Top-Down Visual Attention from Analysis by Synthesis: Appendix

Baifeng Shi
UC Berkeley

Trevor Darrell
UC Berkeley

Xin Wang
Microsoft Research

1. Derivation of Eq. (10)

From Eq. (6-7) we have

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

Then Eq. (10) is derived by

$$\begin{aligned}
\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}) \\
&= -\nabla_{\widetilde{\mathbf{u}}_\ell}\frac{1}{2}||\mathbf{P}_{\ell-1}\widetilde{\mathbf{u}}_{\ell-1} - g_{\ell-1}(\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell)||_2^2 - \nabla_{\widetilde{\mathbf{u}}_\ell}\frac{1}{2}||\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell - g_\ell(\mathbf{P}_{\ell+1}\widetilde{\mathbf{u}}_{\ell+1})||_2^2 - \nabla_{\widetilde{\mathbf{u}}_\ell}\lambda||\widetilde{\mathbf{u}}_\ell||_1 \\
&= \mathbf{P}_\ell^T\mathbf{J}_{\ell-1}^T\left(\mathbf{P}_{\ell-1}\widetilde{\mathbf{u}}_{\ell-1} - g_{\ell-1}(\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell)\right) - \mathbf{P}_\ell^T\left(\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell - g_\ell(\mathbf{P}_{\ell+1}\widetilde{\mathbf{u}}_{\ell+1})\right) - \nabla_{\widetilde{\mathbf{u}}_\ell}\lambda||\widetilde{\mathbf{u}}_\ell||_1 \\
&= -\mathbf{P}_\ell^T\left(\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell - g_\ell(\mathbf{P}_{\ell+1}\widetilde{\mathbf{u}}_{\ell+1}) - \mathbf{J}_{\ell-1}^T\mathbf{P}_{\ell-1}\widetilde{\mathbf{u}}_{\ell-1}\right) - \nabla_{\widetilde{\mathbf{u}}_\ell}\lambda||\widetilde{\mathbf{u}}_\ell||_1 - \mathbf{P}_\ell^T\mathbf{J}_{\ell-1}^T g_{\ell-1}(\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell) \\
&= -\nabla_{\widetilde{\mathbf{u}}_\ell}\frac{1}{2}||\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell - g_\ell(\mathbf{z}_{\ell+1}) - \mathbf{J}_{\ell-1}^T\mathbf{z}_{\ell-1}||_2^2 - \nabla_{\widetilde{\mathbf{u}}_\ell}\lambda||\widetilde{\mathbf{u}}_\ell||_1 - \nabla_{\widetilde{\mathbf{u}}_\ell}\frac{1}{2}||g_{\ell-1}(\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell)||_2^2 \\
&= -\nabla_{\widetilde{\mathbf{u}}_\ell}\frac{1}{2}||\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell - (\mathbf{x}_\ell^{td} + \mathbf{x}_\ell^{bu})||_2^2 - \nabla_{\widetilde{\mathbf{u}}_\ell}\lambda||\widetilde{\mathbf{u}}_\ell||_1 - \nabla_{\widetilde{\mathbf{u}}_\ell}\frac{1}{2}||g_{\ell-1}(\mathbf{P}_\ell\widetilde{\mathbf{u}}_\ell)||_2^2.
\end{aligned} \tag{2}$$

We informally use ∇ for subgradients as well.

1.1. Additional Results on Semantic Segmentation

Model	PASCAL VOC	Cityscapes	ADE20K
ResNet-101 [2]	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 1. Semantic segmentation results on three datasets.

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 [8] as the segmentation head for all the backbones. Results are shown in Tab. 1. We can see that when using AbSViT as the backbone, we can achieve 1.2-2.0% improvements over the ViT baseline with approximately the same number of parameters. This indicates that AbSViT can be used as a general backbone for different vision tasks.

2. Additional Results on Natural Images

In Fig.4-5 in the paper, 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 (Fig. 1). We also compare AbSViT's top-down attention with several baseline methods (Fig. 2) and observe that AbSViT has cleaner attention maps than other methods.

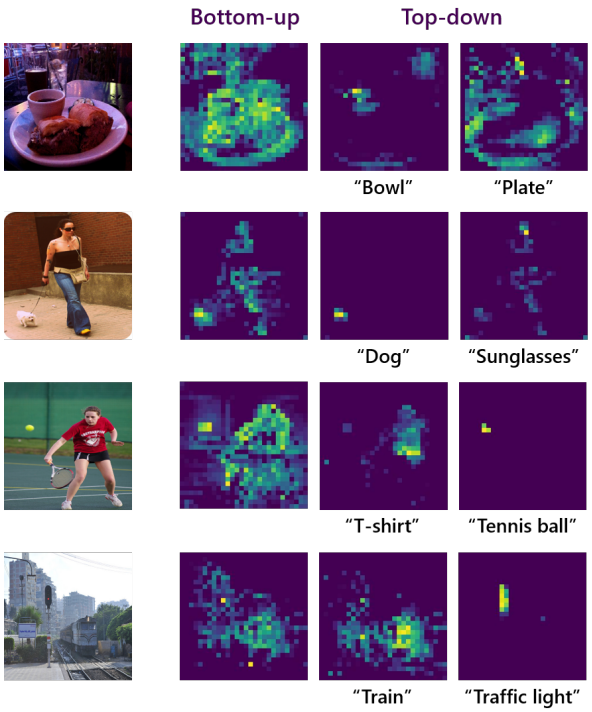

Figure 1. Visualization of top-down attention on natural images. From left to right, we show the original images, the bottom-up attention, as well as the top-down attention regarding to different objects in each image.

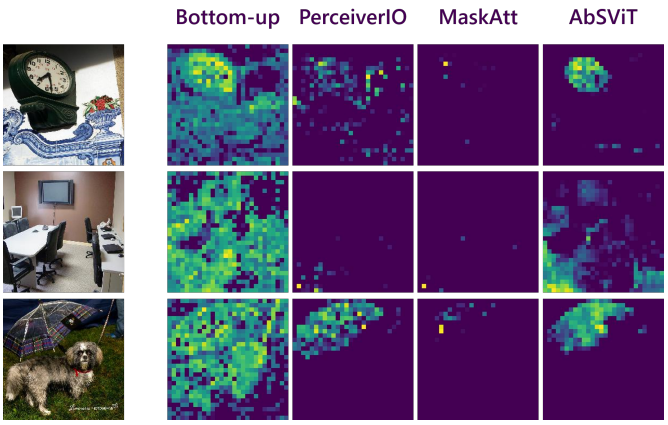

Figure 2. Comparison of top-down attention map between AbSViT and different baselines.

3. Ablation on Variational Loss

	$\mathcal{L}_{var}$	Clean	IN-C (↓)	IN-A	IN-R	IN-SK
AbSViT	✗	73.1	69.0	9.5	33.5	20.8
AbSViT	✓	**74.1**	**66.7**	**10.1**	**34.9**	**22.6**

Table 2. Ablation on the variational loss $\mathcal{L}_{var}$.

We 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 Tab. 2, adding $\mathcal{L}_{var}$ largely improves the clean accuracy and robustness. Note that, as discussed in Sec. 5.1 of the paper, 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 $\mathbf{z}_\ell$ from $\mathbf{z}_{\ell+1}$. This implies that a generative model ("synthesis") is important to high-quality top-down attention in visual recognition ("analysis").

4. Additional Implementation Details

ImageNet Pretraining. The ViT and RVT baselines as well as our AbSViT model are trained using the recipe in [6], and FAN is trained using the recipe in its original paper [11]. 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 [6] 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. [11]) 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 [2] 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 [8] 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 [3], 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 to align the text and visual tokens.

5. Limitations and Future Work

5.1. ImageNet Classification Is a Poor Teacher of Top-Down Attention

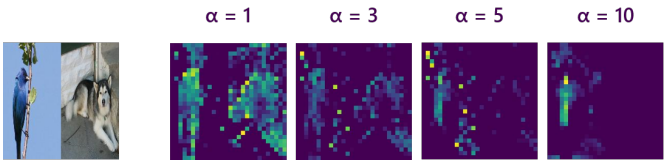

Figure 3. Visualization of top-down attention with different scaling factor α. Prior corresponds to the bird. The top-down attention gets more and more biased on the bird when increasing α.

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 Fig. 3, 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 [5, 9] or vision-language pretraining [7, 10].

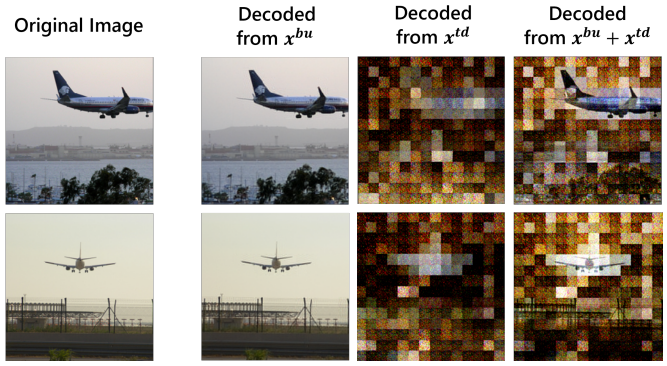

Figure 4. Examples of images decoded from the bottom-up, top-down, or the combination of bottom-up and top-down signals. The decoder can reconstruct the whole image from the bottom-up signal while failing to generate anything recognizable from the top-down signal alone. When decoding from the combination of bottom-up and top-down signals, only the foreground object is reconstructed.

5.2. How Many Syntheses Do We Need for Analysis?

In Sec. 5 of the paper, we mention that enforcing strong generative capability on the features $\mathbf{z}_\ell$ will downgrade the discriminative power regarding classification accuracy. There is a similar observation in recent self-supervised learning work [4], where reconstruction-based algorithms have worse linear-probing performance [1]. However, the empirical results in Tab. 2 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 (Fig. 4). 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 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.

References

[1] 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. 4

[2] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 1, 3

[3] 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. 3

[4] 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. 4

[5] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. *arXiv preprint arXiv:2203.08777*, 2022. 3

[6] 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. 3

[7] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. *arXiv preprint arXiv:2212.04994*, 2022. 3

[8] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 1, 3

[9] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10539–10548, 2021. 3

[10] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 3

[11] 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. 3