attention module as top-down input for the final feedforward pass (Figure 2.3).

When only pretrained on ImageNet [25], which contains mostly single-object images, AbSViT can attend to different objects in multi-object scenes controllably. For real-world applications, we observe consistent improvements from AbSViT on Vision-Language tasks such as VQA [3] and zero-shot image retrieval, where language is used as a prior to guide attention. For tasks without a strong prior, such as ImageNet classification and semantic segmentation, AbSViT can also serve as a general backbone and achieve substantial improvements. Additionally, the object-centric representation resulting from the top-down attention design enables better generalization to corrupted, adversarial, and out-of-distribution images. We hope this work can encourage future exploration of task-guided attention designs and

visual representation learning.

2.2 Related Work

Top-down visual attention endows us with the crucial ability to selectively collect information related to the behavioral goal. Several attempts have been made towards understanding the mechanism of top-down attention from experimental observations such as multiplicative tuning [84] and contrast responses [83, 105] in V4, and extra-classical receptive fields in V1 [2, 109, 12]. Other work tries to build a principled computational model for top-down attention [145, 15, 86].

Top-down attention has also found numerous applications in computer vision tasks where additional guidance (*e.g.*, language) is available aside from the image. Previous work employs top-down attention for object detection [92], image captioning [142], and visual question answering [140, 1]. However, these algorithms are either incompatible with current self-attention-based models or show inferior performance, as indicated by our experiments. Other work [78, 77, 143, 30] uses a feedforward model that takes both image and the high-level guidance (*e.g.*, text tokens or [cls] token) as input, which we show is suboptimal compared to our top-down model design. Dou *et al.* [31] propose to extract image and text features with separate encoders and combine them with a multi-modal fusion module during vision-language pretraining, which works better than using a single multi-modal feedforward model on vision language tasks. However, in this way, the visual encoder is still bottom-up. We show that augmenting it with the proposed top-down attention further improves model performance on standard benchmarks.

Top-down attention explained as Analysis by Synthesis. Analysis by Synthesis (AbS) is hypothesized as a potential computational model behind top-down attention. lee2002top starts from a Bayesian inference perspective and explains the top-down modulation in examples such as illusory contours and shapes from shading. Yu and Dayan [145] focus on the top-down attention in Ponser’s task [99] and build a hierarchical model where each layer corresponds to a computational step of Bayesian inference. Subsequent work [102, 15] assumes each object is generated by an appearance variable and a location variable and uses Bayesian inference to perform spatial attention and feature attention. Borji *et al.* [8] adopt a Dynamic Bayesian Network to simulate eye fixation in top-down attention. However, these models do not apply to practical designs in modern deep learning.

Generative model for discriminative learning. It has been widely explored in using generative models to assist discriminative learning. Specifically, the belief that representation with strong generative capability can better capture the structure of visual signals has inspired numerous unsupervised learning algorithms, from the early Restricted Boltzmann Machine [51, 52] and Helmholtz Machine [24], to the following auto-encoder models such as DAE [123] and VAE [62]. Recent work [45, 119] has shown impressive results on generative unsupervised learning. Generative models can also help with supervised learning, *e.g.*, by refining object

detection [72] or detecting errors in semantic segmentation [135]. Feedforward models with generative feedback are also more robust to input corruptions [56]. In our work, AbSViT also contains a generative feedback path that is able to refine the intermediate representation and attention and thus improves the performance.

2.3 Preliminaries: Attention as Sparse Reconstruction

Chapter 1 shows that a sparse reconstruction (SR) module functionally resembles visual attention. An SR module takes an input $\mathbf{x} \in \mathbb{R}^d$ and outputs $\mathbf{z} = \mathbf{P}\tilde{\mathbf{u}}^*$ where $\mathbf{P} \in \mathbb{R}^{d \times d'}$ is the dictionary and $\tilde{\mathbf{u}}^*$ is the sparse code, *i.e.*,

$$\tilde{\mathbf{u}}^* = \arg \min_{\tilde{\mathbf{u}} \in \mathbb{R}^{d'}} \frac{1}{2} \|\mathbf{P}\tilde{\mathbf{u}} - \mathbf{x}\|_2^2 + \lambda \|\tilde{\mathbf{u}}\|_1. \quad (2.1)$$

Each atom (column) of \mathbf{P} contains a template pattern and each element in $\tilde{\mathbf{u}}$ is the activation of the corresponding template. The objective is to reconstruct the input using as few templates as possible. To solve Equation (2.1), one may adopt a first-order optimization [108, 113] with dynamics at time t of

$$\frac{d\mathbf{u}}{dt} \propto -\mathbf{u} - (\mathbf{P}^T\mathbf{P} - \mathbf{I})\tilde{\mathbf{u}} + \mathbf{P}^T\mathbf{x}, \quad (2.2)$$

where the optimization is over an auxiliary variable \mathbf{u} and $\tilde{\mathbf{u}} = g_\lambda(\mathbf{u}) = \text{sgn}(\mathbf{u})(|\mathbf{u}| - \lambda)_+$ with $\text{sgn}(\cdot)$ as the sign function and $(\cdot)_+$ as ReLU. Here \mathbf{u} is activated by the template matching $\mathbf{P}^T\mathbf{x}$ between the dictionary and the input, and different elements in \mathbf{u} inhibit each other through $-(\mathbf{P}^T\mathbf{P} - \mathbf{I})\tilde{\mathbf{u}}$ to promote sparsity.

To see the connection between visual attention and sparse reconstruction, recall that attention in the human visual system is achieved via two steps [26]: (i) *grouping* features into separate objects or regions, and (ii) *selecting* the most salient objects or regions while repressing the distracting ones. A similar process is also happening in SR, *i.e.*, if each atom in \mathbf{P} is a template of every single object, then each element in \mathbf{u} groups the input features belonging to that object through $\mathbf{P}^T\mathbf{x}$, while the sparsity constraint promoted by the lateral inhibition $-(\mathbf{P}^T\mathbf{P} - \mathbf{I})\tilde{\mathbf{u}}$ selects the object that is most activated. As shown in [113], SR modules achieve similar attention effects as self-attention (SA) [121] while being more robust against image corruptions.

Interestingly, it is also pointed out in [113] that under certain constraints (*e.g.*, the key and query transform is the same), SA can be viewed as solving a similar SR problem but without sparsity. After adding the sparsity back, SA is an approximation of

$$\begin{cases} \tilde{\mathbf{U}}^* = \arg \min_{\tilde{\mathbf{U}}} \frac{1}{2} \|\Phi(\mathbf{K})\tilde{\mathbf{U}} - \mathbf{V}\|_2^2 + \lambda \|\tilde{\mathbf{U}}\|_1, & (2.3) \\ \mathbf{Z} = \Phi(\mathbf{Q})\tilde{\mathbf{U}}^*, & (2.4) \end{cases}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{(hw) \times c}$ are the query, key, and value matrices, $\Phi(\mathbf{Q}), \Phi(\mathbf{K}) \in \mathbb{R}^{(hw) \times d'}$ are the random features [16] that approximate the softmax kernel $\Phi(\mathbf{Q})_i \Phi(\mathbf{K})_j^T \approx e^{\mathbf{Q}_i \mathbf{K}_j^T}$,

$\tilde{\mathbf{U}}^* \in \mathbb{R}^{d' \times c}$ is the sparse code and \mathbf{Z} is the output. This provides a novel perspective on the mechanism of SA, *i.e.*, it is solving a channel-wise sparse reconstruction of the value matrix \mathbf{V} using an input-dependent dictionary $\Phi(\mathbf{K})$. Visualization of $\Phi(\mathbf{K})$ shows each atom contains a mask for one single object or region, which means that SA is trying to reconstruct the input with as few masks as possible, thus only the salient objects are selected and highlighted (Figure 2.2 (a)).

2.4 Top-Down Attention from AbS

We consider top-down visual attention from an Analysis by Synthesis (AbS) view of vision. We start from the hierarchical AbS formulation of visual perception (Section 2.4) and show that it is equivalently optimizing a sparse reconstruction objective that is modulated by a top-down signal, thus entailing top-down attention (Section 2.4).

Hierarchical AbS

AbS formulates visual perception as a Bayesian inference process. Given the image generation process $p(\mathbf{h}|\mathbf{z})$ and a prior $p(\mathbf{z})$, where \mathbf{h} is the image and \mathbf{z} is the latent code, AbS finds $\mathbf{z}^* = \arg \max_{\mathbf{z}} p(\mathbf{h}|\mathbf{z})p(\mathbf{z})$.

In this work, we assume the generation is hierarchical, *i.e.*, $\mathbf{z}_L \rightarrow \mathbf{z}_{L-1} \rightarrow \dots \rightarrow \mathbf{z}_1 \rightarrow \mathbf{h}$, where \mathbf{z}_ℓ is the latent at ℓ -th layer. The MAP estimation is

$$\mathbf{z}_L^*, \dots, \mathbf{z}_1^* = \arg \max_{\mathbf{z}_L, \dots, \mathbf{z}_1} p(\mathbf{h}|\mathbf{z}_1) \cdots p(\mathbf{z}_{L-1}|\mathbf{z}_L)p(\mathbf{z}_L). \quad (2.5)$$

For each generation process $\mathbf{z}_{\ell+1} \rightarrow \mathbf{z}_\ell$ between layer ℓ and $\ell + 1$, we further assume that \mathbf{z}_ℓ is constructed by a sparse code $\tilde{\mathbf{u}}_\ell$ which is generated from $\mathbf{z}_{\ell+1}$ via a non-linear function $g_\ell(\cdot)$, *i.e.*,

$$\tilde{\mathbf{u}}_\ell \sim p(\tilde{\mathbf{u}}_\ell|\mathbf{z}_{\ell+1}) \propto \exp\left\{-\frac{1}{2}\|\mathbf{P}_\ell \tilde{\mathbf{u}}_\ell - g_\ell(\mathbf{z}_{\ell+1})\|_2^2 - \lambda\|\tilde{\mathbf{u}}_\ell\|_1\right\} \quad (2.6)$$

$$\mathbf{z}_\ell = \mathbf{P}_\ell \tilde{\mathbf{u}}_\ell, \quad (2.7)$$

where \mathbf{P}_ℓ is the dictionary. Intuitively, it first generates $g_\ell(\mathbf{z}_{\ell+1})$ as a blurry and noisy version of \mathbf{z}_ℓ , then find the sparse code $\tilde{\mathbf{u}}_\ell$ to construct a sharper and cleaner version.

Since \mathbf{z}_ℓ is decided by $\tilde{\mathbf{u}}_\ell$, it suffices to optimize the MAP estimation over $\{\tilde{\mathbf{u}}_\ell\}_{\ell=1}^L$, *i.e.*,

$$\tilde{\mathbf{u}}_L^*, \dots, \tilde{\mathbf{u}}_1^* = \arg \max_{\tilde{\mathbf{u}}_L, \dots, \tilde{\mathbf{u}}_1} p(\mathbf{h}|\tilde{\mathbf{u}}_1) \cdots p(\tilde{\mathbf{u}}_{L-1}|\tilde{\mathbf{u}}_L)p(\tilde{\mathbf{u}}_L). \quad (2.8)$$

Solving Equation (2.8) by simple gradient ascent (of the logarithm) gives the dynamics

$$\frac{d\tilde{\mathbf{u}}_\ell}{dt} \propto \nabla_{\tilde{\mathbf{u}}_\ell} \log p(\tilde{\mathbf{u}}_{\ell-1}|\tilde{\mathbf{u}}_\ell) + \nabla_{\tilde{\mathbf{u}}_\ell} \log p(\tilde{\mathbf{u}}_\ell|\tilde{\mathbf{u}}_{\ell+1}) \quad (2.9)$$

where $\tilde{\mathbf{u}}_\ell$ is affected by both $\tilde{\mathbf{u}}_{\ell-1}$ and $\tilde{\mathbf{u}}_{\ell+1}$.

Top-Down Attention from AbS

From AbS (Eq. (2.6-2.9)) we can derive the dynamics of $\tilde{\mathbf{u}}_\ell$ as

$$\frac{d\tilde{\mathbf{u}}_\ell}{dt} \propto \nabla_{\tilde{\mathbf{u}}_\ell} \left(-\frac{1}{2} \|\mathbf{P}_\ell \tilde{\mathbf{u}}_\ell - (\mathbf{x}_\ell^{bu} + \mathbf{x}_\ell^{td})\|_2^2 - \lambda \|\tilde{\mathbf{u}}_\ell\|_1 - r_\ell(\tilde{\mathbf{u}}_\ell) \right) \quad (2.10)$$

where $\mathbf{x}_\ell^{td} = g_\ell(\mathbf{z}_{\ell+1})$ is the top-down signal and $\mathbf{x}_\ell^{bu} = f_\ell(\mathbf{z}_{\ell-1}) = \mathbf{J}_{g_{\ell-1}}^T \mathbf{z}_{\ell-1}$ is the bottom-up signal where $\mathbf{J}_{g_{\ell-1}}$ is the jacobian of $g_{\ell-1}(\mathbf{P}_\ell \tilde{\mathbf{u}}_\ell)$, and $r_\ell(\tilde{\mathbf{u}}_\ell) = \|g_{\ell-1}(\mathbf{P}_\ell \tilde{\mathbf{u}}_\ell)\|_2^2$ is an additional regularization. Details of the derivation are pushed back to Appendix. One may notice from Equation (2.10) that, in AbS each layer is solving a similar sparse reconstruction problem as in Equation (2.1) but with the input of $\mathbf{x}_\ell^{bu} + \mathbf{x}_\ell^{td}$, thus simulating attention that is modulated by both bottom-up and top-down signals. This can also be observed by turning Equation (2.10) into

$$\frac{d\tilde{\mathbf{u}}_\ell}{dt} \propto -\mathbf{u}_\ell - (\mathbf{P}_\ell^T \mathbf{P}_\ell - \mathbf{I}) \tilde{\mathbf{u}}_\ell + \mathbf{P}_\ell^T \mathbf{x}_\ell^{bu} + \mathbf{P}_\ell^T \mathbf{x}_\ell^{td} - \nabla r_\ell(\tilde{\mathbf{u}}_\ell). \quad (2.11)$$

Comparing with Equation (2.2), here $\tilde{\mathbf{u}}_\ell$ is steered by an additional term $\mathbf{P}_\ell^T \mathbf{x}_\ell^{td}$ that acts as a bias on which atom in \mathbf{P}_ℓ to choose. For example, if atoms in \mathbf{P}_ℓ are templates of separate objects (like in self-attention), then $\mathbf{P}_\ell^T \mathbf{x}_\ell^{td}$ highlights the objects that are consistent with the top-down signal (Figure 2.2 (b)).

This implies an AbS system naturally entails top-down attention. Intuitively, the prior reflects which objects the output \mathbf{z}_L should highlight. Then the affected \mathbf{z}_L is fed back to layer $L-1$ through g_{L-1} , as a top-down signal to direct which objects to select in layer $L-1$. The same process repeats until the first layer. Different priors will direct the intermediate layers to select different objects, achieving top-down attention.

Interestingly, if we consider the analogy between self-attention and sparse reconstruction, Equation (2.10) leads to a smooth way of building a top-down version of self-attention, *i.e.*, we only need to add a top-down signal to the value \mathbf{V} , while keeping other parts such as \mathbf{Q} and \mathbf{K} (which decides the dictionary) untouched. We will make it clearer in Section 2.5.

2.5 Analysis-by-Synthesis Vision Transformer

Inspired by the connection between top-down attention and AbS, we propose to achieve top-down attention by building a vision transformer that performs AbS (Equation (2.5)), *i.e.*, if the network has input \mathbf{h} and latent representation \mathbf{z}_ℓ after each layer ℓ (which means \mathbf{z}_L is the output), the final latent representation should approximate $\mathbf{z}_1^*, \dots, \mathbf{z}_L^*$. Since directly solving Equation (2.5) requires an iterative optimization which would be extremely costly, in this work, we adopt a variational approximation to Equation (2.5). Specifically, we optimize a variational loss

$$\begin{aligned} \mathcal{L}_{var} &= - \sum_{\ell=0}^{L-1} \log p(\mathbf{z}_\ell | \mathbf{z}_{\ell+1}) - \log p(\mathbf{z}_L) \\ &= \sum_{\ell=0}^{L-1} \left(\frac{1}{2} \|\mathbf{P}_\ell \tilde{\mathbf{u}}_\ell - g_\ell(\mathbf{z}_{\ell+1})\|_2^2 + \lambda \|\tilde{\mathbf{u}}_\ell\|_1 \right) - \log p(\mathbf{z}_L) \end{aligned} \quad (2.12)$$

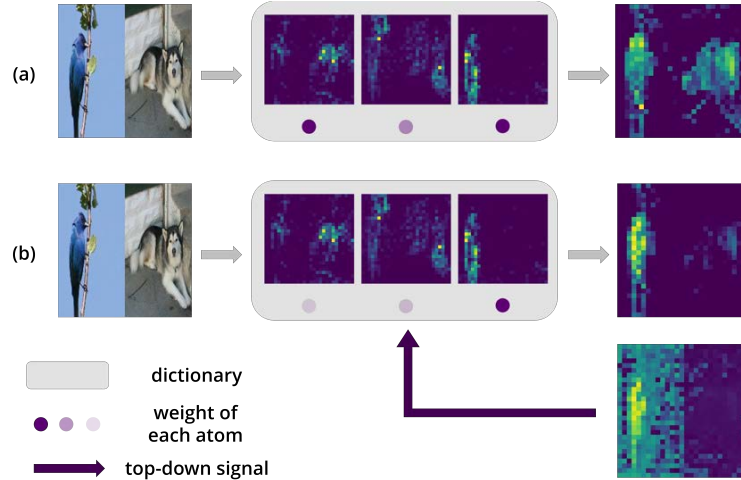


Figure 2.2: (a) Each atom in the dictionary contains masks for separate objects or regions. The sparse reconstruction tries to use as few masks as possible to reconstruct the input feature map, thus only the salient objects are highlighted. (b) The top-down signal \mathbf{x}_ℓ^{td} puts a bias on the weights of the atoms so that only the objects that agree with \mathbf{x}_ℓ^{td} are selected.

where $\mathbf{z}_0 = \mathbf{h}$. However, as stated below, there are several caveats we need to work around when training a network with Equation (2.12) in real-world tasks.

The sparsity regularization. Since the practical model we build in this work is based on self-attention (Section 2.5), which neither has a sparsity constraint nor solves the SR explicitly [113], we remove the sparsity regularization by setting $\lambda = 0$, which makes $-\log p(\mathbf{z}_\ell | \mathbf{z}_{\ell+1}) = \frac{1}{2} \|\mathbf{P}_\ell \tilde{\mathbf{u}}_\ell - g_\ell(\mathbf{z}_{\ell+1})\|_2^2 = \frac{1}{2} \|\mathbf{z}_\ell - g_\ell(\mathbf{z}_{\ell+1})\|_2^2$.

Jointly training the decoder g_ℓ . Normally, optimizing Equation (2.12) requires knowing the generation process g_ℓ beforehand, which in our case is unknown. This can be addressed by training g_ℓ jointly with the whole network, similar to VAE [62]. It is natural to use g_ℓ also as the feedback path of the network, as shown in Section 2.5.

Trade-off between the generative and discriminative power. The variational loss forces each $\mathbf{z}_{\ell+1}$ to be capable of generating \mathbf{z}_ℓ . However, we find empirically that enforcing a strong generative power on the feature will harm its discriminative power in the setting of supervised learning. To address this, for each term $-\log p(\mathbf{z}_\ell | \mathbf{z}_{\ell+1})$ we stop the gradient on \mathbf{z}_ℓ and $\mathbf{z}_{\ell+1}$, *i.e.*, $-\log p(\mathbf{z}_\ell | \mathbf{z}_{\ell+1}) = \frac{1}{2} \|sg(\mathbf{z}_\ell) - g_\ell(sg(\mathbf{z}_{\ell+1}))\|_2^2$, where $sg(\cdot)$ is stop-gradient. In this way, only the decoder g_ℓ receives the gradient.

The variable prior. Rigorously speaking, variational methods only approximate AbS with a fixed prior $p(\mathbf{z}_L)$. However, top-down attention should be able to flexibly attend to different objects by changing different priors. The question is, how can we learn a variational model that generalizes to different priors? In this work, we adopt a simple trick called Meta-amortized VI [130]. Concretely, we assume the prior $p_\xi(\mathbf{z}_L)$ depends on some parameter ξ , which can be a sentence or a class prototype cueing what objects to look at in the image.

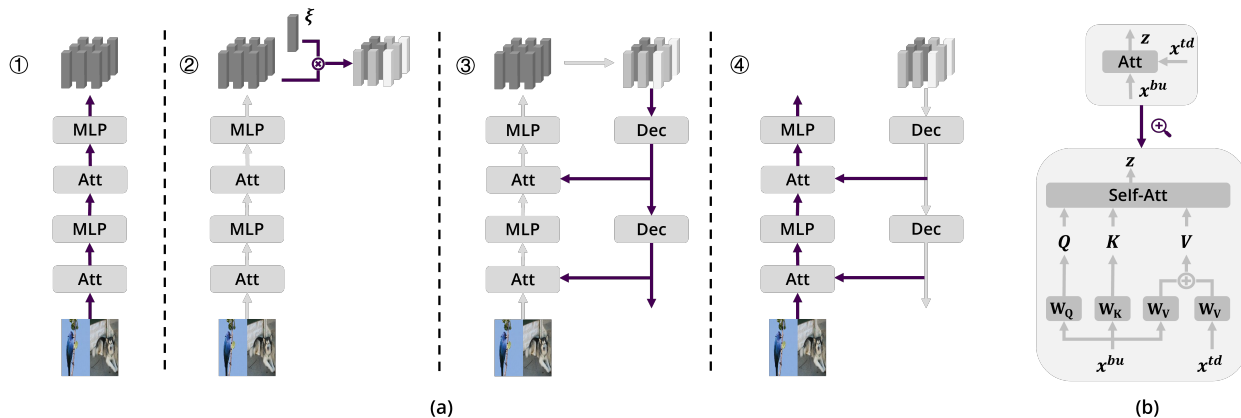


Figure 2.3: **Design of AbSViT.** (a) Four steps to every single inference. The operations in each step are colored as purple and others as gray. AbSViT first passes the image through the feedforward path. The output tokens are then reweighted by their similarity with the prior vector ξ and fed back through the decoders to each self-attention module as the top-down input for the final feedforward run. (b) The top-down input to self-attention is added to the value matrix while other parts stay the same.

Then we make the model adaptable to ξ during inference to approximate AbS with prior $p_\xi(\mathbf{z}_L)$ given any ξ . See the design details in Section 2.5.

After applying these tricks, our variational loss becomes

$$\mathcal{L}_{var} = \frac{1}{2} \sum_{\ell=0}^{L-1} \|sg(\mathbf{z}_\ell) - g_\ell(sg(\mathbf{z}_{\ell+1}))\|_2^2 - \log p_\xi(\mathbf{z}_L), \quad (2.13)$$

which contains layer-wise reconstruction loss and a prior loss. We also try cosine similarity instead of ℓ_2 distance for reconstruction and get similar results. In Section 2.5, we will show how to build a ViT with prior-conditioned top-down modulation and train it with Equation (2.13).

AbSViT Design

Figure 2.3 (a) shows the proposed AbSViT which is built upon ViT [30]. Every single inference consists of 4 steps: (i) pass the image through the feedforward encoder, (ii) modulate the output tokens with a prior vector ξ , (iii) send the tokens back through the feedback decoder to intermediate layers, and (iv) run the feedforward path again but with each self-attention layer also receiving the top-down tokens as input.

Within the whole pipeline, the feedforward encoder has the same architecture as regular ViT. For the feedback path, we use a single token-wise linear transform for each layer-wise decoder g_ℓ . The design of token modulation with prior ξ and the self-attention with top-down input are introduced below:

Design of token modulation with ξ . The purpose is to modify the tokens to carry the information about the prior p_ξ when fed back to the network. The prior is parameterized by ξ , which may be a language embedding or a class prototype telling the network which objects to look at. Therefore, we instantiate the modulation as a simple spatial reweighting, *i.e.*, $\mathbf{z}_L^i \rightarrow \alpha \cdot \text{sim}(\xi, \mathbf{z}_L^i) \cdot \mathbf{z}_L^i$, where \mathbf{z}_L^i is the i -th output token, sim is the cosine similarity clamped to $[0, 1]$, and α is a scaling factor controlling the scale of the top-down signal, which is set to 1 by default. In this way, only the tokens with high similarity to ξ are sent back, and others are (softly) masked out. Note that the design here is for simplicity and may not be suitable for general usage. For example, when dealing with transparent images where two objects overlap, spatial reweighting cannot separate two objects away.

Design of self-attention with top-down input. From the analogy between self-attention and sparse reconstruction (Equation (2.3)), the value matrix in SA corresponds to the reconstructed input signal, and the query and key serve as the dictionary. Since the top-down attention in AbS (Equation (2.10)) adds a top-down signal to the input while keeping the dictionary untouched, it is natural to design the top-down version of self-attention by simply adding the top-down signal to the value and keep query and key as the same, as illustrated in Figure 2.3 (b). We will show in Section 2.6 that this is better than an arbitrary design where we add the top-down signal to the query, key, and value.

In this paper, we focus on supervised learning and train the model on two types of tasks. One is Vision-Language (V&L) tasks such as VQA and zero-shot image retrieval, where the language acts as a prior to cue the model where to look at. The other one is image understanding, such as ImageNet classification and semantic segmentation, which do not have a specific prior. When training the network, we optimize the supervised loss as well as the variational loss (Equation (2.13)), *i.e.*,

$$\mathcal{L} = \frac{1}{2} \sum_{\ell=1}^L \|sg(\mathbf{z}_\ell) - g_\ell(sg(\mathbf{z}_{\ell+1}))\|_2^2 - \log p_\xi(\mathbf{z}_L) + \mathcal{L}_{sup}, \quad (2.14)$$

where \mathbf{z}_ℓ is the ℓ -th layer’s output after the whole inference cycle, sg is stop-gradient, and g_ℓ is the ℓ -th layer’s decoder. The form of prior p_ξ depends on the task. For V&L tasks, ξ is the text embedding and we use a CLIP-style prior [100]:

$$p_\xi(\mathbf{z}_L) = \frac{\exp\{\xi^T \mathbf{z}_L\}}{\exp\{\xi^T \mathbf{z}_L\} + \sum_k \exp\{\xi^T \mathbf{z}_-^k\}}, \quad (2.15)$$

where the negative samples \mathbf{z}_-^k are the output from other images. For image classification and segmentation where no specific prior is available, we set ξ as a trainable query vector that is independent of the input image, and we choose an uninformative prior that does not contribute to the gradient, *i.e.*, $\nabla \log p_\xi(\mathbf{z}_L) = 0$.

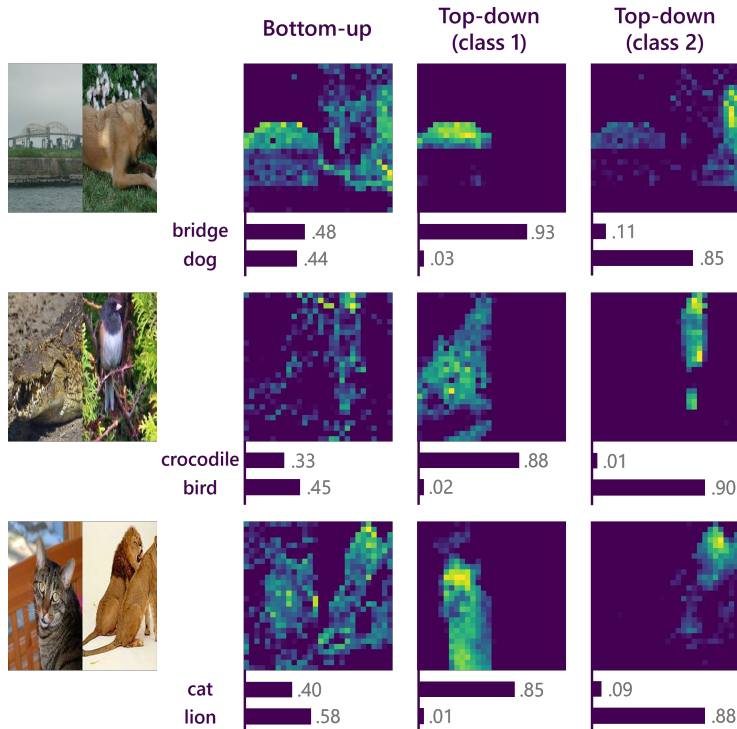


Figure 2.4: Controllable top-down attention in multi-object images. For each image, bottom-up attention will highlight both objects. In contrast, we can use different class prototypes as the prior to control the top-down attention to focus on different objects, and the classification result also changes accordingly.

2.6 Experiments

In this section, we first show that AbSViT achieves controllable top-down attention in multi-object scenes (Section 2.6). Then we test AbSViT on Vision-Language tasks such as VQA and zero-shot image retrieval (Section 2.6), and also on ImageNet classification and model robustness (Section 2.6), as well as semantic segmentation (Section 2.6). Finally, we analyze specific designs of AbSViT in Section 2.6.

Datasets. For VQA, we use VQAv2 [39] for training and testing and compare the attention map with human attention collected by VQA-HAT [22]. For zero-shot image retrieval, we use Flickr30K [98]. For image classification, we train and test on ImageNet-1K (IN) [25], and also test on corrupted images from IN-C [47], adversarial images from IN-A [48], and out-of-distribution images from IN-R [49] and IN-SK [125]. For semantic segmentation, we test on PASCAL VOC [32], Cityscapes [19], and ADE20K [148].

Experimental setup. We compare several baselines for goal-directed attention: (i) **PerceiverIO** [58] uses $e_{\xi}(\cdot)$ to reweight the tokens from feedforward output just like in AbSViT, but directly outputs the reweighted tokens without any feedback, (ii) **MaskAtt** uses the same

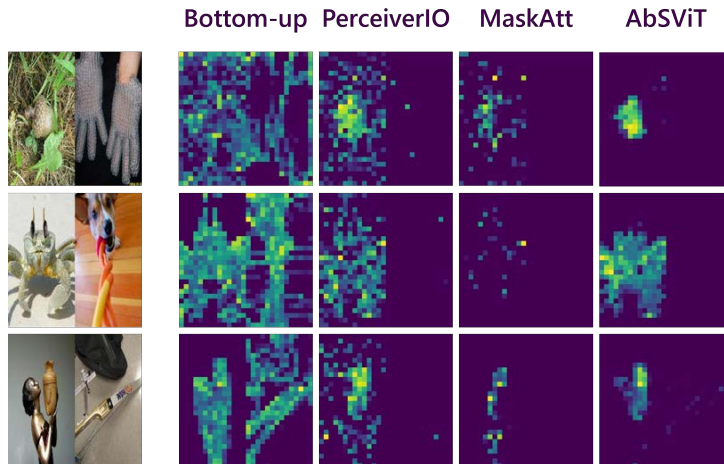


Figure 2.5: Comparison between different top-down attention algorithms. Prior corresponds to the left image. AbSViT has cleaner attention map than other baselines.

soft mask for reweighting the output tokens to reweight the value tokens in intermediate self-attention modules, instead of adding the top-down tokens on them, (iii) **Feedback** directly feeds back the output tokens without reweighting. For V&L tasks, we use the METER [31] framework, which contains a vision backbone, a language backbone, and a multimodal fusion module. We use ViT [30] as the vision backbone and replace it with AbSViT or the baseline models. For image classification, we try the backbones of ViT, RVT [82], and FAN [149], which is state of the art on ImageNet and robustness benchmarks. The scaling factor α is set as 1 during ImageNet pretraining and evaluation and set as 10 for finetuning on V&L tasks because we find AbSViT pretrained on supervised single-object classification only learns weak top-down attention in multi-object scenes (Section 2.7). See the Appendix for additional implementation details.

Controllable Top-Down Attention of AbSViT

To test the top-down attention in multi-object images, we take a AbSViT pretrained on ImageNet (Section 2.6) and create multi-object images by randomly sampling two images from ImageNet and concatenating them side by side. To control the top-down attention, we use the class prototype (from the last linear layer) of the two classes as ξ . Since in regular ViT, the class prototypes only align with the [cls] token but not with other output tokens, here we use a ViT with global average pooling. We set $\alpha = 10$.

To compare the bottom-up and top-down attention, we visualize the norm of output tokens from ViT and AbSViT for each class. As shown in Figure 2.4, bottom-up attention highlights both objects while only the target object is selected by top-down attention. Consequently, the classification result, which has a tie between two classes when no prior is available, is biased towards the target class when we turn on the prior. This indicates AbSViT has the

Model	VQAv2		Flickr-Zero-Shot		
	test-dev	test-std	IR@1	IR@5	IR@10
BEiT-B-16 [5]	68.45	-	32.24	-	-
CLIP-B-32 [100]	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

Table 2.1: Comparison of different top-down attention algorithms on VQA and zero-shot image retrieval. AbSViT achieves consistent improvements on both tasks.

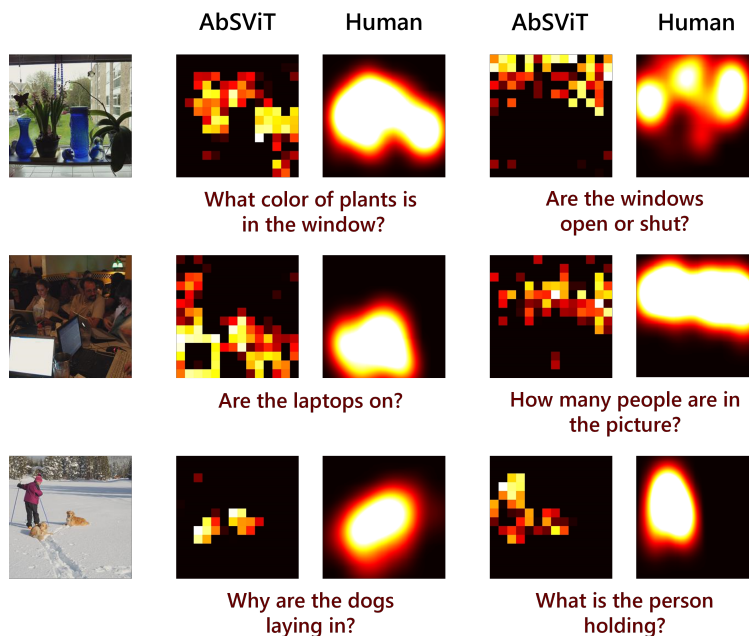


Figure 2.6: Comparison of attention map from AbSViT and human attention on VQA. AbSViT’s attention is adjustable to different questions and is consistent with human attention.

ability to control its attention on different objects given different priors. We also compare the top-down attention of AbSViT with several baselines (Figure 2.5). We can see that the attention of PerceiverIO focuses coarsely on the target object but is noisy, possibly because it lacks a feedback mechanism. MaskAtt, on the other hand, tends to miss parts of the object, implying that masking attention is less suitable for ViTs.

AbSViT for Vision-Language Tasks

We test AbSViT on two V&L tasks, VQA, and zero-shot image retrieval. We use the METER framework and replace the vision backbone with ViT-B, AbSViT-B, and other baselines. All the vision backbones are pretrained on ImageNet (Section 2.6). Results are shown in Table 2.1.

On VQAv2, AbSViT surpasses the baselines on both test splits and reaches the same performance as the unsupervised model (BEiT-B). At the same time, PerceiverIO has no improvement over ViT, probably because the multimodal fusion in METER can already perform token reweighting. The pure feedback network helps a little, mainly due to the feature refinement during the feedback loop. It is worth noticing that MaskAtt, a strategy frequently used in previous work, actually hurts performance when added to the vision transformer. On zero-shot image retrieval, AbSViT also has higher performance than all other baselines. Especially, it has an improvement of $\sim 3\%$ over bottom-up ViT on IR@1.

We also visualize the attention map of AbSViT on VQA and compare it to human attention. As shown in Figure 2.6, AbSViT can adjust its attention to the objects related to the question. The attention map is also consistent with human attention. Nevertheless, the attention map of AbSViT is still not precise enough. For example, in the last example, when the question is “What is the person holding?”, the top-down attention highlights both the person and the dogs. Since the model is only pretrained on ImageNet, it may be further improved by CLIP [100] pretraining.

Image Classification and Robustness

We test AbSViT on ImageNet classification and robustness benchmarks (Table 2.2). We report mCE (lower the better) [47] for IN-C and accuracy for other datasets. On clean images, AbSViT consistently improves over baselines, with a similar number of parameters although higher FLOPs. The clean accuracy on FAN-B is improved to 83.7%, reaching the same level as ConvNext-B with fewer parameters. On corrupted (IN-C) and adversarial (IN-A) images, AbSViT boosts the performance by about 1-5% across all the scales. Especially, the performance on FAN-B is raised by 1% and 5% for IN-C and IN-A, reaching a new state-of-the-art result. On out-of-distribution images, AbSViT also improves by 3% on Tiny and Small models and 0.5% on FAN-B.

Figure 2.7 visualizes the attention map of ViT and AbSViT, as well as token weights generated in $e_{\xi}(\cdot)$. The bottom-up attention in ViT is often noisy and only partly detects the foreground object. On the other hand, the query ξ in AbSViT learns to coarsely detect the foreground and reweight the feedforward output tokens, which are fed back and generate top-down attention that better detects the foreground object.

We compare AbSViT with several baseline algorithms for goal-directed attention in Table 2.3. One may see that a pure feedback model already improves the clean accuracy and robustness, and AbSViT further boosts the performance by better extracting the foreground object. Due to a similar reason, PerceiverIO without feedback also slightly improves the

Model	P/F	Clean	IN-C (↓)	IN-A	IN-SK	IN-R
PiT-Ti [50]	5/0.7	72.9	69.1	6.2	34.6	21.6
ConViT-Ti [20]	6/1.4	73.3	68.4	8.9	35.2	22.4
PVT-Ti [126]	13/1.9	75.0	79.6	7.9	33.9	21.5
GfNet-Ti [103]	8/1.3	74.6	65.9	6.3	40.4	27.0
ViT-Ti [30]	6/1.3	72.5	71.1	7.5	33.0	20.1
Gray - AbS	7/2.6	74.1	66.7	10.1	34.9	22.6
RVT-Ti [82]	9/1.3	78.1	58.8	13.9	42.5	29.1
Gray - AbS	11/2.7	78.6	55.9	17.3	43.2	29.9
FAN-Ti [149]	7/1.3	77.5	59.8	13.1	42.6	29.9
Gray - AbS	9/2.9	78.3	57.4	16.5	42.8	31.2
PiT-S [50]	24/2.9	80.9	52.5	21.7	43.6	30.8
PVT-S [126]	25/3.8	79.9	66.9	18.0	40.1	27.2
Swin-T [75]	28/4.5	81.2	62.0	21.6	41.3	29.1
ConvNext-T [76]	29/4.5	82.1	53.2	24.2	47.2	33.8
ViT-S [30]	22/4.2	80.1	54.6	19.2	41.9	28.9
Gray - AbS	26/9.8	80.7	51.6	24.3	43.1	30.2
RVT-S [82]	22/4.3	81.9	50.5	26.0	47.0	34.5
Gray - AbS	26/10.4	81.9	48.7	31.1	48.5	35.6
FAN-S [149]	28/5.3	82.8	49.1	29.3	47.4	35.6
Gray - AbS	32/11.4	83.0	47.4	34.0	48.3	36.4
PiT-B [50]	74/12.5	82.4	48.2	33.9	43.7	32.3
PVT-L [126]	61/9.8	81.7	59.8	26.6	42.7	30.2
Swin-B [75]	88/15.4	83.4	54.4	35.8	46.6	32.4
ConvNext-B [76]	89/15.4	83.8	46.8	36.7	51.3	38.2
ViT-B [30]	87/17.2	80.8	49.3	25.2	43.3	31.6
Gray - AbS	99/38.9	81.0	48.3	28.2	42.9	31.7
RVT-B [82]	86/17.7	80.9	52.1	26.6	39.6	26.1
Gray - AbS	100/39.5	80.9	51.7	28.5	39.3	26.0
FAN-B [149]	54/10.4	83.5	45.0	33.2	51.4	39.3
Gray - AbS	62/21.8	83.7	44.1	38.4	52.0	39.8

Table 2.2: Results on ImageNet classification and robustness benchmarks. AbSViT improves performance across different benchmarks and backbones. P/F: # of parameters and FLOPs. ↓: lower is better.

performance. On the other hand, MaskAtt is sometimes harmful (on Clean, IN-C, and IN-A for RVT), implying that a mask attention design is unsuitable for vision transformers.

Semantic Segmentation

We evaluate the performance of AbSViT as a backbone for semantic segmentation on three datasets (PASCAL VOC, Cityscapes, and ADE20K). We compare with two baseline backbones, regular ViT and ResNet-101. We use UperNet [138] as the segmentation head for all the backbones. Results are shown in Table 2.4. We can see that when using AbSViT as the backbone, we can achieve 1.2-2.0% improvements over the ViT baseline with approximately

Model	Clean	IN-C (\downarrow)	IN-A	IN-R	IN-SK
ViT-Ti	72.5	71.1	7.5	33.0	20.1
- PerceiverIO	72.8	70.4	8.0	32.8	20.5
- Feedback	73.4	67.8	9.7	34.6	22.4
- MaskAtt	72.5	70.6	8.3	33.4	20.5
- AbS	74.1	66.7	10.1	34.9	22.6
RVT-Ti	78.1	58.8	13.9	42.5	29.1
- PerceiverIO	78.3	57.8	13.7	42.8	29.8
- Feedback	79.1	55.7	18.2	44.1	31.3
- MaskAtt	77.9	59.0	13.5	43.0	29.7
- AbS	79.5	54.8	18.7	44.5	32.5

Table 2.3: Comparison of different top-down attention algorithms on ImageNet classification and robustness.

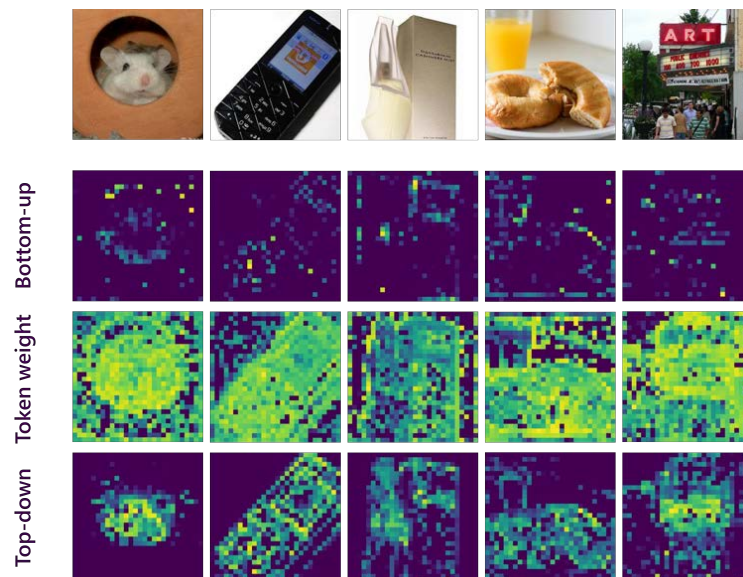


Figure 2.7: Visualization of the bottom-up attention, token weights, and the top-down attention in AbSViT. The bottom-up attention is noisy and fails to detect the complete foreground object. In AbSViT, the query mask can coarsely detect the foreground object and reweight tokens fed back to direct the top-down attention to better extract the foreground object.

the same number of parameters. This indicates that AbSViT can be used as a general backbone for different vision tasks.

Model	PASCAL VOC	Cityscapes	ADE20K
ResNet-101 [18]	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)

Table 2.4: Semantic segmentation results on three datasets.

Model	Clean	IN-C (\downarrow)	IN-A	IN-R	IN-SK
AbSViT-QKV	73.3	68.0	9.4	33.8	21.2
AbSViT	74.1	66.7	10.1	34.9	22.6

Table 2.5: The predicted design of top-down self-attention (AbSViT) is better than an arbitrary design (AbSViT-QKV).

	\mathcal{L}_{var}	Clean	IN-C (\downarrow)	IN-A	IN-R	IN-SK
AbSViT	\times	73.1	69.0	9.5	33.5	20.8
AbSViT	\checkmark	74.1	66.7	10.1	34.9	22.6

Table 2.6: Ablation on the variational loss \mathcal{L}_{var} .

Justification of Model Design

The design of AbSViT follows the principle of AbS. For example, AbSViT adds the top-down signal only to the value matrix considering the analogy between self-attention and sparse reconstruction (Section 2.5). At the same time, an arbitrary design may also add it to the query and key. We also optimize the variational loss to approximate AbS instead of just building a top-down model and training with the supervised loss. In this section, we show the advantage of these “destined” designs compared with an arbitrary design, which also justifies the proposed guiding principle of AbS.

We first try an arbitrary design of self-attention with top-down input by adding the top-down signal on the query, key, and value instead of only on the value. We name this design as AbSViT-QKV. We compare AbSViT and AbSViT-QKV on image classification and robustness (Table 2.5), and we can see that AbSViT is superior to AbSViT-QKV on every benchmark. This is consistent with our analysis in Section 2.4 that the sparse reconstruction AbS is optimizing has an additional top-down input (corresponding to V), while the dictionary (corresponding to Q and K), which contains templates for separate objects, is fixed.

We also test the effect of the variational loss \mathcal{L}_{var} , which ensures the model is approximating AbS. We compare AbSViT with its counterpart without \mathcal{L}_{var} , *i.e.*, a top-down model trained with only supervised loss. As shown in Table 2.6, adding \mathcal{L}_{var} largely improves the clean accuracy and robustness. Note that, as discussed in Section 2.5, we do not have a prior loss $-\log p(\mathbf{z}_L)$ for image classification, which means the improvement completely comes from the reconstruction loss $\frac{1}{2} \sum_{\ell=1}^L \|sg(\mathbf{z}_\ell) - g_\ell(sg(\mathbf{z}_{\ell+1}))\|_2^2$ which forces the decoder to reconstruct \mathbf{z}_ℓ from $\mathbf{z}_{\ell+1}$. This implies that a generative model (“synthesis”) is important to high-quality top-down attention in visual recognition (“analysis”).

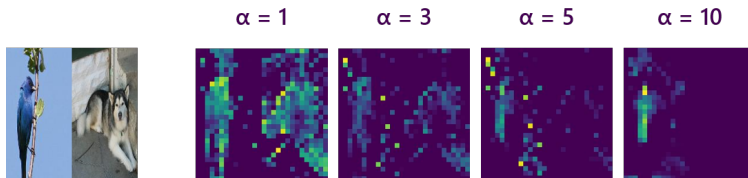


Figure 2.8: Visualization of top-down attention with different scaling factor α . Prior corresponds to the bird. The top-down attention gets more and more biased on the bird when increasing α .

2.7 Limitations and Future Work

ImageNet Classification Is a Poor Teacher of Top-Down Attention

AbSViT is trained to focus on different objects given different priors in multi-object images. However, ImageNet classification targets single object classification without any prior, making it unsuitable for pretraining top-down attention. We find that the ImageNet-supervised AbSViT only learns weak top-down attention. A simple trick to augment the top-down attention for downstream tasks such as VQA is manually setting a larger scaling factor α (e.g., $\alpha = 10$). In Figure 2.8, we visualize the top-down attention with different α . We can see that, with a prior corresponding to the bird, the attention under $\alpha = 1$ still highlights both the bird and the dog but is more and more biased towards the bird as we increase α . For future exploration, we may learn stronger top-down attention through object-level unsupervised learning [137, 46] or vision-language pretraining [141, 88].

How Many Syntheses Do We Need for Analysis?

In Section 2.5, we mention that enforcing strong generative capability on the features \mathbf{z}_ℓ will downgrade the discriminative power regarding classification accuracy. There is a similar observation in recent self-supervised learning work [45], where reconstruction-based algorithms have worse linear-probing performance [10]. However, the empirical results in Table 2.6 indicate that at least some degree of generative power is still helpful. This echoes the classical debate of how much generative capability (“synthesis”) we need for visual discrimination (“analysis”). As a starting point, we measure the generative power of the ImageNet-pretrained AbSViT (Figure 2.9). Specifically, we train a linear decoder that projects the bottom-up input \mathbf{x}_0^{bu} of the first layer to the original image and then visualize the image decoded from the bottom-up signal \mathbf{x}_0^{bu} , the top-down signal \mathbf{x}_0^{td} , or their combination $\mathbf{x}_0^{bu} + \mathbf{x}_0^{td}$. We can see that the bottom-up signal contains full information about the original image and gives a perfect reconstruction. On the other hand, the top-down signal has lost most of the information, which is reasonable considering that \mathbf{x}_0^{td} itself is decoded from the last layer’s feature. Intriguingly, when we combine the bottom-up and the top-down signals, it can reconstruct only the foreground object, implying AbSViT can selectively preserve

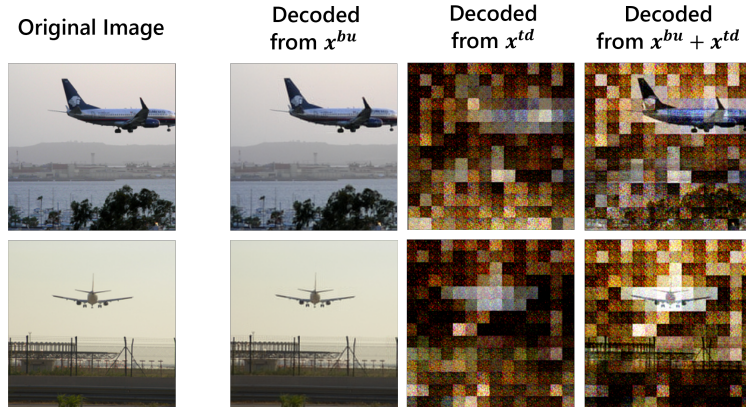


Figure 2.9: Examples of images decoded from the bottom-up, top-down, or the combination of bottom-up and top-down signals. The decoder can reconstruct the whole image from the bottom-up signal while failing to generate anything recognizable from the top-down signal alone. When decoding from the combination of bottom-up and top-down signals, only the foreground object is reconstructed.

partial information in the image, and the selection process is adaptive to different priors. This leaves the question of whether a *selective* generation process is the best companion of the discriminative model and how to control the selective process under different priors adaptively.

2.8 Conclusion

We consider top-down attention by explaining from an Analysis-by-Synthesis (AbS) view of vision. Starting from previous work on the functional equivalence between visual attention and sparse reconstruction, we show that AbS optimizes a similar sparse reconstruction objective but modulates it with a goal-directed top-down modulation, thus simulating top-down attention. We propose AbSViT, a top-down modulated ViT model that variationally approximates AbS. We show that AbSViT achieves controllable top-down attention and improves over baselines on V&L tasks as well as image classification and robustness.

Bibliography

- [1] Peter Anderson et al. “Bottom-up and top-down attention for image captioning and visual question answering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6077–6086.
- [2] Alessandra Angelucci et al. “Circuits for local and global signal integration in primary visual cortex”. In: *Journal of Neuroscience* 22.19 (2002), pp. 8633–8646.
- [3] Stanislaw Antol et al. “Vqa: Visual question answering”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433.
- [4] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. “Deep equilibrium models”. In: *arXiv preprint arXiv:1909.01377* (2019).
- [5] Hangbo Bao, Li Dong, and Furu Wei. “Beit: Bert pre-training of image transformers”. In: *arXiv preprint arXiv:2106.08254* (2021).
- [6] Irving Biederman. “Recognition-by-components: a theory of human image understanding.” In: *Psychological review* 94.2 (1987), p. 115.
- [7] Jeffrey Bisanz, Gay L Bisanz, and Robert Kail. *Learning in children: Progress in cognitive development research*. Springer Science & Business Media, 2012.
- [8] Ali Borji, Dicky Sihite, and Laurent Itti. “An object-based bayesian framework for top-down visual attention”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 26. 1. 2012, pp. 1529–1535.
- [9] William H Bosking et al. “Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex”. In: *Journal of neuroscience* 17.6 (1997), pp. 2112–2127.
- [10] Mathilde Caron et al. “Emerging properties in self-supervised vision transformers”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.
- [11] Marisa Carrasco. “Visual attention: The past 25 years”. In: *Vision research* 51.13 (2011), pp. 1484–1525.
- [12] James R Cavanaugh, Wyeth Bair, and J Anthony Movshon. “Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons”. In: *Journal of neurophysiology* 88.5 (2002), pp. 2530–2546.

- [13] Wesley Chaney, Jason Fischer, and David Whitney. “The hierarchical sparse selection model of visual crowding”. In: *Frontiers in integrative neuroscience* 8 (2014), p. 73.
- [14] Gang Chen. “Where to Look: A Unified Attention Model for Visual Recognition with Reinforcement Learning”. In: *arXiv preprint arXiv:2111.07169* (2021).
- [15] Sharat Chikkerur et al. “What and where: A Bayesian inference theory of attention”. In: *Vision research* 50.22 (2010), pp. 2233–2247.
- [16] Krzysztof Choromanski et al. “Rethinking attention with performers”. In: *arXiv preprint arXiv:2009.14794* (2020).
- [17] Michael A Cohen and Stephen Grossberg. “Absolute stability of global pattern formation and parallel memory storage by competitive neural networks”. In: *IEEE transactions on systems, man, and cybernetics* 5 (1983), pp. 815–826.
- [18] MMSegmentation Contributors. *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. <https://github.com/open-mmlab/mms Segmentation>. 2020.
- [19] Marius Cordts et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
- [20] Stéphane d’Ascoli et al. “Convit: Improving vision transformers with soft convolutional inductive biases”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2286–2296.
- [21] Trevor Darrell, Stan Sclaroff, and Alex Pentland. “Segmentation by minimal description”. In: *Proceedings Third International Conference on Computer Vision*. IEEE Computer Society. 1990, pp. 112–113.
- [22] Abhishek Das et al. “Human attention in visual question answering: Do humans and deep networks look at the same regions?” In: *Computer Vision and Image Understanding* 163 (2017), pp. 90–100.
- [23] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational Neuroscience Series, 2001.
- [24] Peter Dayan et al. “The helmholtz machine”. In: *Neural computation* 7.5 (1995), pp. 889–904.
- [25] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [26] Robert Desimone, John Duncan, et al. “Neural mechanisms of selective visual attention”. In: *Annual review of neuroscience* 18.1 (1995), pp. 193–222.
- [27] Josip Djolonga et al. “On robustness and transferability of convolutional neural networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 16458–16468.

- [28] Samuel Dodge and Lina Karam. “A study and comparison of human and deep learning recognition performance under visual distortions”. In: *2017 26th international conference on computer communication and networks (ICCCN)*. IEEE. 2017, pp. 1–7.
- [29] Minjing Dong et al. “Adversarially robust neural architectures”. In: *arXiv preprint arXiv:2009.00902* (2020).
- [30] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).
- [31] Zi-Yi Dou et al. “An empirical study of training end-to-end vision-and-language transformers”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18166–18176.
- [32] M. Everingham et al. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [33] Mark Everingham et al. “The pascal visual object classes (voc) challenge”. In: *IJCV* 88.2 (2010), pp. 303–338.
- [34] Yuchen Fan et al. “Neural Sparse Representation for Image Restoration”. In: *arXiv preprint arXiv:2006.04357* (2020).
- [35] Daniel J Felleman and David C Van Essen. “Distributed hierarchical processing in the primate cerebral cortex.” In: *Cerebral cortex (New York, NY: 1991)* 1.1 (1991), pp. 1–47.
- [36] Robert Geirhos et al. “Comparing deep neural networks against humans: object recognition when the signal gets weaker”. In: *arXiv preprint arXiv:1706.06969* (2017).
- [37] Robert Geirhos et al. “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”. In: *arXiv preprint arXiv:1811.12231* (2018).
- [38] Charles D Gilbert and Torsten N Wiesel. “Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex”. In: *Journal of Neuroscience* 9.7 (1989), pp. 2432–2442.
- [39] Yash Goyal et al. “Making the v in vqa matter: Elevating the role of image understanding in visual question answering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6904–6913.
- [40] Karol Gregor and Yann LeCun. “Learning fast approximations of sparse coding”. In: *Proceedings of the 27th international conference on international conference on machine learning*. 2010, pp. 399–406.
- [41] Stephen Grossberg and Ennio Mingolla. “Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations”. In: *The adaptive brain II*. Elsevier, 1987, pp. 143–210.
- [42] Yiwen Guo et al. “Sparse dnns with improved adversarial robustness”. In: *arXiv preprint arXiv:1810.09619* (2018).

- [43] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [44] Kaiming He et al. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.
- [45] Kaiming He et al. “Masked autoencoders are scalable vision learners”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16000–16009.
- [46] Olivier J Hénaff et al. “Object discovery and representation networks”. In: *arXiv preprint arXiv:2203.08777* (2022).
- [47] Dan Hendrycks and Thomas Dietterich. “Benchmarking neural network robustness to common corruptions and perturbations”. In: *arXiv preprint arXiv:1903.12261* (2019).
- [48] Dan Hendrycks et al. “Natural adversarial examples”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 15262–15271.
- [49] Dan Hendrycks et al. “The many faces of robustness: A critical analysis of out-of-distribution generalization”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 8340–8349.
- [50] Byeongho Heo et al. “Rethinking spatial dimensions of vision transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11936–11945.
- [51] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. “A fast learning algorithm for deep belief nets”. In: *Neural computation* 18.7 (2006), pp. 1527–1554.
- [52] Geoffrey E Hinton and Ruslan R Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *science* 313.5786 (2006), pp. 504–507.
- [53] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.
- [54] Torsten Hoeffler et al. “Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks”. In: *arXiv preprint arXiv:2102.00554* (2021).
- [55] John J Hopfield. “Neurons with graded response have collective computational properties like those of two-state neurons”. In: *Proceedings of the national academy of sciences* 81.10 (1984), pp. 3088–3092.
- [56] Yujia Huang et al. “Neural networks with recurrent generative feedback”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 535–545.
- [57] Laurent Itti and Christof Koch. “Computational modelling of visual attention”. In: *Nature reviews neuroscience* 2.3 (2001), pp. 194–203.
- [58] Andrew Jaegle et al. “Perceiver io: A general architecture for structured inputs & outputs”. In: *arXiv preprint arXiv:2107.14795* (2021).

- [59] Tilke Judd et al. “Learning to predict where humans look”. In: *2009 IEEE 12th international conference on computer vision*. IEEE. 2009, pp. 2106–2113.
- [60] Kohitij Kar et al. “Evidence that recurrent circuits are critical to the ventral stream’s execution of core object recognition behavior”. In: *Nature neuroscience* 22.6 (2019), pp. 974–983.
- [61] Junkyung Kim et al. “Disentangling neural mechanisms for perceptual grouping”. In: *arXiv preprint arXiv:1906.01558* (2019).
- [62] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).
- [63] David C Knill and Whitman Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.
- [64] Christof Koch and Shimon Ullman. “Shifts in selective visual attention: towards the underlying neural circuitry”. In: *Matters of intelligence*. Springer, 1987, pp. 115–141.
- [65] Victor AF Lamme and Pieter R Roelfsema. “The distinct modes of vision offered by feedforward and recurrent processing”. In: *Trends in neurosciences* 23.11 (2000), pp. 571–579.
- [66] Honglak Lee, Chaitanya Ekanadham, and Andrew Ng. “Sparse deep belief net model for visual area V2”. In: *Advances in neural information processing systems* 20 (2007), pp. 873–880.
- [67] T Sing Lee. “Analysis and synthesis of visual images in the brain: evidence for pattern theory”. In: *IMA VOLUMES IN MATHEMATICS AND ITS APPLICATIONS* 133 (2003), pp. 87–106.
- [68] Tai Sing Lee. “Top-down influence in early visual processing: a Bayesian perspective”. In: *Physiology & behavior* 77.4-5 (2002), pp. 645–650.
- [69] Tai Sing Lee and David Mumford. “Hierarchical Bayesian inference in the visual cortex”. In: *JOSA A* 20.7 (2003), pp. 1434–1448.
- [70] Da Li et al. “Deeper, broader and artier domain generalization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5542–5550.
- [71] Zhaoping Li. “A neural model of contour integration in the primary visual cortex”. In: *Neural computation* 10.4 (1998), pp. 903–940.
- [72] Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.
- [73] Drew Linsley et al. “Recurrent neural circuits for contour detection”. In: *arXiv preprint arXiv:2010.15314* (2020).
- [74] Shiwei Liu, Decebal Constantin Mocanu, and Mykola Pechenizkiy. “On improving deep learning generalization with adaptive sparse connectivity”. In: *arXiv preprint arXiv:1906.11626* (2019).

- [75] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.
- [76] Zhuang Liu et al. “A convnet for the 2020s”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11976–11986.
- [77] Jiasen Lu et al. “Hierarchical question-image co-attention for visual question answering”. In: *Advances in neural information processing systems* 29 (2016).
- [78] Jiasen Lu et al. “Knowing when to look: Adaptive attention via a visual sentinel for image captioning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 375–383.
- [79] Harshitha Machiraju et al. “Bio-inspired Robustness: A Review”. In: *arXiv preprint arXiv:2103.09265* (2021).
- [80] Aleksander Madry et al. “Towards deep learning models resistant to adversarial attacks”. In: *arXiv preprint arXiv:1706.06083* (2017).
- [81] Xiaofeng Mao et al. “Towards robust vision transformer”. In: *arXiv preprint arXiv:2105.07926* (2021).
- [82] Xiaofeng Mao et al. “Towards robust vision transformer”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12042–12051.
- [83] Julio C Martinez-Trujillo and Stefan Treue. “Attentional modulation strength in cortical area MT depends on stimulus contrast”. In: *Neuron* 35.2 (2002), pp. 365–370.
- [84] Carrie J McAdams and John HR Maunsell. “Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4”. In: *Journal of Neuroscience* 19.1 (1999), pp. 431–441.
- [85] David A Mély, Drew Linsley, and Thomas Serre. “Complementary surrounds explain diverse contextual phenomena across visual modalities.” In: *Psychological review* 125.5 (2018), p. 769.
- [86] M Berk Mirza et al. “Introducing a Bayesian model of selective attention based on active inference”. In: *Scientific reports* 9.1 (2019), pp. 1–22.
- [87] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. “Recurrent models of visual attention”. In: *Advances in neural information processing systems*. 2014, pp. 2204–2212.
- [88] Jishnu Mukhoti et al. “Open Vocabulary Semantic Segmentation with Patch Aligned Contrastive Learning”. In: *arXiv preprint arXiv:2212.04994* (2022).
- [89] Muzammal Naseer et al. “Intriguing Properties of Vision Transformers”. In: *arXiv preprint arXiv:2105.10497* (2021).

- [90] Aran Nayebi et al. “Task-driven convolutional recurrent models of the visual system”. In: *arXiv preprint arXiv:1807.00053* (2018).
- [91] Randall C O’Reilly et al. “Recurrent processing during object recognition”. In: *Frontiers in psychology* 4 (2013), p. 124.
- [92] Aude Oliva et al. “Top-down control of visual attention in object detection”. In: *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*. Vol. 1. IEEE. 2003, pp. I–253.
- [93] Bruno A Olshausen and David J Field. “Emergence of simple-cell receptive field properties by learning a sparse code for natural images”. In: *Nature* 381.6583 (1996), pp. 607–609.
- [94] Bruno A Olshausen and David J Field. “Sparse coding with an overcomplete basis set: A strategy employed by V1?” In: *Vision research* 37.23 (1997), pp. 3311–3325.
- [95] Bo Pang et al. “Tdaf: Top-down attention framework for vision tasks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 3. 2021, pp. 2384–2392.
- [96] Sayak Paul and Pin-Yu Chen. “Vision transformers are robust learners”. In: *arXiv preprint arXiv:2105.07581* (2021).
- [97] Robert J Peters et al. “Components of bottom-up gaze allocation in natural images”. In: *Vision research* 45.18 (2005), pp. 2397–2416.
- [98] Bryan A Plummer et al. “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2641–2649.
- [99] Michael I Posner. “Orienting of attention”. In: *Quarterly journal of experimental psychology* 32.1 (1980), pp. 3–25.
- [100] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8748–8763.
- [101] Ilija Radosavovic et al. “Designing network design spaces”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10428–10436.
- [102] Rajesh PN Rao. “Bayesian inference and attentional modulation in the visual cortex”. In: *Neuroreport* 16.16 (2005), pp. 1843–1848.
- [103] Yongming Rao et al. “Global filter networks for image classification”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 980–993.
- [104] Sylvestre-Alvise Rebuffi et al. “Data Augmentation Can Improve Robustness”. In: *Advances in Neural Information Processing Systems* 34 (2021).
- [105] John H Reynolds and David J Heeger. “The normalization model of attention”. In: *Neuron* 61.2 (2009), pp. 168–185.

- [106] Pieter R Roelfsema. “Cortical algorithms for perceptual grouping”. In: *Annu. Rev. Neurosci.* 29 (2006), pp. 203–227.
- [107] Pieter R Roelfsema et al. “Figure—ground segregation in a recurrent network architecture”. In: *Journal of cognitive neuroscience* 14.4 (2002), pp. 525–537.
- [108] Christopher J Rozell et al. “Sparse coding via thresholding and local competition in neural circuits”. In: *Neural computation* 20.10 (2008), pp. 2526–2563.
- [109] Michael P Sceniak, Michael J Hawken, and Robert Shapley. “Visual spatial characterization of macaque V1 neurons”. In: *Journal of neurophysiology* 85.5 (2001), pp. 1873–1887.
- [110] Brian J Scholl. “Objects and attention: The state of the art”. In: *Cognition* 80.1-2 (2001), pp. 1–46.
- [111] Paul R Schrater and Rashmi Sundareswara. “Theory and dynamics of perceptual bistability”. In: *Advances in neural information processing systems* 19 (2006).
- [112] Baifeng Shi et al. “Informative dropout for robust representation learning: A shape-bias perspective”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8828–8839.
- [113] Baifeng Shi et al. “Visual Attention Emerges from Recurrent Sparse Reconstruction”. In: *arXiv preprint arXiv:2204.10962* (2022).
- [114] Ekta Sood et al. “Interpreting attention models with human visual attention in machine reading comprehension”. In: *arXiv preprint arXiv:2010.06396* (2020).
- [115] Dan D Stettler et al. “Lateral connectivity and contextual interactions in macaque primary visual cortex”. In: *Neuron* 36.4 (2002), pp. 739–750.
- [116] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *arXiv preprint arXiv:1312.6199* (2013).
- [117] Christian Szegedy et al. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.
- [118] Hanlin Tang et al. “Recurrent computations for visual pattern completion”. In: *Proceedings of the National Academy of Sciences* 115.35 (2018), pp. 8835–8840.
- [119] Shengbang Tong et al. “Unsupervised Learning of Structured Representations via Closed-Loop Transcription”. In: *arXiv preprint arXiv:2210.16782* (2022).
- [120] Hugo Touvron et al. “Training data-efficient image transformers & distillation through attention”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10347–10357.
- [121] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

- [122] Shaun P Vecera and Martha J Farah. “Is visual image segmentation a bottom-up or an interactive process?” In: *Perception & Psychophysics* 59.8 (1997), pp. 1280–1296.
- [123] Pascal Vincent et al. “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.” In: *Journal of machine learning research* 11.12 (2010).
- [124] Dirk Walther et al. “Selective visual attention enables learning and recognition of multiple objects in cluttered scenes”. In: *Computer Vision and Image Understanding* 100.1-2 (2005), pp. 41–63.
- [125] Haohan Wang et al. “Learning robust global representations by penalizing local predictive power”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [126] Wenhai Wang et al. “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 568–578.
- [127] Xiaolong Wang et al. “Non-local neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7794–7803.
- [128] Xin Wang et al. “Robust Object Detection via Instance-Level Temporal Cycle Confusion”. In: *arXiv preprint arXiv:2104.08381* (2021).
- [129] John Wright et al. “Robust face recognition via sparse representation”. In: *IEEE transactions on pattern analysis and machine intelligence* 31.2 (2008), pp. 210–227.
- [130] Mike Wu et al. “Meta-amortized variational inference and learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 6404–6412.
- [131] Si Wu et al. “Continuous attractor neural networks: candidate of a canonical model for neural information representation”. In: *F1000Research* 5 (2016).
- [132] Yan Wu, Mihaela Rosca, and Timothy Lillicrap. “Deep compressed sensing”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6850–6860.
- [133] Dean Wyatte, Tim Curran, and Randall O’Reilly. “The limits of feedforward vision: Recurrent processing promotes robust object recognition when objects are degraded”. In: *Journal of Cognitive Neuroscience* 24.11 (2012), pp. 2248–2261.
- [134] Dean Wyatte, David J Jilk, and Randall C O’Reilly. “Early recurrent feedback facilitates visual object recognition under challenging conditions”. In: *Frontiers in psychology* 5 (2014), p. 674.
- [135] Yingda Xia et al. “Synthesize then compare: Detecting failures and anomalies for semantic segmentation”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 145–161.
- [136] Tete Xiao et al. “Early convolutions help transformers see better”. In: *Advances in Neural Information Processing Systems* 34 (2021).

- [137] Tete Xiao et al. “Region similarity representation learning”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10539–10548.
- [138] Tete Xiao et al. “Unified perceptual parsing for scene understanding”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 418–434.
- [139] Saining Xie et al. “Aggregated residual transformations for deep neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1492–1500.
- [140] Huijuan Xu and Kate Saenko. “Ask, attend and answer: Exploring question-guided spatial attention for visual question answering”. In: *European conference on computer vision*. Springer. 2016, pp. 451–466.
- [141] Jiarui Xu et al. “Groupvit: Semantic segmentation emerges from text supervision”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18134–18144.
- [142] Kelvin Xu et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057.
- [143] Zichao Yang et al. “Stacked attention networks for image question answering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 21–29.
- [144] Shih-Cheng Yen and Leif H Finkel. “Extraction of perceptually salient contours by striate cortical networks”. In: *Vision research* 38.5 (1998), pp. 719–741.
- [145] Angela J Yu and Peter Dayan. “Inference, attention, and decision in a Bayesian neural architecture”. In: *Advances in neural information processing systems* 17 (2004).
- [146] Alan Yuille and Daniel Kersten. “Vision as Bayesian inference: analysis by synthesis?”. In: *Trends in cognitive sciences* 10.7 (2006), pp. 301–308.
- [147] Li Zhaoping. *Understanding vision: theory, models, and data*. OUP Oxford, 2014.
- [148] Bolei Zhou et al. “Scene parsing through ade20k dataset”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 633–641.
- [149] Daquan Zhou et al. “Understanding the robustness in vision transformers”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 27378–27394.
- [150] Daniel Zoran et al. “Towards robust image classification using sequential attention models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9483–9492.
- [151] Steven W Zucker, Allan Dobbins, and Lee Iverson. “Two stages of curve detection suggest two styles of visual computation”. In: *Neural computation* 1.1 (1989), pp. 68–81.

Appendix A

Appendix

A.1 Additional Qualitative Results

Evolution of Attention Maps during Recurrent Updates in VARS

In Figure A.1 we show more examples of the evolution of attention maps in each step of the recurrent update in VARS. Here we choose VARS-D built upon the RVT baseline and visualize the attention map of the last layer in the first block. We can observe that the attention map is more sharp and concentrated on the salient objects after each iteration.

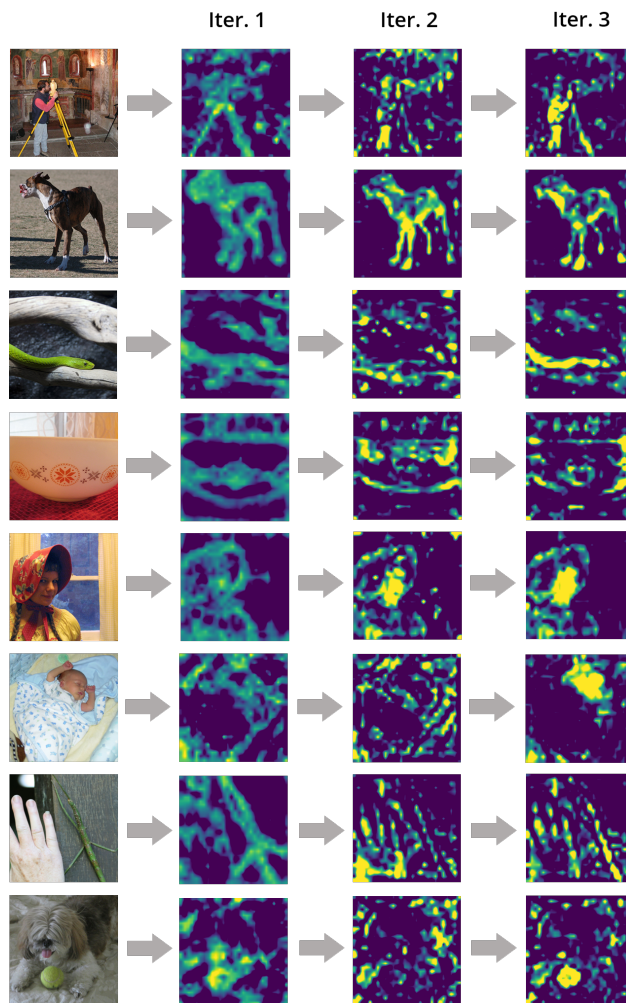


Figure A.1: Visualization of the attention maps after each iteration of update in VARS. One can see that the attention will be more concentrated on the salient objects in the image after each update.

Visualization of Dynamic Dictionaries in VARS-D and VARS-SD

In VARS-D and VARS-SD, we use the input-dependent dictionary $\Phi(\mathbf{X})$ for sparse reconstruction. Here we visualize the dynamic dictionaries for a deeper understanding (Figure A.2-A.4). We can see that most atoms in the dictionary are approximately uniform masks on either foreground or background regions. This is a direct consequence from the design of self-attention and random features [16].

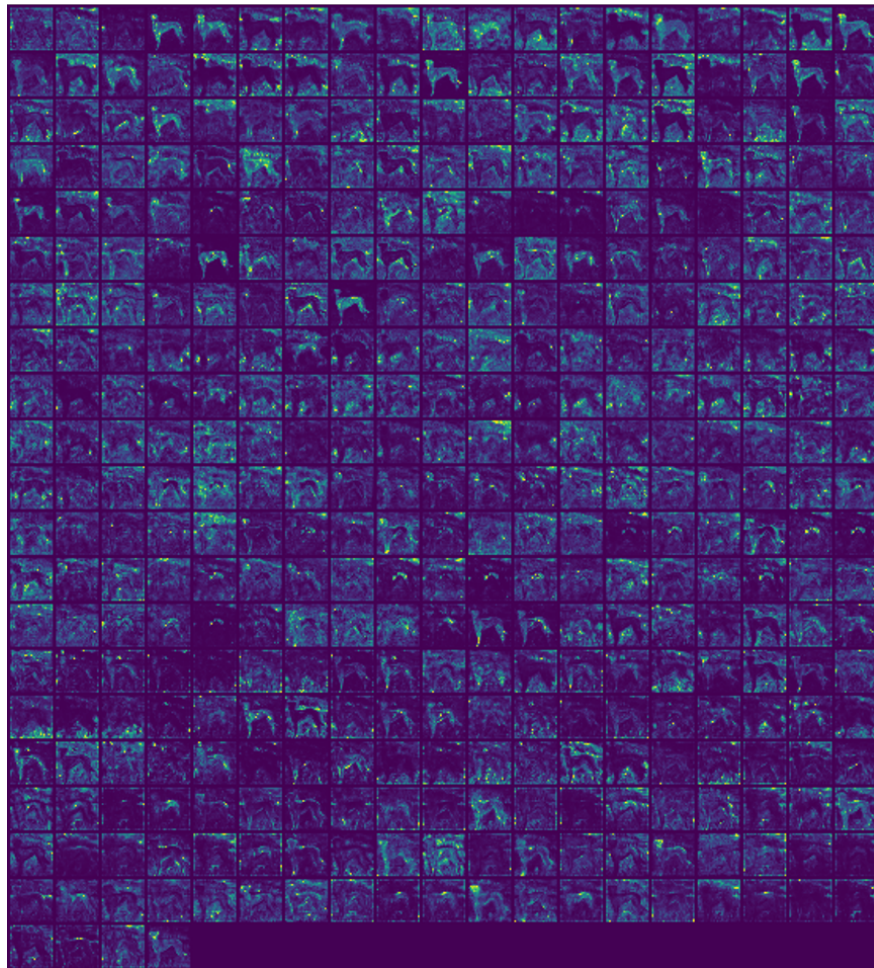


Figure A.2: Visualization of the dynamic dictionary. The atoms are mostly masks on either foreground objects or backgrounds.

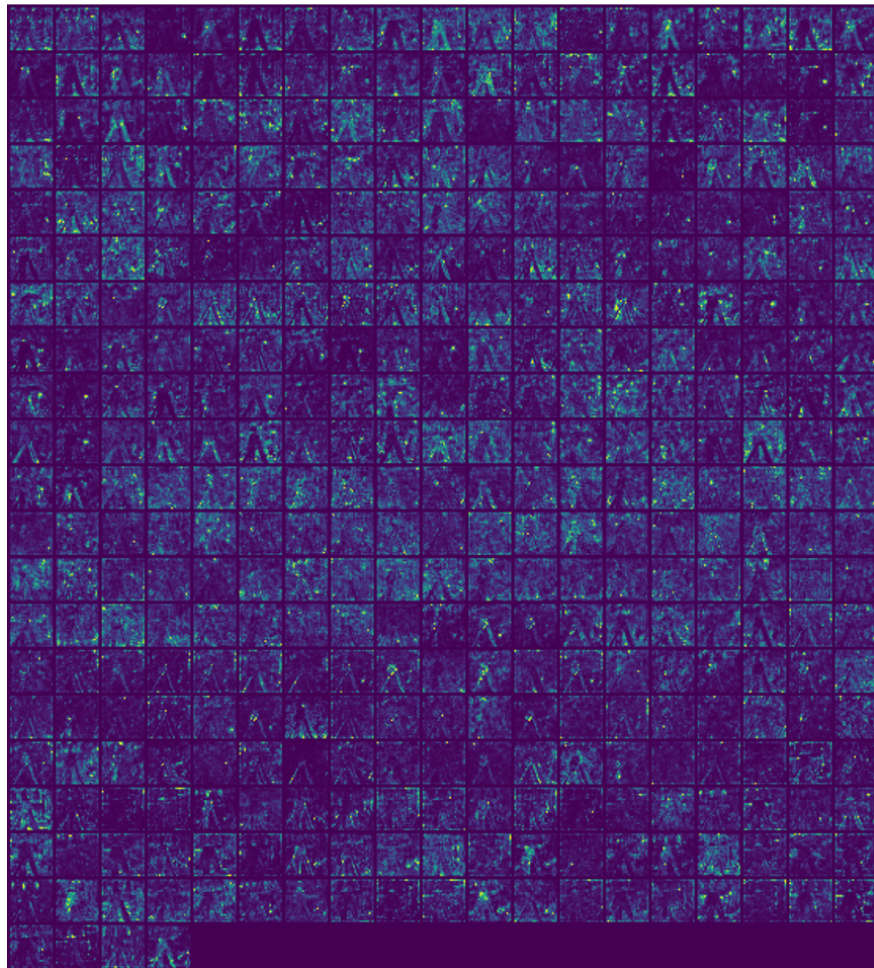


Figure A.3: Visualization of the dynamic dictionary. The atoms are mostly masks on either foreground objects or backgrounds.

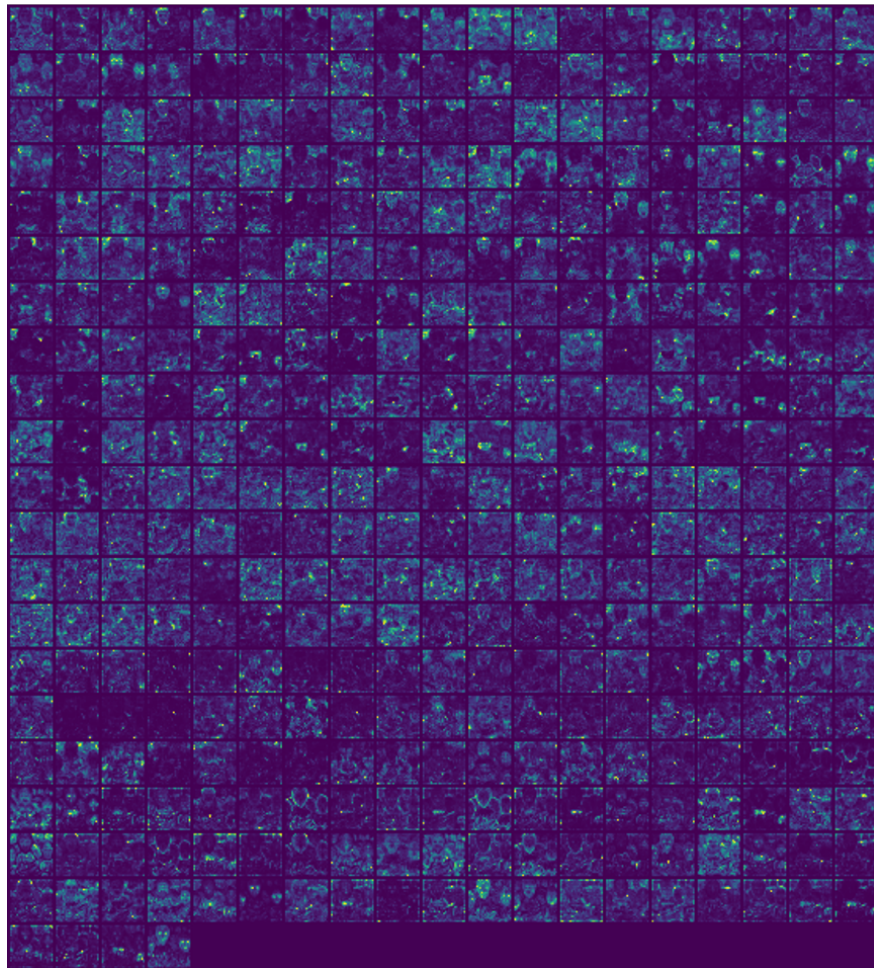


Figure A.4: Visualization of the dynamic dictionary. The atoms are mostly masks on either foreground objects or backgrounds.

Visualization of Attention Maps under Different Image Corruptions

We show additional visualization of the attention maps under different image corruptions in Figure A.5, where each block contains attention maps of clean images as well as images under noise, blur, digital, and weather corruptions (from top to down). We show the attention maps of RVT*, VARS-S, VARS-D, and VARS-SD (from left to right). One can see that the attention maps of self-attention baseline is more sensitive to image corruptions, while variants of VARS tend to output stable attention maps. Meanwhile, the attention maps of VARS-S are steady but not as sharp as VARS-D and VARS-SD.

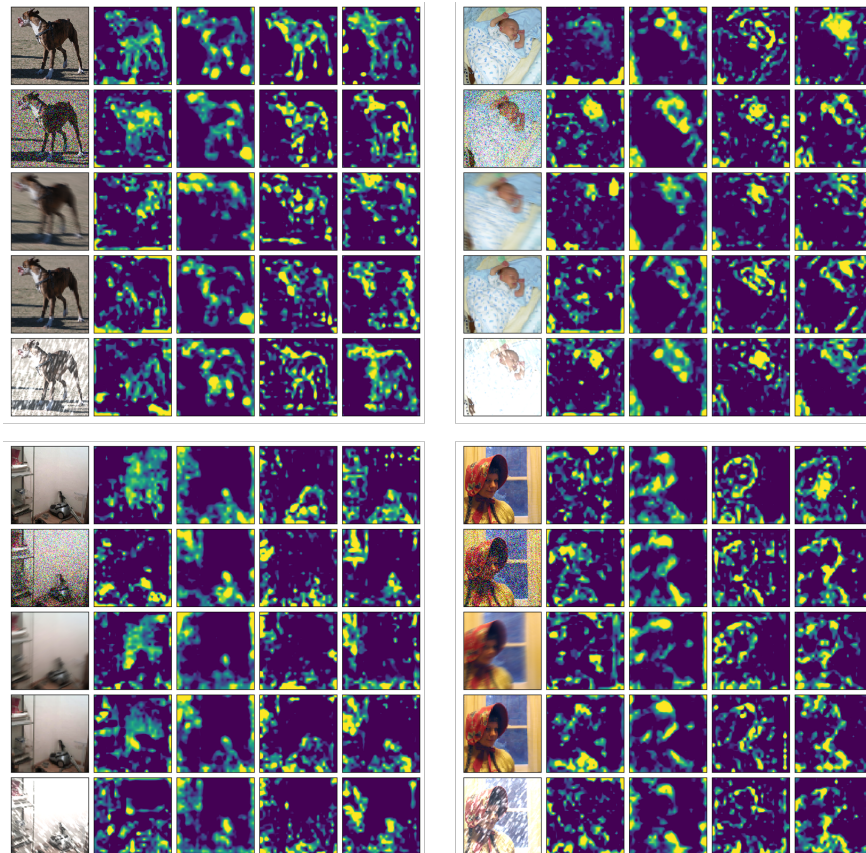


Figure A.5: Visualization of the attention maps under image corruptions. Each block contains clean images as well as images under corruption of noise, blur, digital, and weather (from top to down). We visualize the attention map of RVT*, VARS-S, VARS-D, and VARS-SD (from left to right in each block). Across different images, self-attention is usually more unstable than the variants of VARS. Meanwhile, VARS-S has attention maps that are consistent under different corruptions but are not as sharp as those of VARS-D and VARS-SD.

Comparing Attention Maps with Human Eye Fixation

In Figure A.6 we show additional results on comparing the attention maps of different models with the human eye fixation data. We can see that, the attention maps of VARS are more consistent with human eye fixation than self-attention.

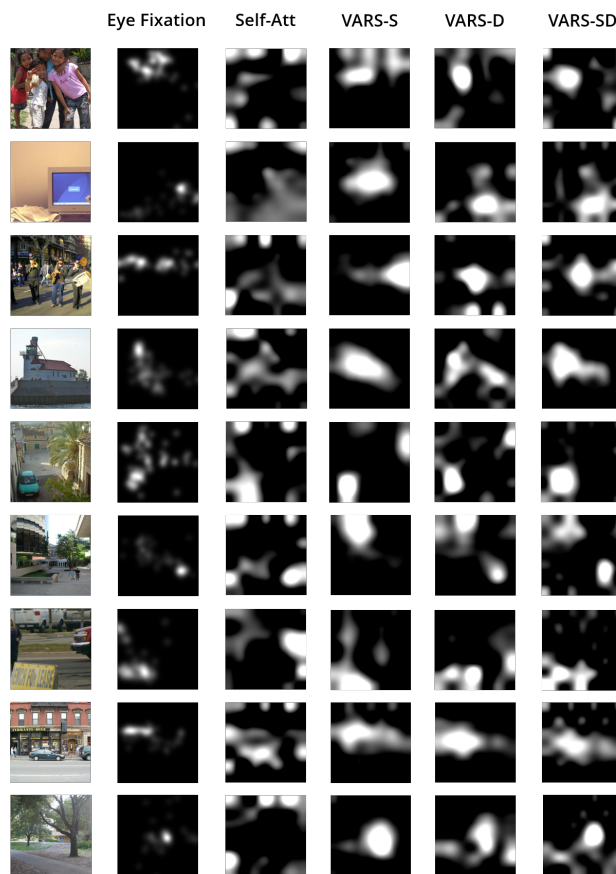


Figure A.6: Comparison between attention maps of different models and human eye fixation probabilities. The variants of VARS have attention maps that are more consistent with human eye fixation.

A.2 Details on Derivation of the Sparse Reconstruction Problem

We follow [108] and add recurrent connections between \mathbf{u} , modeled by the weight matrix $-\gamma(\mathbf{P}^T\mathbf{P} - \mathbf{I})$. We also add hyperparameters α and β to control the strength of self-leakage and the element-wise activation functions $g(\cdot)$ to gate the output from the neurons [23]. As a result, we fix the dynamics of the recurrent networks as:

$$\begin{cases} \frac{d\mathbf{z}}{dt} = -\alpha\mathbf{z} + \mathbf{P}g(\mathbf{u}) + \mathbf{x}, \\ \frac{d\mathbf{u}}{dt} = -\beta\mathbf{u} - \gamma(\mathbf{P}^T\mathbf{P} - \mathbf{I})g(\mathbf{u}) + \mathbf{P}^T\mathbf{z}. \end{cases} \quad (\text{A.1})$$

$$\frac{d\mathbf{u}}{dt} = -\beta\mathbf{u} - \gamma(\mathbf{P}^T\mathbf{P} - \mathbf{I})g(\mathbf{u}) + \mathbf{P}^T\mathbf{z}. \quad (\text{A.2})$$

By taking $\alpha = 1$ and $\beta = \gamma = 2$, it has the same steady state solution as

$$\begin{cases} \frac{d\mathbf{u}}{dt} = -2(\mathbf{u} - \tilde{\mathbf{u}}) - \mathbf{P}^T\mathbf{P}\tilde{\mathbf{u}} + \mathbf{P}^T\mathbf{x}, \\ \mathbf{z} = \mathbf{P}\tilde{\mathbf{u}} + \mathbf{x}, \end{cases} \quad (\text{A.3})$$

$$\mathbf{z} = \mathbf{P}\tilde{\mathbf{u}} + \mathbf{x}, \quad (\text{A.4})$$

where $\tilde{\mathbf{u}} = g(\mathbf{u})$. Now we choose $g(\cdot)$ as the thresholding function $g(\mathbf{u}_i) = \text{sgn}(\mathbf{u}_i) \cdot (|\mathbf{u}_i| - \lambda)_+$, where $\text{sgn}(\cdot)$ is the sign function and $(\cdot)_+$ is ReLU. Under the assumption that $g(\cdot)$ is monotonically non-decreasing, Eq. A.3 is actually minimizing the energy function

$$E(\tilde{\mathbf{u}}) = \frac{1}{2} \|\mathbf{P}\tilde{\mathbf{u}} - \mathbf{x}\|^2 + 2\lambda \|\tilde{\mathbf{u}}\|_1. \quad (\text{A.5})$$

To see this, one can verify that when \mathbf{u} evolves by Eq. A.3, $E(\tilde{\mathbf{u}})$ is non-increasing, *i.e.*,

$$\frac{dE}{dt} = -(2\lambda \cdot \text{sgn}(\mathbf{u}) + \mathbf{P}^T\mathbf{P}\tilde{\mathbf{u}} - \mathbf{P}^T\mathbf{x})^T \cdot \mathbf{K}(\mathbf{u}) \cdot (2\lambda \cdot \text{sgn}(\mathbf{u}) + \mathbf{P}^T\mathbf{P}\tilde{\mathbf{u}} - \mathbf{P}^T\mathbf{x}) = \left(\frac{d\mathbf{u}}{dt}\right)^T \mathbf{K}(\mathbf{u}) \frac{d\mathbf{u}}{dt}, \quad (\text{A.6})$$

where $\mathbf{K}(\mathbf{u})$ is a diagonal matrix with $\mathbf{K}(\mathbf{u})_{ii} = 1$ when $(i) |\mathbf{u}_i| > 1$, and $(ii) |\mathbf{u}_i| = 1$ and $\frac{d\mathbf{u}_i}{dt} \cdot \text{sgn}(\mathbf{u}_i) > 0$, otherwise $\mathbf{K}(\mathbf{u})_{ii} = 0$. Since $\mathbf{K}(\mathbf{u})$ is positive semi-definite, Eq. A.6 is non-positive, which means the energy is non-increasing. Then Eq. A.3 equivalently optimizes the sparse reconstruction:

$$\begin{cases} \tilde{\mathbf{u}}^* = \arg \min_{\tilde{\mathbf{u}} \in \mathbb{R}^{d'}} \frac{1}{2} \|\mathbf{P}\tilde{\mathbf{u}} - \mathbf{x}\|^2 + 2\lambda \|\tilde{\mathbf{u}}\|_1 \\ \mathbf{z}^* = \mathbf{P}\tilde{\mathbf{u}}^* + \mathbf{x}. \end{cases} \quad (\text{A.7})$$

$$\mathbf{z}^* = \mathbf{P}\tilde{\mathbf{u}}^* + \mathbf{x}. \quad (\text{A.8})$$

A.3 Derivation of Eq. (10)

From Eq. (6-7) we have

$$p(\tilde{\mathbf{u}}_\ell | \tilde{\mathbf{u}}_{\ell+1}) \propto \exp\left\{-\frac{1}{2} \|\mathbf{P}_\ell \tilde{\mathbf{u}}_\ell - g_\ell(\mathbf{P}_{\ell+1} \tilde{\mathbf{u}}_{\ell+1})\|_2^2 - \lambda \|\tilde{\mathbf{u}}_\ell\|_1\right\}. \quad (\text{A.9})$$

Then Eq. (10) is derived by

$$\begin{aligned}
\frac{d\tilde{\mathbf{u}}_\ell}{dt} &\propto \nabla_{\tilde{\mathbf{u}}_\ell} \log p(\tilde{\mathbf{u}}_{\ell-1}|\tilde{\mathbf{u}}_\ell) + \nabla_{\tilde{\mathbf{u}}_\ell} \log p(\tilde{\mathbf{u}}_\ell|\tilde{\mathbf{u}}_{\ell-1}) \\
&= -\nabla_{\tilde{\mathbf{u}}_\ell} \frac{1}{2} \|\mathbf{P}_{\ell-1}\tilde{\mathbf{u}}_{\ell-1} - g_{\ell-1}(\mathbf{P}_\ell\tilde{\mathbf{u}}_\ell)\|_2^2 - \nabla_{\tilde{\mathbf{u}}_\ell} \frac{1}{2} \|\mathbf{P}_\ell\tilde{\mathbf{u}}_\ell - g_\ell(\mathbf{P}_{\ell+1}\tilde{\mathbf{u}}_{\ell+1})\|_2^2 - \nabla_{\tilde{\mathbf{u}}_\ell} \lambda \|\tilde{\mathbf{u}}_\ell\|_1 \\
&= \mathbf{P}_\ell^T \mathbf{J}_{\ell-1}^T (\mathbf{P}_{\ell-1}\tilde{\mathbf{u}}_{\ell-1} - g_{\ell-1}(\mathbf{P}_\ell\tilde{\mathbf{u}}_\ell)) - \mathbf{P}_\ell^T (\mathbf{P}_\ell\tilde{\mathbf{u}}_\ell - g_\ell(\mathbf{P}_{\ell+1}\tilde{\mathbf{u}}_{\ell+1})) - \nabla_{\tilde{\mathbf{u}}_\ell} \lambda \|\tilde{\mathbf{u}}_\ell\|_1 \\
&= -\mathbf{P}_\ell^T (\mathbf{P}_\ell\tilde{\mathbf{u}}_\ell - g_\ell(\mathbf{P}_{\ell+1}\tilde{\mathbf{u}}_{\ell+1})) - \mathbf{J}_{\ell-1}^T \mathbf{P}_{\ell-1}\tilde{\mathbf{u}}_{\ell-1} - \nabla_{\tilde{\mathbf{u}}_\ell} \lambda \|\tilde{\mathbf{u}}_\ell\|_1 - \mathbf{P}_\ell^T \mathbf{J}_{\ell-1}^T g_{\ell-1}(\mathbf{P}_\ell\tilde{\mathbf{u}}_\ell) \\
&= -\nabla_{\tilde{\mathbf{u}}_\ell} \frac{1}{2} \|\mathbf{P}_\ell\tilde{\mathbf{u}}_\ell - g_\ell(\mathbf{z}_{\ell+1}) - \mathbf{J}_{\ell-1}^T \mathbf{z}_{\ell-1}\|_2^2 - \nabla_{\tilde{\mathbf{u}}_\ell} \lambda \|\tilde{\mathbf{u}}_\ell\|_1 - \nabla_{\tilde{\mathbf{u}}_\ell} \frac{1}{2} \|g_{\ell-1}(\mathbf{P}_\ell\tilde{\mathbf{u}}_\ell)\|_2^2 \\
&= -\nabla_{\tilde{\mathbf{u}}_\ell} \frac{1}{2} \|\mathbf{P}_\ell\tilde{\mathbf{u}}_\ell - (\mathbf{x}_\ell^{td} + \mathbf{x}_\ell^{bu})\|_2^2 - \nabla_{\tilde{\mathbf{u}}_\ell} \lambda \|\tilde{\mathbf{u}}_\ell\|_1 - \nabla_{\tilde{\mathbf{u}}_\ell} \frac{1}{2} \|g_{\ell-1}(\mathbf{P}_\ell\tilde{\mathbf{u}}_\ell)\|_2^2.
\end{aligned} \tag{A.10}$$

We informally use ∇ for subgradients as well.

A.4 Additional Results on Natural Images

In Fig.2.4-2.5, we show examples of top-down attention on artificial images. Here we show more results on natural images containing multiple objects. We borrow the LVIS dataset and collect images that contain object categories that also appear in ImageNet. We demonstrate that given different prior, AbSViT is able to focus on different objects in the same image (Figure A.7). We also compare AbSViT’s top-down attention with several baseline methods (Figure A.8) and observe that AbSViT has cleaner attention maps than other methods.

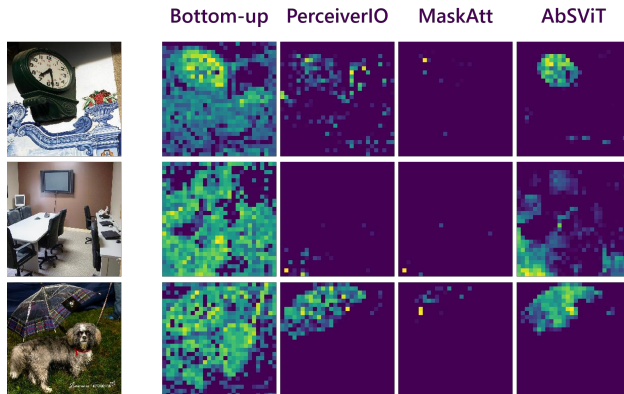


Figure A.8: Comparison of top-down attention map between AbSViT and different baselines.

A.5 Additional Implementation Details

ImageNet Pretraining. The ViT and RVT baselines as well as our AbSViT model are trained using the recipe in [82], and FAN is trained using the recipe in its original paper [149].

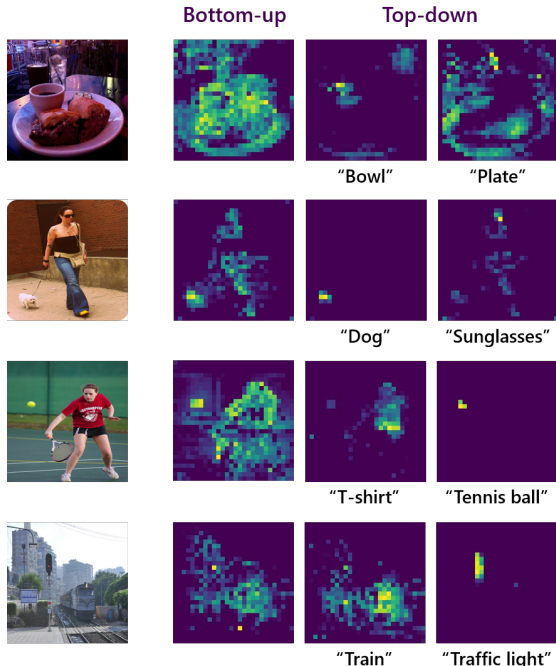


Figure A.7: Visualization of top-down attention on natural images. From left to right, we show the original images, the bottom-up attention, as well as the top-down attention regarding to different objects in each image.

Specifically, we use AdamW optimizer to train AbSViT for 300 epochs, with a batch size of 512, a base learning rate of $5e-4$, and 5 warm-up epochs. One may use different batch-size and adjust the learning rate by the linear scaling rule. We use a cosine learning rate scheduling and weight decay of 0.05. We use the default setting of data augmentation, which includes Mixup, Cutmix, ColorJittering, AutoAugmentation, and Random Erasing. For AbSViT, the weights of supervised loss and variational loss are set as 1 and 0.1.

Robustness against Image Corruptions. We evaluate model robustness against image corruption on ImageNet-C, which contains a total of 19 corruption types. We follow [82] and evaluate 15 types of corruption including Brightness, Contrast, Defocus Blur, Elastic Transform, Fog, Frost, Gaussian Noise, Glass Blur, Impulse Noise, JPEG Compression, Motion Blur, Pixelate, Shot Noise, Snow, and Zoom Blur. Note that other work (e.g. [149]) tests on a different subset of corruption types. To make a fair comparison, all the models are tested under the aforementioned 15 corruption types.

Semantic Segmentation. We use MMSegmentation [18] as our test bed. We take the ImageNet pretrained ViT-B and AbSViT-B and finetune them on semantic segmentation on PASCAL VOC, Cityscapes, and ADE20K. For all the experiments, we use UperNet [138] as the decoder head and FCNHead as the auxiliary head. We train on 2 GPUs with a total batch size of 16, using AdamW optimizer, a learning rate of 0.00006, and weight decay of

0.01. We train for 20k, 40k, and 160k iterations for three datasets, respectively. We use image resolution of 512x512 for PASCAL VOC and ADE20K, and 512x1024 for Cityscapes.

V&L Finetuning. Following [31], the whole model contains a pretrained visual encoder, a pretrained text encoder, and a multimodal encoder to merge vision and language. We use the ImageNet pretrained ViT or AbSViT for the visual encoder, a pretrained RoBERTa for the text encoder, and the multimodal encoder is trained from scratch. We use a learning rate of $1e - 5$ for visual and text encoders and $5e - 5$ for the multimodal encoder. For top-down attention, we use the [cls] token as the prior ξ . Since the text and visual tokens are not aligned initially, we train a linear transform to project the text tokens into the same space as the visual tokens. This is trained by the prior loss, which is set as a CLIP-style loss (Equation (2.15)) to align the text and visual tokens.