Supplementary Materials for

BIFRNet: A Brain-Inspired Feature Restoration DNN for Partially Occluded Image Recognition

Jiahong Zhang, Lihong Cao, Qiuxia Lai, Bingyao Li, Yunxiao Qin,

Datasets Details

Occluded-Vehicles dataset is proposed in (Wang et al. 2015) and extended in (Kortylewski et al. 2021). Images in this dataset are sampled from the PASCAL3D+ dataset (Xiang, Mottaghi, and Savarese 2014), and the occluded images are generated artificially by covering four different types of occluders (*i.e.*, white, noise, textures, and segmented objects) to the clean image. There are four levels to measure the percentage of occluded area: Level0 (0%), Level1 (20-40%), Level2 (40-60%), and Level3 (60-80%). Several example images are shown in Figure S1.

Occluded-COCO Vehicles dataset. Images in Occluded-COCO Vehicles dataset are selected from MS-COCO (Lin et al. 2014) dataset. There are also four occlusion levels (see Figure S2 for example).

Image occluders used in training. The image occluders we used in training are from the first 20 categories in ImageNet valuation set(Deng et al. 2009) and the KTH-TIPS2-a dataset(Mallikarjuna et al. 2006). Figure S3 shows a part of the images.

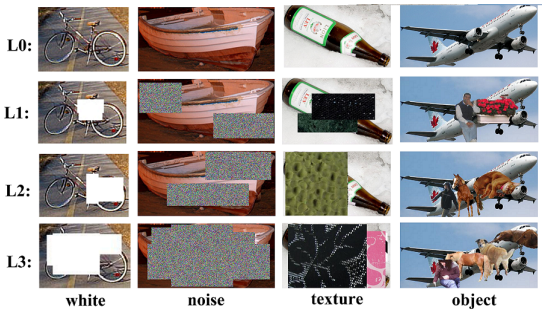


Figure S1: Images from the Occluded-Vehicles dataset.

Figure S2: Images from the Occluded-COCO Vehicles dataset.

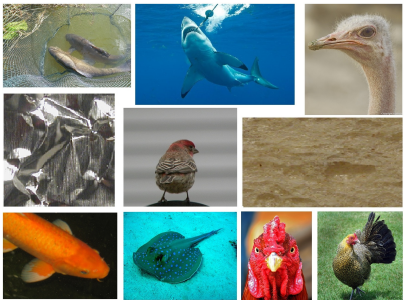


Figure S3: Occluders used in training.

Experimental Details

Model details

Since VVP is VGG16(Simonyan and Zisserman 2014), the size of the input image is 224×224. The configuration details of VVP are shown in Figure S4. For the encoder in knowledge module, the first convolution layer is with kernel size 3, stride 1, and padding 1, denoted as $(3, 1, 1)$, and the second convolution layer is set as $(1, 1, 0)$. In the completion module, convolution layers in ConvLSTM are set as $(5, 1, 2)$ and other convolution layers are $(3, 1, 1)$.

Supplementary Experiments

Discussion about α We multiply the attention loss by $\alpha = 0.1$ for better training because we observed that the attention loss is larger than the other three losses by one order in magnitude. Figure S5 the learning curves during training. We observe that the model converges faster with $\alpha = 0.1$ than $\alpha = 1$.

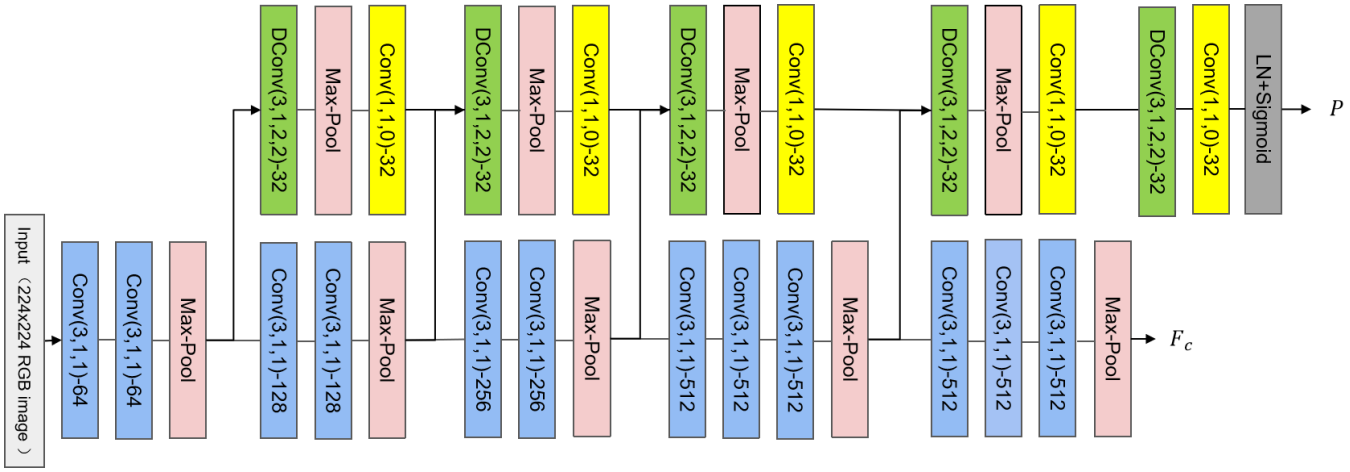
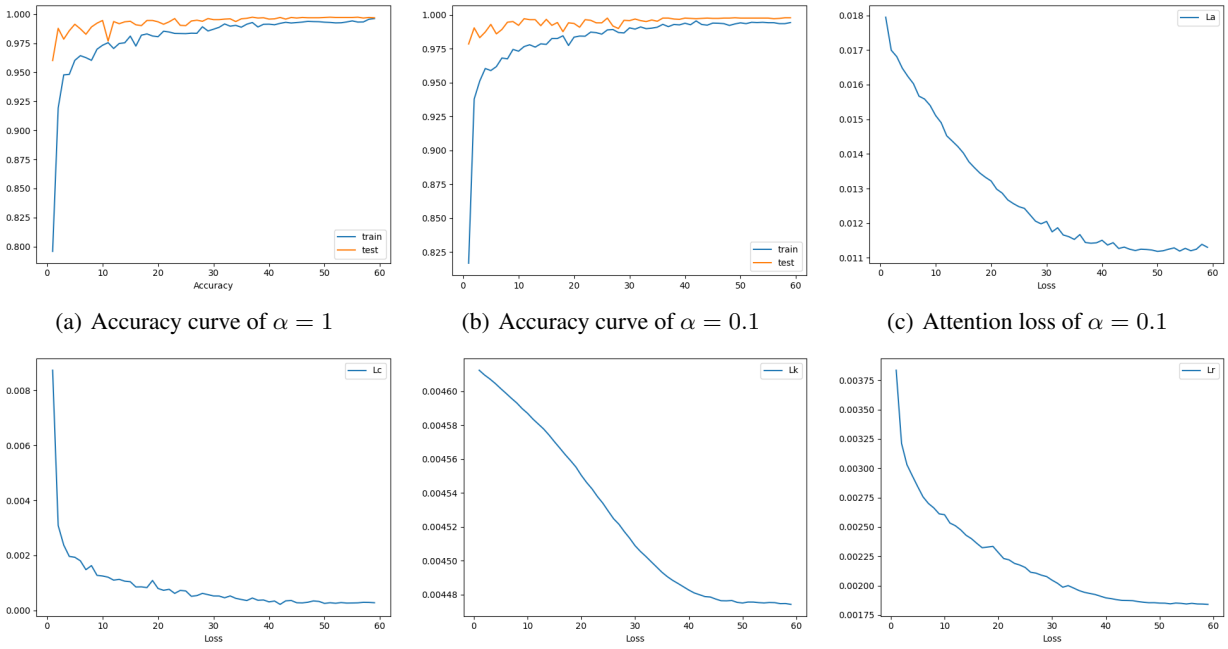


Figure S4: Configuration of the visual pathways. DConv(3,1,2,2)-32 denotes the dilated convolution layer with output channels 32, kernel size 3, stride 1, dilation 2 and padding 2. Conv(1,1,0)-32 denotes the convolution layer with output channels 32, kernel size 1, stride 1 and padding 2. The pool size of Max-Pool used in visual pathways is $(2, 2)$.



(a) Accuracy curve of $\alpha = 1$ (b) Accuracy curve of $\alpha = 0.1$ (c) Attention loss of $\alpha = 0.1$

(d) Classification loss of $\alpha = 0.1$ (e) Knowledge loss of $\alpha = 0.1$ (f) Recovery loss of $\alpha = 0.1$

Figure S5: Figures in (a) and (b) show the recognition accuracy curve of $\alpha = 1$ and $\alpha = 0.1$, respectively. (c), (d), (e) and (f) present different loss curves during training, respectively.

Study of the attention map of DVP DVP can distinguish which pixels of the image are occluded. Figure S6 visualizes some examples from the Occluded-COCO Vehicles dataset, demonstrating that DVP can tackle real-world occlusion well.

Furthermore, we present these results quantitatively by using the score distribution of the attention map. Ideally, the attention values are all < 0.5 in occluded areas and > 0.5 in non-occluded ones. We calculate the probability density function of the attention map, named as score distribution. Figure S6 shows the attention map and score distribution for real-world images from the Occluded-COCO Vehicles dataset. The ground-truth of the occlusion area is unavailable in the dataset, so we manually measured its occluded area. In Figure S6, we found that the peak values of the distributions are all < 0.5 for the occluded area and > 0.6 for the non-occluded area, which is reasonably separated.

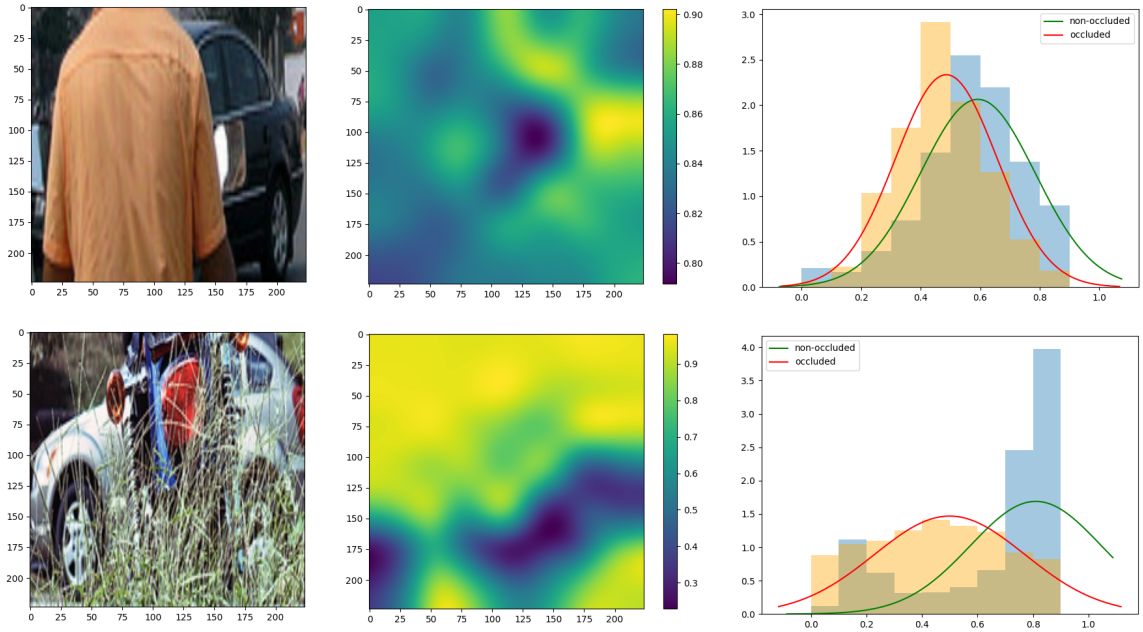


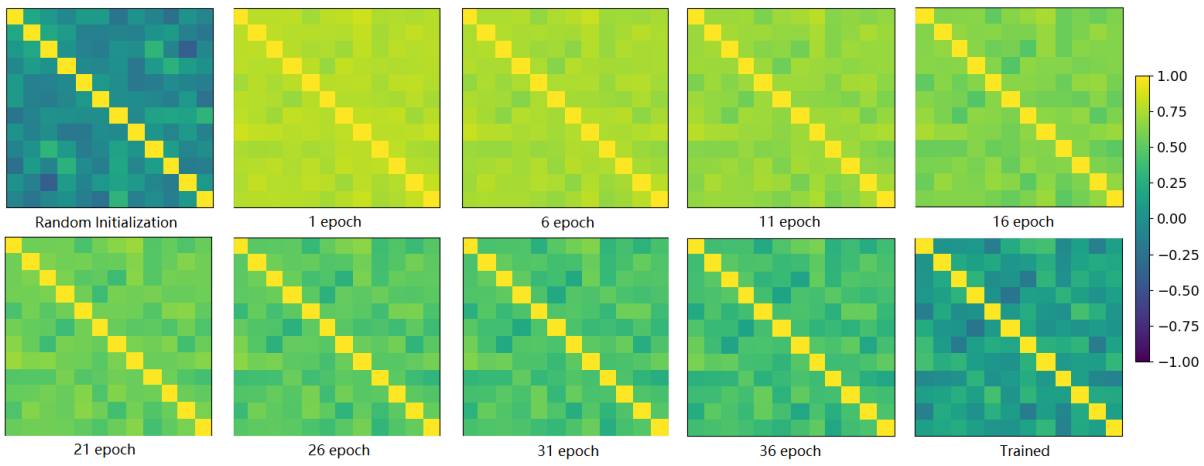
Figure S6: Visualization of the attention map and the score distribution. The left and middle columns show occluded images and attention maps from DVP. The right column presents the probability density of the attention map.

Separable representations of $\mathcal{K}$ The knowledge matrix $\mathcal{K}$ has a dimension $7 \times 7 \times 12$, representing 12 categories with a 7×7 knowledge matrix for each category. So we take a 7×7 matrix as a vector of dimension 49, and calculate the cosine distances among all 12 knowledge matrices to get the similarity matrix. We find the similarity matrix of $\mathcal{K}$ after training (Figure S7 (a), last one) shows clear separable representations among the 12 categories. However, the standard normal distribution random matrix used in initialization also has this separability. So, we study the changes in the similarity matrix during training to verify whether the separability of $\mathcal{K}$ is obtained by learning. As shown in Figure S7 (a), in the training course, the similarity matrices are obviously different from the randomly initialized matrix and gradually exhibit better separability as training continues. Furthermore, class-specific knowledge representations stabilize over training epochs, and the steady state is significantly different from the initialized value. For example, the representations of aeroplane and bicycle are shown in Figure S7 (b) and Figure S7 (c), respectively. These results illustrate that the knowledge module represented by $\mathcal{K}$ achieves good separability among different categories and stability for each category through learning.

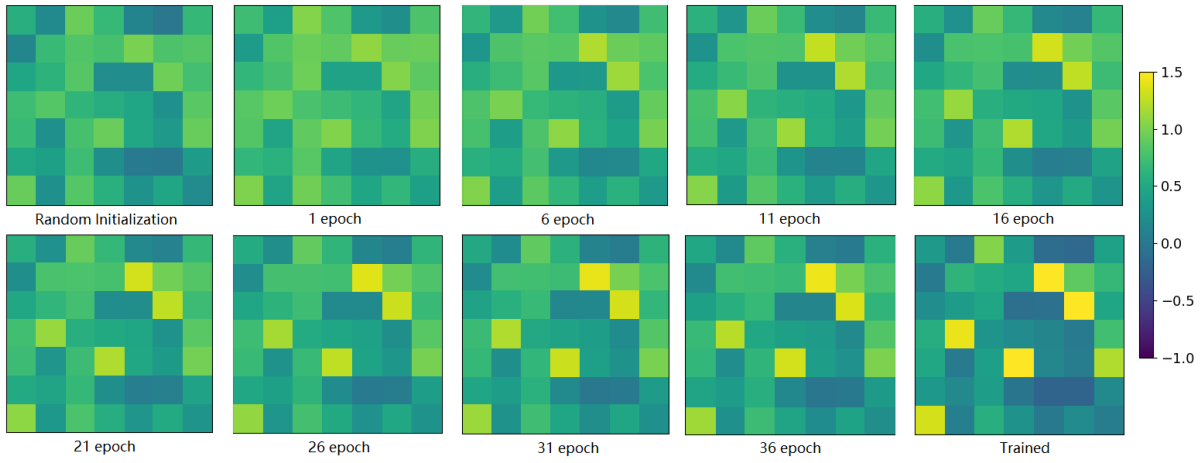
Knowledge modulated recognition It has been suggested that perception depends on the bottom-up propagation of information from sensory organs and the top-down influence of prior knowledge in cortical hierarchies (Oliva and Torralba 2007; Rauss, Schwartz, and Pourtois 2011; Gilbert and Li 2013; Summerfield and De Lange 2014; Kozunov et al. 2020). When the bottom-up sensory signal is vague or very noisy, the top-down signal, maybe from another modality, can help to make good judgment. Top-down influences in the brain are believed that higher levels of the cortex generate knowledge-based predictions about representations at lower levels (Rao and Ballard 1999; Friston 2005). We further study whether top-down signals affect the recognition results of BIFRNet through the change of knowledge matrix.

In BIFRNet, visual pathways, the encoder of the knowledge module, and the completion module pertain to the bottom-up propagation, and $\mathcal{K}$ can be considered as the top-down influence modulated by prior knowledge. We made two severely occluded images that can easily confuse human judgment. Experiments on easily confused images shown in FigureS8 confirm that BIFRNet's recognition can be affected by the top-down signal generated by prior knowledge, similar to human-like behavior.

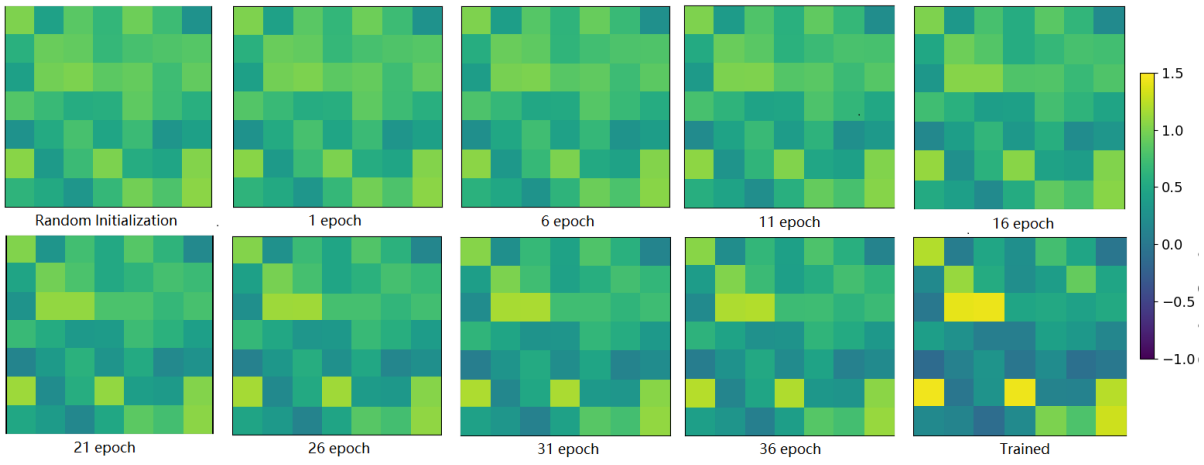
For the occluded image in FigureS8 (a), after training, the BIFRNet with trained $\mathcal{K}$ will recognize it as a bicycle and put motorbike as a secondary candidate. We then adjust $\mathcal{K}$ so that it only contains the motorbike knowledge, which models a top-down modulation signal of a motorbike. The recognition result shows a higher probability for the motorbike, and put bicycle as a secondary candidate. To further verify this modulation effect, we adjust $\mathcal{K}$ with the top-down knowledge of bicycles, and the recognition result changes back to the bicycle. When $\mathcal{K}$ provides wrong information, such as the car knowledge, the recognition result is still bicycle, again like humans will do. Moreover, BIFRNet exhibits similar properties on another example, shown in FigureS8 (b).



(a)

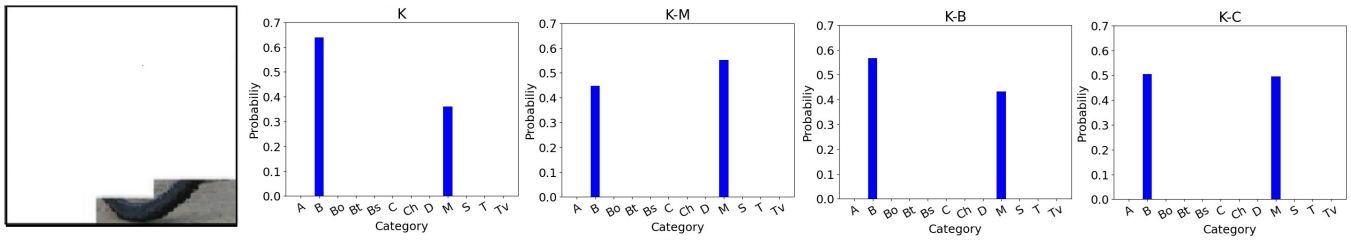


(b)

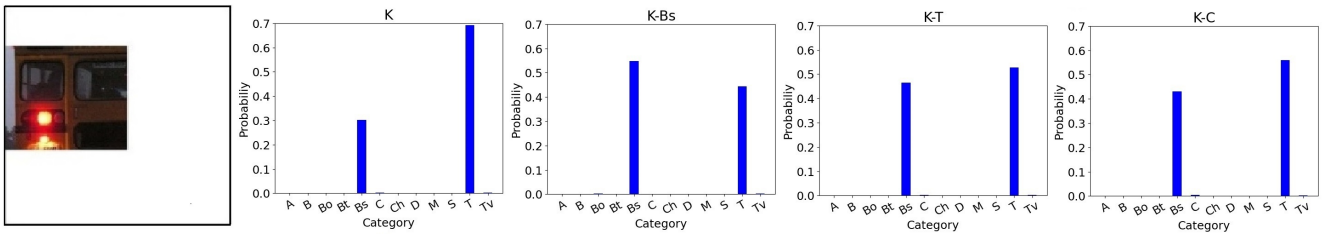


(c)

Figure S7: Figures in (a) show the variations of the similarity matrix in different training epochs. (b) and (c) shows that the representations of aeroplane and bicycle in $\mathcal{K}$ tend to stabilize over training epochs, respectively. And stable representations are significantly different from their initial values.



(a)



(b)

Figure S8: Top-down knowledge modulation effect of BIFRNet on two confusing images. (a) left image shows an occluded bicycle, possibly identified as a motorbike. The occluded object in (b) left is easily confused between bus and train. Subfigures to the right of the occlusion image show the recognition results as probably for each of 12 categories. The title of subfigure K denotes the recognition result without top-down signals. K-M denotes that $\mathcal{K}$ only contains motorbike knowledge, that is, generating a top-down recognition signal of motorbike. B, C, Bs, T refer to bicycle, car, bus and train, respectively.

References

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Friston, K. 2005. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences* 360(1456):815–836.

Gilbert, C. D., and Li, W. 2013. Top-down influences on visual processing. *Nature Reviews Neuroscience* 14(5):350–363.

Kortylewski, A.; He, J.; Liu, Q.; Cosgrove, C.; Yang, C.; and Yuille, A. L. 2021. Compositional generative networks and robustness to perceptible image changes. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 1–8.

Kozunov, V. V.; West, T. O.; Nikolaeva, A. Y.; Stroganova, T. A.; and Friston, K. J. 2020. Object recognition is enabled by an experience-dependent appraisal of visual features in the brain's value system. *Neuroimage* 221:117143.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Mallikarjuna, P.; Targhi, A. T.; Fritz, M.; Hayman, E.; Caputo, B.; and Eklundh, J.-O. 2006. The kth-tips2 database. *Computational Vision and Active Perception Laboratory, Stockholm, Sweden* 11.

Oliva, A., and Torralba, A. 2007. The role of context in object recognition. *Trends in cognitive sciences* 11(12):520–527.

Rao, R. P., and Ballard, D. H. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* 2(1):79–87.

Rauss, K.; Schwartz, S.; and Pourtois, G. 2011. Top-down effects on early visual processing in humans: A predictive coding framework. *Neuroscience & Biobehavioral Reviews* 35(5):1237–1253.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Summerfield, C., and De Lange, F. P. 2014. Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience* 15(11):745–756.

Wang, J.; Zhang, Z.; Xie, C.; Premachandran, V.; and Yuille, A. 2015. Unsupervised learning of object semantic parts from internal states of cnns by population encoding. *Computer Science*.

Xiang, Y.; Mottaghi, R.; and Savarese, S. 2014. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, 75–82. IEEE.