# Learning High-Level Concepts by Training A Deep Network on Eye Fixations

**Chengyao Shen**
NUS Graduate School for Integrative Science and Engineering
National University of Singapore
scyscyao@gmail.com

**Mingli Song**
College of Computer Science
Zhejiang University
brooksong@ieee.org

**Qi Zhao** *
Department of Electrical and Computer Engineering
National University of Singapore
eleqiz@nus.edu.sg

## Abstract

Visual attention is the ability to select visual stimuli that are most behaviorally relevant among the many others. It allows us to allocate our limited processing resources to the most informative part of the visual scene. In this paper, we learn general high-level concepts with the aid of selective attention in a principled unsupervised framework, where a three layer deep network is built and greedy layer-wise training is applied to learn mid- and high- level features from salient regions of images. The network is demonstrated to be able to successfully learn meaningful high-level concepts such as faces and texts in the third-layer and mid-level features like junctions, textures, and parallelism in the second-layer. Unlike pre-trained object detectors that are recently included in saliency models to predict semantic objects, the higher-level features we learned are general base features that are not restricted to one or few object categories. A saliency model built upon the learned features demonstrates its competitive predictive power in natural scenes compared with existing methods.

## 1 Introduction

Visual attention is a fundamental process of our visual system. It happens in our everyday life and allows us to bring our fovea, the high-resolution part of retina, to sample the important parts of a scene. Over the past decades, a large amount of efforts have been devoted to the research of visual attention, yet its neural mechanism remains unclear.

Early computational models of visual attention mostly follow the "Feature Integration Theory" [1, 2] and try to explain the mechanism of the attention based on low-level features such as intensity, color, orientation [3, 4]. These models work well to a certain extent, but are usually insufficient in predicting accurate eye fixations, especially when the scene contains strong semantic objects such as faces, texts, or other socially meaningful contents [5, 6].

To approach this so-called "semantic gap", improved models [6, 7, 8] have been proposed to better predict human fixations by integrating higher-level features (e.g., a common practice is to add specific object detectors) into the original low-level feature based models. However, regarding the fact that there are thousands of object categories existing in our daily life, simply adding detectors would

*Corresponding Author. Tel. (65) 6516-6658.

make the saliency models more complex and even infeasible in implementation. Hence, a unified framework that naturally integrates features at various levels is desirable.

Recent advances on deep learning and unsupervised feature learning [9, 10, 11] provide useful tools for unified feature integration. Deep learning models are usually multilayer generative networks trained to maximize the likelihood of input data with sparse priors on the responses of each layer. When exposed to natural images, hierarchies of target-relevant features with increasing complexity could be learnt in an unsupervised way and multiple levels of sparse representations can then be generated as the efficient coding of input signals. Such properties of deep learning models are attractive in that they to some extent resemble early processing stage of the primate visual system [12, 13].

In this paper, we propose a model upon the deep learning framework to learn from natural images higher-level features that normally attract attention. The main inspirations of our work are from the observation that humans tend to frequently look at semantic objects like faces, texts, animals, and cars, which are showed to be more important than other parts of the visual input. Further, recent advances in deep network on the unsupervised learning of high-level features like faces [11] pointed a promising direction of learning more general high-level feature that may be inherent in visual perception, in an unsupervised manner.

The model is built by stacking three layers of sparse coding units and pooling units together. To mimic the fixed size of image projected to the fovea during eye fixations, we train the network purely on salient regions extracted from the MIT [6] and FIFA [14] dataset. Results show that this uneven sampling based on eye fixations is the key to learn out meaningful high-level concepts. In the inference stage, full images are taken as the input of the network and a hierarchy of sparse codes are obtained according to the features learned in the training stage. Visualization and experimental results show that this model is able to encode high-level concepts like faces and texts and is competitive among existing saliency models to predict where humans look at.

The main contributions of our work are:

1. We learn meaningful high-level visual features using the principled framework of deep networks by modeling the way humans sample the visual scene.
2. We show that visual saliency plays an important role in the learning process. On one hand, it allows the learning of more general higher-level features, not restricting to a particular/predefined set of object categories; on the other, it selects the most informative part of the visual input thus greatly enhances the signal-to-noise ratio of the learning input.
3. We propose a unified feature integration framework for saliency detection that could integrate low-, mid- and high-level features in a biologically-plausible way.

The rest of the paper is organized as follows. In Section 2, we first review some related works on saliency detection and deep network. Then, we present the model of multi-layer sparse network and the way of training and testing the model in Section 3. In Section 4, experiments are conducted on FIFA and MIT dataset and quantitative results are given. Section 5 concludes the article.

## 2 Related Works

In recent years, there have been growing interests in modeling eye fixations by integrating mid-/high- level features [14, 8, 6, 15]. Cerf *et. al.* [14] refine the Itti and Koch's model [3] by adding a face detector. Zhao and Koch [8] further improve the Itti and Koch's model [3] by using a least square technique to learn the weights of face and low-level feature maps from different eye tracking datasets. In Judd *et. al.*'s work [6], low-level features including statistics of local orientations, luminance and colors, mid-level feature such as a horizon line detector, and high-level features such as face detector and a person detector are integrated by a linear SVM to predict where humans look. Based on Judd *et. al.*'s work, Lu *et. al.* [15] further improve the saliency computation by including Gestalt cues such as convexity, symmetry and surroundedness into their model. All these works indicate that mid-/high- level features play an important role in predicting human fixations, but there still lacks a unified framework that could integrate various low-, mid- and high- level features that have been mentioned or not mentioned above.

Also closely related are deep learning models that aim to learn mid-/high-level features from natural images. In one seminal work [9], Lee *et. al.* show that, by training on well-aligned images from the Caltech 101 dataset [16], hierarchies of representations which correspond to object parts and objects could be learned with a convolutional Restricted Boltzmann Machine (RBM). In [10], Zeiler *et. al.* propose a hierarchical sparse network in which each layer reconstructs the input and show that edges, junctions, and even object parts can be learned out from the images that contain objects. In one recent work[11], Le *et. al.* build a three-layer deep auto-encoder and prove that neurons representing faces, human bodies, and cats can be learned out in a fully unsupervised way on images sampled from 10 million YouTube videos. These models all validate that, by training on natural images, meaningful high-level features can be learned out using a deep network. However, none of them has considered the influence of visual attention on the feature learning in deeper levels. Furthermore, compared with existing works, our model is able to learn out meaningful high-level neurons in relatively few samples.

# 3 The Model

In this section, we describe a multilayer network that is used to learn features from salient regions. Normally the model is composed of three layers of sparse coding units and pooling units stacking together with a linear classifier at the end to read out the response of the network. This hierarchical model shares similarity with several hierarchical models that aim to model the structure of the ventral stream [17, 18, 19].

## 3.1 Sparse Coding Algorithm

Sparse coding is an unsupervised scheme that learns to represent input data using a small set of bases (or features). It is the core computational algorithm in our model.

The idea of sparse coding originates from Barlow's principle of redundancy reduction [20], which states that a useful goal of sensory coding is to transform the input in such a manner that reduces the redundancy of the input stream. In its original form of modeling image patches [21], it can be described as a generative image model as:

$$E = \|\mathbf{x} - \Phi\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1 \tag{1}$$

where $\mathbf{x}$ is the input data, $\Phi$ denotes the bases or features learnt from the data, $\mathbf{a}$ is the sparse codes for the data, and $\lambda$ is the penalty constant for sparsity. Here $\|\mathbf{a}\|_p = (\sum_m |a_m|^p)^{\frac{1}{p}}$ is called $L_p$ norm.

In (1), if we see $\mathbf{x}$ as an image, the first item $\|\mathbf{x} - \Phi\mathbf{a}\|_2^2$ can be seen as the difference between the original image and the reconstructed image and the second item $\lambda\|\mathbf{a}\|_1$ can be seen as the sparse penalty which regularizes the sparseness of the output codes. The features $\Phi$ and the sparse codes $\mathbf{a}$ can be found by iteratively minimizing the energy function:

$$\Phi = arg\min_{\Phi}\langle\min_{\mathbf{a}} E\rangle \tag{2}$$

In our model, we update sparse codes $\mathbf{a}$ with coordinate descent [22] by fixing basis $\Phi$ and updating $\Phi$ with the Lagrange dual method [23] by fixing $\mathbf{a}$.

## 3.2 Spatial Pooling

Spatial pooling is an operation that integrates the responses of nearby feature detectors into one. It is often used in image recognition models to obtain a more compact representation that preserves the important information in the input signal while discarding noises and irrelevant details.

In our model, we implement the max-pooling in the pooling layer. We use the max-pooling here mainly because of its good performance for sparse codes and simplicity in implementation [24].

Given a disjoint local neighborhood $W$ of size $l \times l$ in the sparse response maps, the max-pooling responses $\mathbf{z}$ can be obtained by:

$$\mathbf{z} = \max_{i \in W}(|a_i|) \tag{3}$$

Here $a_i$ indicates the a local neighborhood of sparse responses in $\mathbf{a}$.

After this operation, the sparse responses of the layer would shrink in a scale of $l$ and become more tolerant to minor translation and scaling.

### 3.3 Multi-layer Architecture

To model the hierarchical structure of the ventral stream, we stack three layers of sparse coding units and pooling units together to construct a hierarchical sparse coding network. The input data of the network is sampled from natural images. Before reaching the first layer, the raw data $\mathbf{x}$ is whitened with local contrast normalization to have zero mean and unit variance.

$$\mathbf{x}_1 = \frac{\mathbf{x} - \overline{\mathbf{x}}}{\max(std(\mathbf{x}), t)} \tag{4}$$

where $t$ is a small constant to avoid numerical errors. This operation approximates the visual processing in retina and LGN and is important for unsupervised learning of the first layer feature.

In the training stage, the network is trained by greedy layer-wise training method. In each layer, a large number of patches in the size of the features are extracted randomly from the input of the layer and features are learned by alternatively updating $\Phi$ and $\mathbf{a}$ according to the rule derived from sparse coding. The sparse codes of the current layer are pooled with max-pooling operation and then used as the input to the next layer. In the inference stage, full images are used as the input of the network and a hierarchy of sparse codes are obtained in a convolutional way by fixing the features learned in the training stage.

## 4 Experiments

This section reports experimental results to validate our model. We first discuss the learned higher-level features with visualization results, and then train a saliency model using the learned features and compare it quantitatively with existing models.

### 4.1 Dataset

We evaluate our model on MIT [6] and FIFA [14] datasets which contain fixations on strong semantic contents such as faces and texts. The MIT dataset [6] includes 1003 landscape and portrait images mostly in $36^\circ \times 27^\circ$ and the images in the dataset contain a variety of objects like cars, people, faces, animals, etc. These images are randomly collected from Flickr creative commons and LabelMe dataset and the fixation data were collected from 15 subjects with 3-sec-long "free-viewing". The FIFA dataset [14] contains 181 colored natural images ($28^\circ \times 21^\circ$) with fixation data. The fixation data were collected from 8 subjects with 2 second long"free-viewing" and most of the images in FIFA dataset contain faces in various with different postures.

### 4.2 Feature Learning

In order to learn mid- and high-level features from eye fixations, we collect salient regions from each image according to the eye fixation data in the MIT and FIFA datasets. Particularly we convolve a gaussian mask with accumulated fixation map from all the subjects and crop a square bounding boxes of size $150 \times 150$ centered at positions of large local maxima. This way we collect 2178 salient regions from MIT dataset and 424 salient regions from FIFA dataset and use them to train the three-layer sparse coding network with greedy layer-wise training. In each layer, each sparse coding unit would be optimized to generate a sparse output code for its input and these codes would be pooled by a max-pooling operation and then used as the input to the next layer. The parameters of the network are listed in Table. 1 and Table. 2.

### 4.3 Visualization

We validate the features learned in deep layers by visualizing their most responsive stimuli in the effective receptive field. The effective receptive field are computed by remapping one unit in the deep layer to the input pixel space. To ensure that optimal stimuli in the input space were found, we

Table 1: Parameters of the Network on MIT Dataset [6]

| | Layer 1 | Layer 2 | Layer 3 |
|---|---|---|---|
| Feature Size | $8 \times 8$ | $4 \times 4 \times 16$ | $8 \times 8 \times 200$ |
| Pooling Size | $4 \times 4$ | $4 \times 4$ | $1 \times 1$ |
| Number of Features | 16 | 200 | 200 |

Table 2: Parameters of the Network on FIFA Dataset [14]

| | Layer 1 | Layer 2 | Layer 3 |
|---|---|---|---|
| Feature Size | $6 \times 6 \times 3$ | $6 \times 6 \times 25$ | $6 \times 6 \times 100$ |
| Pooling Size | $3 \times 3$ | $3 \times 3$ | $1 \times 1$ |
| Number of Features | 25 | 100 | 225 |

traverse the whole response space of second layer and third layer for the all the full images in the datasets.

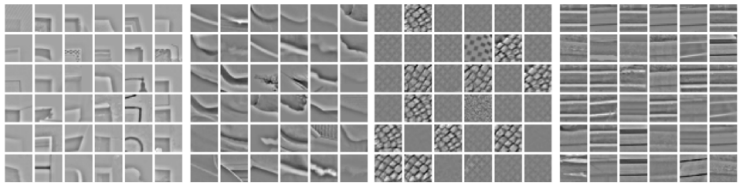

Figure 1: Illustration of top 36 stimuli of four second-layer neurons in whitened image space of MIT Dataset.

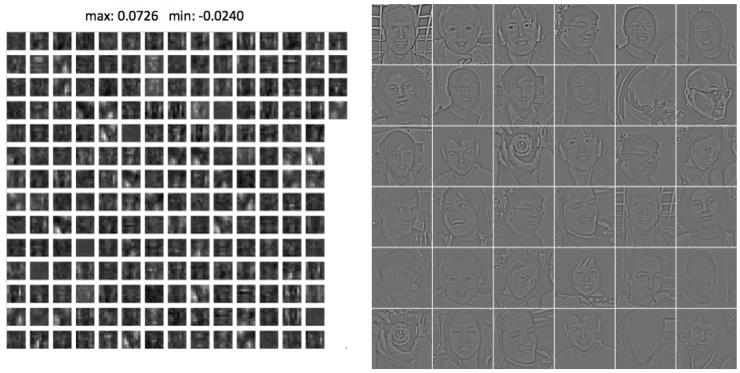


Figure 2: Illustration of the third-layer neurons that encode high-level concepts of faces by visualizing its receptive field and top 36 responsive stimuli (MIT Dataset).

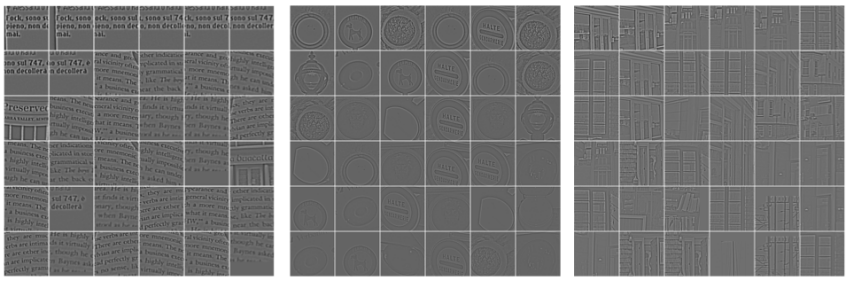

Figure 3: Illustration of three third-layer neurons that encode high-level concepts of texts, round objects, and windows (MIT Dataset).

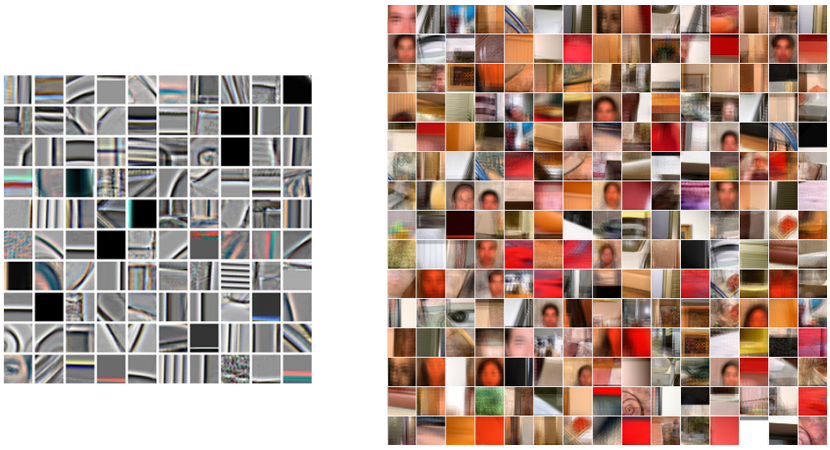

Figure 4: Average of top 36 stimuli for all the second-layer (left) and third-layer (right) neurons trained on FIFA Dataset.

Through visualization, we found that, by training on salient regions, neurons in the second-layer encode mid-level features like junctions, contours, textures, and parallelism (as shown in Figure. 1 and Figure. 4) and neurons in the third-layer are able to learn high-level concepts like faces, texts, windows, and round objects (as illustrated in Figure. 2, Figure. 3 and Figure. 4). To further verify the role of salient regions on the results of feature learning, we train the network by sampling random patches and visualize the second-level and third-level features learned. We found that without salient region sampling, the second-level neurons tend to learn features like long edges and the third-level neurons fail to learn out meaningful features after optimization.

### 4.4 Saliency Prediction

We then integrate the features learned in previous section to predict visual saliency on the two datasets. Here we take an approach similar to Judd *et. al.*[6], using a linear SVM to learn optimal weights for feature integration. To train the linear SVM, we divide the two datasets into two halves. Positive samples are collected from salient regions and negative ones are randomly sampled from non-fixated are of the training set. The saliency map is then constructed by the output value of the linear SVM on each local region:

$$s = g \circ \max(\mathbf{w}^T \mathbf{x}, 0) \tag{5}$$

Here $\mathbf{w}$ denotes the weight of the linear SVM, $\mathbf{x}$ represents the vectorized feature responses for the local region, and $g$ is a gaussian mask with a standard deviation of 1 visual degree in the input space. To compensate the boundary loss after stages of convolution, a zero-value boundary is added according to the effective receptive size of the high-level neuron. Since there is a strong bias for human fixations to be near the center of the image [6, 8], we also compare our model with a center bias modeling (i.e., adding a Gaussian mask centered in the middle of the image on the final saliency map) with that of Judd et. al.s model which also includes a distance to center channel to account for center bias. All the saliency maps are resized to the original image size in the final evaluation. It is worth emphasizing that our model does not include particular well-trained object detectors, but learn all features in an unsupervised manner.

We evaluate our model using ROC curve. The ROC curve is obtained by varying the threshold saliency map and calculating the true positive rate with respect to fixations across all subjects. The first fixation for each image is eliminated as it is always the center of the image. The ROC curve of human fixation data is also provided for comparison. This curve is computed by iterating all the subjects and averaging the ROC Curve on whether the fixations of this subject can be predicted by the saliency map generated by the other $n - 1$ subjects.

**MIT Dataset** For the MIT Dataset, we divide it into 501 training images and 502 testing images and train a linear SVM based on the second layer responses. We then compare our algorithm with

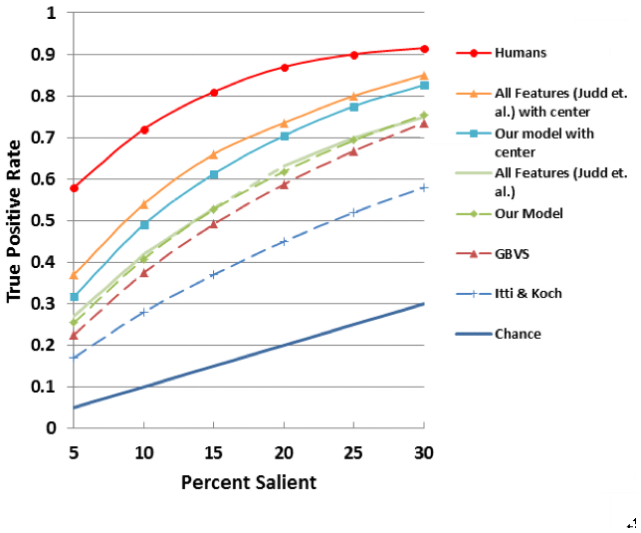


Figure 5: ROC curve of Different Saliency Models on the MIT Dataset [6].

classical saliency algorithms based on low level features [3, 4] and the benchmark algorithms on MIT dataset [6] which combines classical low-level features, mid-level features (a horizon detector) and high-level features (face and people detectors). From Figure. 5, we can see that, although we just use one layer of feature, our model outperforms the models based on low-level features, and work comparably well to the benchmark algorithm.

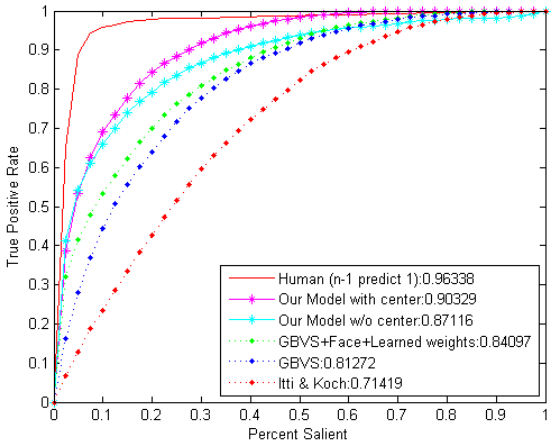


Figure 6: ROC curve of Different Saliency Models on the FIFA Dataset [14].

**FIFA Dataset** For the FIFA Dataset, we divide it into 90 training images and 91 testing images and train a linear SVM based on the third layer responses. For comparison, we also compute the saliency map using classical saliency algorithms based on low level features [3, 4] and the benchmark algorithms on FIFA dataset [8, 25] which combines low-level features with a face channels and learns the weights from data[1]. The ROC Curve and Area Under Curve (AUC) for all the algorithms are shown in Figure. 6. From Figure. 6, we can see that, our model outperforms all the previous models and is close to the human fixation data.

[1]We use 0.027 for color, 0.024 for intensity, 0.222 for orientation and 0.727 for face channel according to Table 1 in [25]

## 5 Conclusion

This paper presents a novel algorithm to effectively learn base feature at various levels by training the deep network on salient regions. A saliency model based on the deep learning framework is further proposed and demonstrated to be competitive and promising in predicting where people look at. As far as we know, this model is the first saliency model that attempts to utilize hierarchies of features learned directly from natural images and naturally integrate these features to tackle the problem of object/social saliency. Without pre-trained detectors designed for specific object detection, this model can still perform competitively on dataset with a lot of semantic content. Results demonstrate that, through unsupervised learning, it is possible to learn semantic-related features with a hierarchical architecture and link them with saliency by a simple linear classifier.

## 6 Acknowledgement

The authors would like to thank Yen Shih-Cheng and Jiang Ming for helpful advices on the work. This work is partially funded by the Sensor-enhanced Social Media Centre (SeSaMe) and the start-up grant at ECE of NUS (No.R-263-000-648-133).

## References

[1] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[2] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Hum. Neurobiol.*, vol. 4, no. 4, pp. 219–27, 1985.

[3] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10-12, pp. 1489–1506, 2000.

[4] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in neural information processing systems*, vol. 19, p. 545, 2007.

[5] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 14, 2008.

[6] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 2106–2113, IEEE, 2009.

[7] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *Advances in neural information processing systems*, vol. 20, 2008.

[8] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of Vision*, vol. 11, no. 3, 2011.

[9] H. Lee, R. Grosse, R. Ranganath, and A. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 609–616, ACM, 2009.

[10] M. Zeiler, G. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2018–2025, IEEE, 2011.

[11] Q. Le, R. Monga, M. Devin, G. Corrado, K. Chen, M. Ranzato, J. Dean, and A. Ng, "Building high-level features using large scale unsupervised learning," *Arxiv preprint arXiv:1112.6209*, 2011.

[12] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio, "A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex," tech. rep., DTIC Document, 2005.

[13] Y. Bengio, "Learning deep architectures for ai," *Foundations and Trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[14] M. Cerf, E. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *Journal of Vision*, vol. 9, no. 12, 2009.

[15] Y. Lu, W. Zhang, C. Jin, and X. Xue, "Learning attention map from images," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1067–1074, IEEE, 2012.

[16] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[17] M. Riesenhuber, T. Poggio, *et al.*, "Hierarchical models of object recognition in cortex," *Nature neuroscience*, vol. 2, pp. 1019–1025, 1999.

[18] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[19] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[20] H. Barlow, "Possible principles underlying the transformation of sensory messages," *Sensory communication*, pp. 217–234, 1961.

[21] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[22] J. Yang and M. Yang, "Learning hierarchical image representation with sparsity, saliency and locality," in *Jesse Hoey, Stephen McKenna and Emanuele Trucco, Proceedings of the British Machine Vision Conference, pages*, pp. 19–1.

[23] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," *Advances in neural information processing systems*, vol. 19, p. 801, 2007.

[24] Y. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *International Conference on Machine Learning*, pp. 111–118, 2010.

[25] Q. Zhao and C. Koch, "Learning visual saliency," in *Information Sciences and Systems (CISS), 2011 45th Annual Conference on*, pp. 1–6, IEEE, 2011.