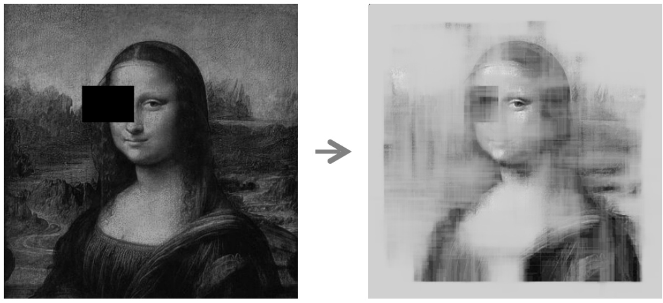
Figure-Ground Organization Emerges in a Deep Net with a Feedback Loop

Karl Zipser[1], Stella X. Yu[2], Bruno A. Olshausen[1,3]
[1] *Helen Wills Neuroscience Institute, U.C. Berkeley, CA 94709, USA*
[2] *ICSI, Computer Science Department, U.C. Berkeley, CA 94704, USA*
[3] *School of Optometry, U.C. Berkeley, CA 94720, USA*

A feedforward 'deep network' has no explicit representation of surfaces of the type proposed to characterize ‘mid-level vision’ (Nakayama, 1999); despite this, these networks do a good job of recognizing objects. This result has two important implications for biological visual processing. First, there is no clear need for a 'mid-level vision' representation of surfaces *prior* to basic-level object classification. Second, information about object classification derived in a feed-forward manner should be available, via feedback, for the development of perceptually rich representations in early visual cortex. We used a deep net to model how object-specific activation at high levels in the network could be fed back to modify representation in early levels. We first identified a subset of nodes in the uppermost hidden layer that were preferentially activated by images of people. Next, we took a degraded test image of a person and modified it recursively based on the 'person-selective' signals of this subset of nodes.



One example of a test image is the Mona Lisa with a black rectangle paced over part of the face. After sampling the 'person-selective' activation for this test image, we began a process of modifying the image by choosing a rectangular region (of random size and position) and reducing contrast in that region (tending the region slightly toward gray). After each random modification, we sampled 'person-selective' activation in the top hidden layer. If this activation became larger relative to the activation of the remaining nodes in that layer, then the modification was kept. Otherwise, it was discarded. This was repeated until 10-20 thousand modifications were accumulated. This process led to appearance modification according to learned statistics, which includes: (i) recovery of figural details in the occlusion zone, and (ii) modification of figural details in un-occluded zone according to what is consistent with object category statistics, and suppression of distractors in the background. We also tried this process with the classic ambiguous face-vase image of Rubin. When the feedback was a person-specific signal, the left and right face profiles were modified, somewhat crudely, to have person features (mouth, eyes, glasses). When instead the feedback signal was made specific for the network's initial classification of the image (which was identified as a ‘pedestal’ or a ‘vase’ by the network), an entirely different outcome occurred, the central form being modified to have shading like a pedestal.

These results indicate that feedback of object-specific information can be used to facilitate figure-ground segregation and drive low-level representation towards enhancing perceptual interpretation.