Top-Down Visual Attention from Analysis by Synthesis

Baifeng Shi
UC Berkeley

Trevor Darrell
UC Berkeley

Xin Wang
Microsoft Research

Abstract

Current attention algorithms (e.g., self-attention) are stimulus-driven and highlight all the salient objects in an image. However, 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 this paper, we consider top-down attention from a classic Analysis-by-Synthesis (AbS) perspective of vision. Prior work 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. Project page: https://sites.google.com/view/absvit.

1. Introduction

Human visual attention is often *task-guided*, *i.e.*, we tend to focus on different objects when processing different tasks [7, 69]. For example, when we answer different questions about one image, we only attend to the objects that are relevant to the question (Fig. 1 (b-c)). This stands in contrast with the widely-used self-attention [17], which is completely *stimulus-driven*, *i.e.*, it highlights all the salient objects in the image without task-guided selection (Fig. 1 (a)). While the stimulus-driven bottom-up attention has shown promising results in visual representation learning [6], 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, 64, 65]. Although some algorithms of top-down attention are proposed in the literature [1, 9, 46, 64, 65], they are incompatible with self-attention-based transformers and principled and unified designs are still missing.

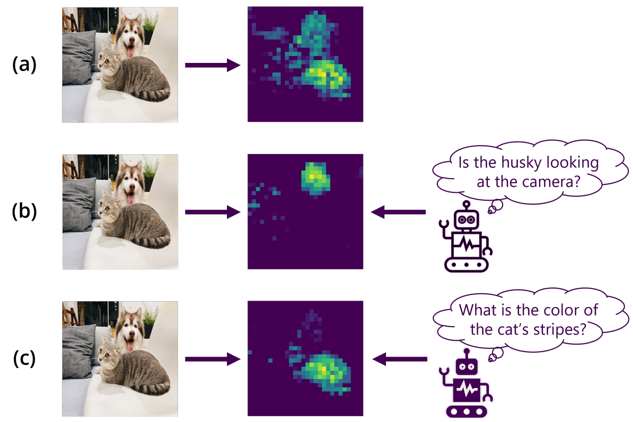

Figure 1. **Top-down *vs*. bottom-up attention**. (a) Bottom-up attention is stimulus-driven, *i.e.*, any salient objects (dog and cat) in the image may attract attention. (b-c) Top-down attention is task-guided. For example, when the task is to answer a question about a specific object, the attention will only center on that object and ignore the others. In this way, a more focused representation can be extracted for the current goal.

Previous work [5, 10, 34, 35, 50] 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 [32, 68] 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 [33] and bistable perception [55]). 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 [10, 44, 67] 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 [56], which shows that visual attention (*e.g.*, self-attention) 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 [17] 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 (Fig. 3).

When only pretrained on ImageNet [15], 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. 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 [43] and contrast responses [42, 52] in V4, and extra-classical receptive fields in V1 [2, 8, 54]. Other work tries to build a principled computational model for top-down attention [10, 44, 67].

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 [45], image captioning [65], and visual question answering [1, 64]. However, these algorithms are either incompatible with current self-attention-based models or show inferior performance, as indicated by our experiments. Other work [17, 39, 40, 66] 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*. [18] 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. Lee [34] starts from a Bayesian inference perspective and explains the top-down modulation in examples such as illusionary contours and shapes from shading. Yu and Dayan [67] focus on the top-down attention in Ponser's task [48] and build a hierarchical model where each layer corresponds to a computational step of Bayesian inference. Subsequent work [10, 50] 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. [5] 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 [27, 28] and Helmholtz Machine [14], to the following auto-encoder models such as DAE [59] and VAE [31]. Recent work [22, 57] has shown impressive results on generative unsupervised learning. Generative models can also help with supervised learning, *e.g.*, by refining object detection [36] or detecting errors in semantic segmentation [63]. Feedforward models with generative feedback are also more robust to input corruptions [29]. 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.

3. Preliminaries: Attention as Sparse Reconstruction

Shi et al. [56] show that a sparse reconstruction (SR) module functionally resembles visual attention. An SR module takes an input $\mathbf{x} \in \mathbb{R}^d$ and outputs $\mathbf{z} = \mathbf{P}\widetilde{\mathbf{u}}^*$ where $\mathbf{P} \in \mathbb{R}^{d\times d'}$ is the dictionary and $\widetilde{\mathbf{u}}^*$ is the sparse code, *i.e.*,

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

Each atom (column) of $\mathbf{P}$ contains a template pattern and each element in $\widetilde{\mathbf{u}}$ is the activation of the corresponding template. The objective is to reconstruct the input using as few templates as possible. To solve Eq. (1), one may adopt a first-order optimization [53, 56] with dynamics at time t of

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} \propto -\mathbf{u} - (\mathbf{P}^T\mathbf{P} - \mathbf{I})\widetilde{\mathbf{u}} + \mathbf{P}^T\mathbf{x}, \tag{2}$$

where the optimization is over an auxiliary variable $\mathbf{u}$ and $\widetilde{\mathbf{u}} = g_\lambda(\mathbf{u}) = sgn(\mathbf{u})(|\mathbf{u}| - \lambda)_+$ with $sgn(\cdot)$ 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 $-(\mathbf{P}^T\mathbf{P} - \mathbf{I})\widetilde{\mathbf{u}}$ 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 [16]: (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 $-(\mathbf{P}^T\mathbf{P} - \mathbf{I})\widetilde{\mathbf{u}}$ selects the object that is most activated. As shown in [56], SR modules achieve similar attention effects as self-attention (SA) [58] while being more robust against image corruptions.

Interestingly, it is also pointed out in [56] 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

$$\begin{cases} \widetilde{\mathbf{U}}^* = \arg\min_{\widetilde{\mathbf{U}}} \frac{1}{2}||\Phi(\mathbf{K})\widetilde{\mathbf{U}} - \mathbf{V}||_2^2 + \lambda||\widetilde{\mathbf{U}}||_1, & (3) \\ \mathbf{Z} = \Phi(\mathbf{Q})\widetilde{\mathbf{U}}^*, & (4) \end{cases}$$

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 [11] that approximate the softmax kernel $\Phi(\mathbf{Q})_i\Phi(\mathbf{K})_j^T \approx e^{\mathbf{Q}_i\mathbf{K}_j^T}$, $\widetilde{\mathbf{U}}^* \in \mathbb{R}^{d'\times c}$ 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 (Fig. 2 (a)).

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 (Sec. 4.1) and show that it is equivalently optimizing a sparse reconstruction objective that is modulated by a top-down signal, thus entailing top-down attention (Sec. 4.2).

4.1. 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.*, $\mathbf{z}_L \to \mathbf{z}_{L-1} \to \cdots \to \mathbf{z}_1 \to \mathbf{h}$, where $\mathbf{z}_\ell$ is the latent at ℓ-th layer. The MAP estimation is

$$\mathbf{z}_L^*, \cdots, \mathbf{z}_1^* = \arg\max_{\mathbf{z}_L,\cdots,\mathbf{z}_1} p(\mathbf{h}|\mathbf{z}_1)\cdots p(\mathbf{z}_{L-1}|\mathbf{z}_L)p(\mathbf{z}_L). \tag{5}$$

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

$$\widetilde{\mathbf{u}}_\ell \sim p(\widetilde{\mathbf{u}}_\ell|\mathbf{z}_{\ell+1}) \propto \exp\{-\frac{1}{2}||\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell - g_\ell(\mathbf{z}_{\ell+1})||_2^2 - \lambda||\widetilde{\mathbf{u}}_\ell||_1\} \tag{6}$$

$$\mathbf{z}_\ell = \mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell, \tag{7}$$

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

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

$$\widetilde{\mathbf{u}}_L^*, \cdots, \widetilde{\mathbf{u}}_1^* = \arg\max_{\widetilde{\mathbf{u}}_L,\cdots,\widetilde{\mathbf{u}}_1} p(\mathbf{h}|\widetilde{\mathbf{u}}_1)\cdots p(\widetilde{\mathbf{u}}_{L-1}|\widetilde{\mathbf{u}}_L)p(\widetilde{\mathbf{u}}_L). \tag{8}$$

Solving Eq. (8) by simple gradient ascent (of the logarithm) gives the dynamics

$$\frac{\mathrm{d}\widetilde{\mathbf{u}}_\ell}{\mathrm{d}t} \propto \nabla_{\widetilde{\mathbf{u}}_\ell} \log p(\widetilde{\mathbf{u}}_{\ell-1}|\widetilde{\mathbf{u}}_\ell) + \nabla_{\widetilde{\mathbf{u}}_\ell} \log p(\widetilde{\mathbf{u}}_\ell|\widetilde{\mathbf{u}}_{\ell+1}) \tag{9}$$

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

4.2. Top-Down Attention from AbS

From AbS (Eq. (6-9)) we can derive the dynamics of $\widetilde{\mathbf{u}}_\ell$ as

$$\frac{\mathrm{d}\widetilde{\mathbf{u}}_\ell}{\mathrm{d}t} \propto \nabla_{\widetilde{\mathbf{u}}_\ell}\left(-\frac{1}{2}||\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell - (\mathbf{x}_\ell^{bu} + \mathbf{x}_\ell^{td})||_2^2 - \lambda||\widetilde{\mathbf{u}}_\ell||_1 - r_\ell(\widetilde{\mathbf{u}}_\ell)\right) \tag{10}$$

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 $g_{\ell-1}(\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell)$, and $r_\ell(\widetilde{\mathbf{u}}_\ell) = ||g_{\ell-1}(\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell)||_2^2$ is an additional regularization. Details of the derivation are pushed back to Appendix. One may notice from Eq. (10) that, in AbS each layer is solving a similar

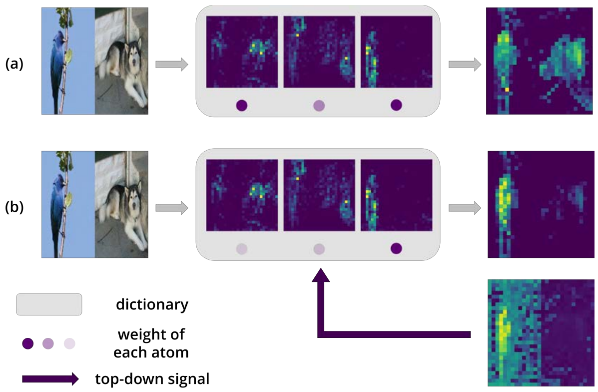

Figure 2. (a) Each atom in the dictionary contains masks for separate objects or regions. The sparse reconstruction tries to use as few masks as possible to reconstruct the input feature map, thus only the salient objects are highlighted. (b) The top-down signal $\mathbf{x}_\ell^{td}$ puts a bias on the weights of the atoms so that only the objects that agree with $\mathbf{x}_\ell^{td}$ are selected.

sparse reconstruction problem as in Eq. (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 Eq. (10) into

$$\frac{\mathrm{d}\widetilde{\mathbf{u}}_\ell}{\mathrm{d}t} \propto -\mathbf{u}_\ell - (\mathbf{P}_\ell^T\mathbf{P}_\ell - \mathbf{I})\widetilde{\mathbf{u}}_\ell + \mathbf{P}_\ell^T\mathbf{x}_\ell^{bu} + \mathbf{P}_\ell^T\mathbf{x}_\ell^{td} - \nabla r_\ell(\widetilde{\mathbf{u}}_\ell). \tag{11}$$

Comparing with Eq. (2), here $\widetilde{\mathbf{u}}_\ell$ is steered by an additional term $\mathbf{P}_\ell^T\mathbf{x}_\ell^{td}$ that acts as a bias on which atom in $\mathbf{P}_\ell$ to choose. For example, if atoms in $\mathbf{P}_\ell$ 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 (Fig. 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, Eq. (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 Sec. 5.

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 (Eq. (5)), *i.e.*, if the network has input $\mathbf{h}$ and latent representation $\mathbf{z}_\ell$ after each layer ℓ (which means $\mathbf{z}_L$ is the output), the final latent representation should approximate $\mathbf{z}_1^*, \cdots, \mathbf{z}_L^*$. Since directly solving Eq. (5) requires an iterative optimization which would be extremely costly, in this work, we adopt a variational approximation to Eq. (5). Specifically, we optimize a variational loss

$$\begin{aligned}\mathcal{L}_{var} &= -\sum_{\ell=0}^{L-1} \log p(\mathbf{z}_\ell|\mathbf{z}_{\ell+1}) - \log p(\mathbf{z}_L) \\ &= \sum_{\ell=0}^{L-1}\left(\frac{1}{2}||\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell - g_\ell(\mathbf{z}_{\ell+1})||_2^2 + \lambda||\widetilde{\mathbf{u}}_\ell||_1\right) - \log p(\mathbf{z}_L)\end{aligned} \tag{12}$$

where $\mathbf{z}_0 = \mathbf{h}$. However, as stated below, there are several caveats we need to work around when training a network with Eq. (12) in real-world tasks.

The sparsity regularization. Since the practical model we build in this work is based on self-attention (Sec. 5.1), which neither has a sparsity constraint nor solves the SR explicitly [56], we remove the sparsity regularization by setting $\lambda = 0$, which makes $-\log p(\mathbf{z}_\ell|\mathbf{z}_{\ell+1}) = \frac{1}{2}||\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell - g_\ell(\mathbf{z}_{\ell+1})||_2^2 = \frac{1}{2}||\mathbf{z}_\ell - g_\ell(\mathbf{z}_{\ell+1})||_2^2$.

Jointly training the decoder g_ℓ. Normally, optimizing Eq. (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 [31]. It is natural to use g_ℓ also as the feedback path of the network, as shown in Sec. 5.1.

Trade-off between the generative and discriminative power. The variational loss forces each $\mathbf{z}_{\ell+1}$ to be capable of generating $\mathbf{z}_\ell$. 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 $\mathbf{z}_\ell$ 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 [62]. 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. 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 Sec. 5.1.

After applying these tricks, our variational loss becomes

$$\mathcal{L}_{var} = \frac{1}{2}\sum_{\ell=0}^{L-1}||sg(\mathbf{z}_\ell) - g_\ell(sg(\mathbf{z}_{\ell+1}))||_2^2 - \log p_\xi(\mathbf{z}_L), \tag{13}$$

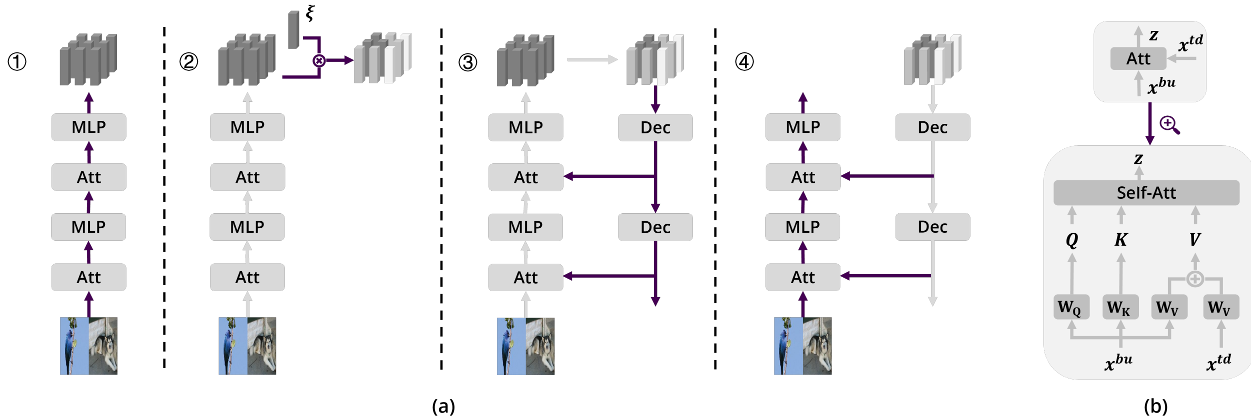

Figure 3. **Design of AbSViT**. (a) Four steps to every single inference. The operations in each step are colored as purple and others as gray. AbSViT first passes the image through the feedforward path. The output tokens are then reweighted by their similarity with the prior vector ξ and fed back through 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.

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 Sec. 5.1, we will show how to build a ViT with prior-conditioned top-down modulation and train it with Eq. (13).

5.1. AbSViT Design

Fig. 3 (a) shows the proposed AbSViT which is built upon ViT [17]. 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.*, $\mathbf{z}_L^i \rightarrow \alpha \cdot sim(\xi, \mathbf{z}_L^i) \cdot \mathbf{z}_L^i$, 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 (Eq. (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 (Eq. (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 Fig. 3 (b). We will show in Sec. 6.4 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 (Eq. (13)), *i.e.*,

$$\mathcal{L} = \frac{1}{2}\sum_{\ell=1}^{L} ||sg(\mathbf{z}_\ell) - g_\ell(sg(\mathbf{z}_{\ell+1}))||_2^2 - \log p_\xi(\mathbf{z}_L) + \mathcal{L}_{sup}, \quad (14)$$

where $\mathbf{z}_\ell$ 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 [49]:

$$p_\xi(\mathbf{z}_L) = \frac{\exp\{\xi^T \mathbf{z}_L\}}{\exp\{\xi^T \mathbf{z}_L\} + \sum_k \exp\{\xi^T \mathbf{z}_-^k\}}, \quad (15)$$

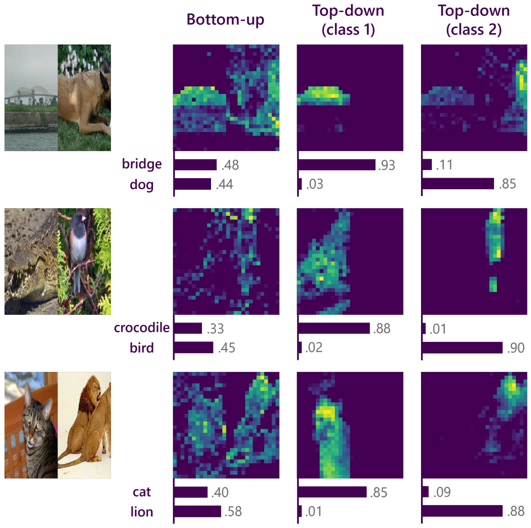

Figure 4. Controllable top-down attention in multi-object images. For each image, bottom-up attention will highlight both objects. In contrast, we can use different class prototypes as the prior to control the top-down attention to focus on different objects, and the classification result also changes accordingly.

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.*, $\nabla \log p_\xi(\mathbf{z}_L) = 0$.

6. Experiments

In this section, we first show that AbSViT achieves controllable top-down attention in multi-object scenes (Sec. 6.1). Then we test AbSViT on Vision-Language tasks such as VQA and zero-shot image retrieval (Sec. 6.2), and also on ImageNet classification and model robustness (Sec. 6.3). Finally, we analyze specific designs of AbSViT in Sec. 6.4. See Appendix for experiments on semantic segmentation.

Datasets. For VQA, we use VQAv2 [21] for training and testing and compare the attention map with human attention collected by VQA-HAT [13]. For zero-shot image retrieval, we use Flickr30K [47]. For image classification, we train and test on ImageNet-1K (IN) [15], and also test on corrupted images from IN-C [23], adversarial images from IN-A [25], and out-of-distribution images from IN-R [24] and IN-SK [60]. For semantic segmentation, we test on PASCAL VOC [20], Cityscapes [12], and ADE20K [70].

Experimental setup. We compare several baselines for goal-directed attention: (i) **PerceiverIO** [30] 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 soft mask for

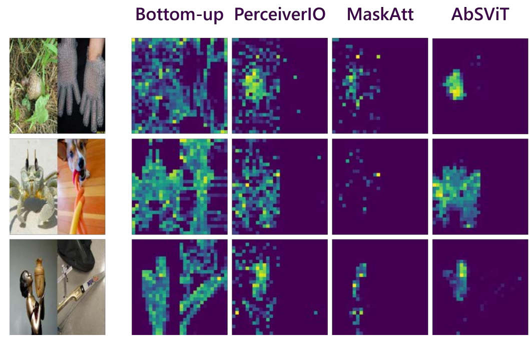

Figure 5. Comparison between different top-down attention algorithms. Prior corresponds to the left image. AbSViT has cleaner attention map than other baselines.

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 [18] framework, which contains a vision backbone, a language backbone, and a multimodal fusion module. We use ViT [17] as the vision backbone and replace it with AbSViT or the baseline models. For image classification, we try the backbones of ViT, RVT [41], and FAN [71], 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 (**??**). See the Appendix for additional implementation details.

6.1. Controllable Top-Down Attention of AbSViT

To test the top-down attention in multi-object images, we take a AbSViT pretrained on ImageNet (Sec. 6.3) 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 Fig. 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 ability to control its attention on different objects given different priors. We also compare the top-down attention of AbSViT with several baselines (Fig. 5). We can see that the attention of

Model	VQAv2		Flickr-Zero-Shot		
	test-dev	test-std	IR@1	IR@5	IR@10
BEiT-B-16 [4]	68.45	-	32.24	-	-
CLIP-B-32 [49]	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 1. Comparison of different top-down attention algorithms on VQA and zero-shot image retrieval. AbSViT achieves consistent improvements on both tasks.

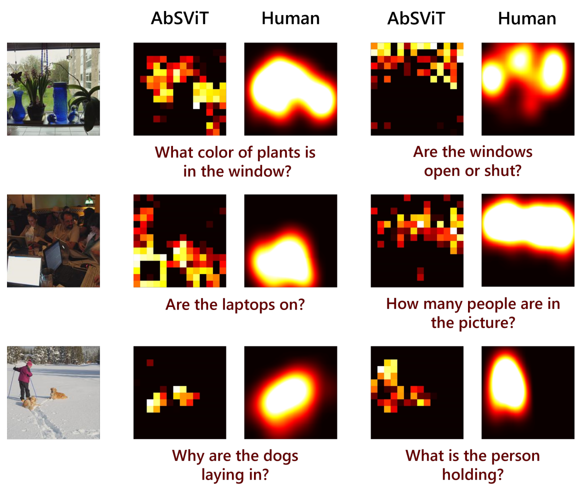

Figure 6. Comparison of attention map from AbSViT and human attention on VQA. AbSViT's attention is adjustable to different questions and is consistent with human attention.

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.

6.2. 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 (Sec. 6.3). Results are shown in Tab. 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.

Model	P/F	Clean	IN-C (↓)	IN-A	IN-SK	IN-R
PiT-Ti [26]	5/0.7	72.9	69.1	6.2	34.6	21.6
ConViT-Ti [19]	6/1.4	73.3	68.4	8.9	35.2	22.4
PVT-Ti [61]	13/1.9	75.0	79.6	7.9	33.9	21.5
GFNet-Ti [51]	8/1.3	74.6	65.9	6.3	40.4	27.0
ViT-Ti [17]	6/1.3	72.5	71.1	7.5	33.0	20.1
- AbS	7/2.6	**74.1**	**66.7**	**10.1**	**34.9**	**22.6**
RVT-Ti [41]	9/1.3	78.1	58.8	13.9	42.5	29.1
- AbS	11/2.7	**78.6**	**55.9**	**17.3**	**43.2**	**29.9**
FAN-Ti [71]	7/1.3	77.5	59.8	13.1	42.6	29.9
- AbS	9/2.9	**78.3**	**57.4**	**16.5**	**42.8**	**31.2**
PiT-S [26]	24/2.9	80.9	52.5	21.7	43.6	30.8
PVT-S [61]	25/3.8	79.9	66.9	18.0	40.1	27.2
Swin-T [37]	28/4.5	81.2	62.0	21.6	41.3	29.1
ConvNext-T [38]	29/4.5	82.1	53.2	24.2	47.2	33.8
ViT-S [17]	22/4.2	80.1	54.6	19.2	41.9	28.9
- AbS	26/9.8	**80.7**	**51.6**	**24.3**	**43.1**	**30.2**
RVT-S [41]	22/4.3	81.9	50.5	26.0	47.0	34.5
- AbS	26/10.4	81.9	**48.7**	**31.1**	**48.5**	**35.6**
FAN-S [71]	28/5.3	82.8	49.1	29.3	47.4	35.6
- AbS	32/11.4	**83.0**	**47.4**	**34.0**	**48.3**	**36.4**
PiT-B [26]	74/12.5	82.4	48.2	33.9	43.7	32.3
PVT-L [61]	61/9.8	81.7	59.8	26.6	42.7	30.2
Swin-B [37]	88/15.4	83.4	54.4	35.8	46.6	32.4
ConvNext-B [38]	89/15.4	83.8	46.8	36.7	51.3	38.2
ViT-B [17]	87/17.2	80.8	49.3	25.2	**43.3**	31.6
- AbS	99/38.9	**81.0**	**48.3**	**28.2**	42.9	31.7
RVT-B [41]	86/17.7	80.9	52.1	26.6	**39.6**	26.1
- AbS	100/39.5	80.9	**51.7**	**28.5**	39.3	26.0
FAN-B [71]	54/10.4	83.5	45.0	33.2	51.4	39.3
- AbS	62/21.8	**83.7**	**44.1**	**38.4**	**52.0**	**39.8**

Table 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.

On zero-shot image retrieval, we have a similar observation that AbSViT 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 Fig. 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 [49] pretraining.

6.3. Image Classification and Robustness

We test AbSViT on ImageNet classification and robustness benchmarks (Tab. 2). We report mCE (lower the better) [23] 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

Model	Clean	IN-C (↓)	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 3. Comparison of different top-down attention algorithms on ImageNet classification and robustness.

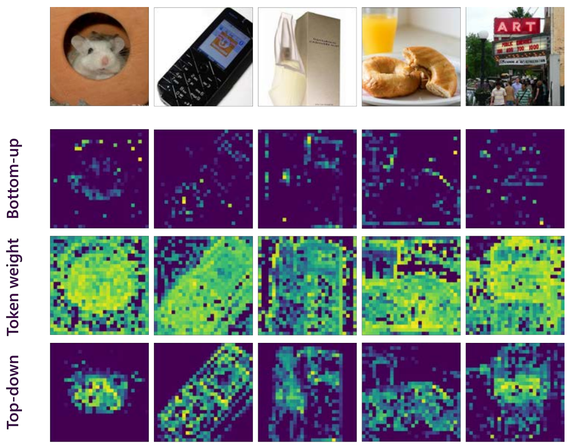

Figure 7. Visualization of the bottom-up attention, token weights, and the top-down attention in AbSViT. The bottom-up attention is noisy and fails to detect the complete foreground object. In AbSViT, the query mask can coarsely detect the foreground object and reweight tokens fed back to direct the top-down attention to better extract the foreground object.

the same 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.

Fig. 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 Tab. 3. One may see that a pure feedback model already improves the clean accuracy and robustness, and AbSViT further boosts the performance

Model	Clean	IN-C (↓)	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 4. The predicted design of top-down self-attention (AbSViT) is better than an arbitrary design (AbSViT-QKV).

by better extracting the foreground object. Due to a similar reason, PerceiverIO without feedback also slightly improves the 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.

6.4. 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 (Sec. 5.1). 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. We show advantages of these “destined” designs compared with arbitrary designs, which also justifies the proposed guiding principle of AbS. We discuss the design of top-down self-attention in this section, and push the ablation on variational loss to Appendix.

We 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 (Tab. 4), and we can see that AbSViT is superior to AbSViT-QKV on every benchmark. This is consistent with our analysis in Sec. 4.2 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.

7. 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.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086, 2018. 1, 2

[2] Alessandra Angelucci, Jonathan B Levitt, Emma JS Walton, Jean-Michel Hupe, Jean Bullier, and Jennifer S Lund. Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience*, 22(19):8633–8646, 2002. 2

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 7

[5] 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*, volume 26, pages 1529–1535, 2012. 1, 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[7] Marisa Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525, 2011. 1

[8] 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. *Journal of neurophysiology*, 88(5):2530–2546, 2002. 2

[9] Gang Chen. Where to look: A unified attention model for visual recognition with reinforcement learning. *arXiv preprint arXiv:2111.07169*, 2021. 1

[10] Sharat Chikkerur, Thomas Serre, Cheston Tan, and Tomaso Poggio. What and where: A bayesian inference theory of attention. *Vision research*, 50(22):2233–2247, 2010. 1, 2

[11] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. 3

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[13] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163: 90–100, 2017. 6

[14] Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995. 2

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 6

[16] Robert Desimone, John Duncan, et al. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18 (1):193–222, 1995. 3

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 5, 6, 7

[18] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, 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*, pages 18166–18176, 2022. 2, 6

[19] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 7

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 6

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, pages 6904–6913, 2017. 6

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[23] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 6, 7

[24] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, 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*, pages 8340–8349, 2021. 6

[25] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 6

[26] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11936–11945, 2021. 7

[27] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313 (5786):504–507, 2006. 2

[28] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006. 2

[29] Yujia Huang, James Gornet, Sihui Dai, Zhiding Yu, Tan Nguyen, Doris Tsao, and Anima Anandkumar. Neural networks with recurrent generative feedback. *Advances in Neural Information Processing Systems*, 33:535–545, 2020. 2

[30] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 6

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 4

[32] David C Knill and Whitman Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996. 1

[33] T Sing Lee. Analysis and synthesis of visual images in the brain: evidence for pattern theory. *IMA VOLUMES IN MATHEMATICS AND ITS APPLICATIONS*, 133:87–106, 2003. 1

[34] Tai Sing Lee. Top-down influence in early visual processing: a bayesian perspective. *Physiology & behavior*, 77(4-5):645–650, 2002. 1, 2

[35] Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003. 1

[36] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 7

[38] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 7

[39] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016. 2

[40] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 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*, pages 375–383, 2017. 2

[41] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12042–12051, 2022. 6, 7

[42] Julio C Martınez-Trujillo and Stefan Treue. Attentional modulation strength in cortical area mt depends on stimulus contrast. *Neuron*, 35(2):365–370, 2002. 2

[43] Carrie J McAdams and John HR Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area v4. *Journal of Neuroscience*, 19(1): 431–441, 1999. 2

[44] M Berk Mirza, Rick A Adams, Karl Friston, and Thomas Parr. Introducing a bayesian model of selective attention based on active inference. *Scientific reports*, 9(1):1–22, 2019. 1, 2

[45] Aude Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. Top-down control of visual attention in object detection. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 1, pages I–253. IEEE, 2003. 2

[46] Bo Pang, Yizhuo Li, Jiefeng Li, Muchen Li, Hanwen Cao, and Cewu Lu. Tdaf: Top-down attention framework for vision tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2384–2392, 2021. 1

[47] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 6

[48] Michael I Posner. Orienting of attention. *Quarterly journal of experimental psychology*, 32(1):3–25, 1980. 2

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5, 7

[50] Rajesh PN Rao. Bayesian inference and attentional modulation in the visual cortex. *Neuroreport*, 16(16):1843–1848, 2005. 1, 2

[51] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021. 7

[52] John H Reynolds and David J Heeger. The normalization model of attention. *Neuron*, 61(2):168–185, 2009. 2

[53] Christopher J Rozell, Don H Johnson, Richard G Baraniuk, and Bruno A Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural computation*, 20 (10):2526–2563, 2008. 3

[54] Michael P Sceniak, Michael J Hawken, and Robert Shapley. Visual spatial characterization of macaque v1 neurons. *Journal of neurophysiology*, 85(5):1873–1887, 2001. 2

[55] Paul R Schrater and Rashmi Sundareswara. Theory and dynamics of perceptual bistability. *Advances in neural information processing systems*, 19, 2006. 1

[56] Baifeng Shi, Yale Song, Neel Joshi, Trevor Darrell, and Xin Wang. Visual attention emerges from recurrent sparse reconstruction. *arXiv preprint arXiv:2204.10962*, 2022. 2, 3, 4

[57] Shengbang Tong, Xili Dai, Yubei Chen, Mingyang Li, Zengyi Li, Brent Yi, Yann LeCun, and Yi Ma. Unsupervised learning of structured representations via closed-loop transcription. *arXiv preprint arXiv:2210.16782*, 2022. 2

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[59] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 2

[60] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[61] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 7

[62] Mike Wu, Kristy Choi, Noah Goodman, and Stefano Ermon. Meta-amortized variational inference and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6404–6412, 2020. 4

[63] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *European Conference on Computer Vision*, pages 145–161. Springer, 2020. 2

[64] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European conference on computer vision*, pages 451–466. Springer, 2016. 1, 2

[65] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 1, 2

[66] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016. 2

[67] Angela J Yu and Peter Dayan. Inference, attention, and decision in a bayesian neural architecture. *Advances in neural information processing systems*, 17, 2004. 1, 2

[68] Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7): 301–308, 2006. 1

[69] Li Zhaoping. *Understanding vision: theory, models, and data*. OUP Oxford, 2014. 1

[70] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 6

[71] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022. 6, 7