Predicting Visual Saliency: Where Do People Look?

Austin Le
Princeton University
austinle@princeton.edu

Abstract

The ability to understand and predict where humans naturally fixate their eyes in images is a valuable component towards advancing computer vision systems. In recent years, the problem of predicting visual saliency has advanced from working with low-level hand-crafted features to data-driven approaches via neural networks, given their recent successful advances in capturing and understanding high-level information. Further, recent publications have made available saliency datasets that are sufficiently large enough to meaningful train and test end-to-end neural networks.

This literature survey will first introduce and explore the visual saliency task, Then, we will survey some of the most relevant methods and datasets in the research area, including seminal methods from the early years as well as the most recent state-of-the-art.

1. Introduction

One of the many challenges we face in computer vision is to be able to mimic human levels of visual understanding. Humans outperform computer counterparts when it comes to looking at an image, identifying its most important parts, and piecing together the story conveyed by the image. In short, the human visual system is incredibly fast and reliable at detecting visual saliency, but computer vision systems still struggle to model this same high-level understanding at the same speed and level of accuracy. In recent years, the high-level visual task of predicting saliency has risen to much greater interest within the computer vision community, as we begin to tackle increasingly harder challenges.

1.1. What is visual saliency?

At any moment in time, the human visual system receives an enormous amount of information that it must quickly filter through, process, and communicate to the rest of the body in a never-ending loop. Usually, we can process only a small fraction of this information, with the most important bits actually presented to our conscious awareness. How does the human system go about the selection process of what visual information is important enough to be further processed? In more colloquial terms, what are the most interesting and important parts of the image? Where are human eyes most likely to fixate? In visual saliency, the question is more about *where*, and less so about *what*.

Figure 1. Challenging scenes for saliency prediction: woman lying in a flower field

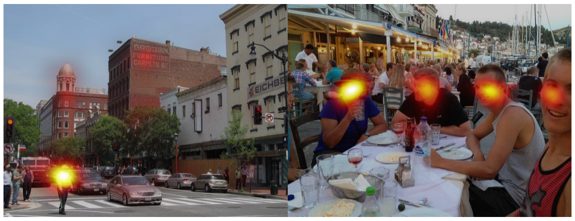


Figure 2. Two example saliency heatmaps overlaid on their original source images from the SALICON dataset. Notice that while there are many different objects in the scene that could be interesting, only some major locations stand out as being the most eye-grabbing.

Although this task slightly de-emphasizes the importance of an object (*where* rather than *what*), regions of high saliency within an image still tend to be around objects. In other words, a prior on detected objects can often be helpful when predicting saliency. An object's saliency in the context of an image is a combination of multiple factors, some of which include local and global contrast, overall spatial distribution, how ”in focus” the object is, its background-

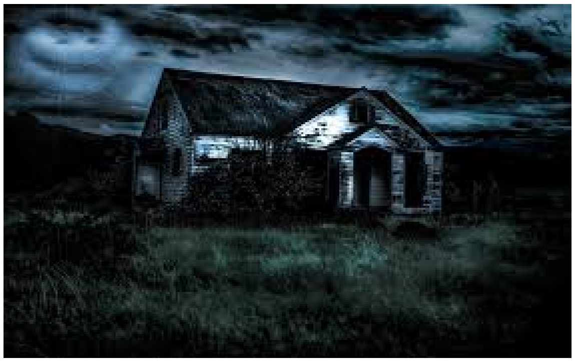


Figure 3. Challenging scenes for saliency prediction: house in low-contrast environment

edness, and how much it is affected by central bias. These properties often appear when examining human visual attention, and therefore often serve as a basis for building systems for predicting visual sailency.

Thus, visual saliency is fundamentally related to the process of selective attention in the human visual system. The objective of a saliency prediction system is to approximate which parts of an image tend to attract human visual attention, which correspond to where human eyes would likely fixate when viewing the image. The predicted information is then aggregated into a single saliency map over the original image space, where the values in the heatmap correspond to how salient that pixel is (i.e. how likely it is to capture the human eye's attention). Figure 2 illustrates what a ground-truth saliency heatmap might look like for a given image, and is what saliency prediction systems strive to reproduce.

1.2. Why is predicting saliency difficult?

As with most computer vision tasks, early and simple models usually revolve around low-level features, such as contrast, edges, and color. For instance, contrast could be used to generally detect objects, which might be more likely to be associated with a region with higher relative salience than the rest of the image.

However, it has been shown that the task of predicting visual saliency requires far more than simple low-level features to perform reasonably well at all. For example, the house in Figure 3 does not particularly stand out from its background from the perspective of low-level contrast. The house could easily be interpreted as part of the background or otherwise dismissed as not salient, even though it is semantically salient in a higher-level human cognition. This is an example scenario in which traditional prediction systems that rely on low-level saliency cues would fail. Thus, an effective image saliency predictor should be able to learn and understand that not all detected objects are worthy of attention. Even within this, the rules of what does and what does not deserve attention are highly complex and can be considered a high-level visual understanding task.

The recent trend in visual recognition tasks has been that computer systems are being continuously pushed towards learning increasingly higher-level and more complicated features. These include tasks such as object classification, video activity recognition, and visual question-answering. Much like these sister tasks, the visual saliency prediction task also demands higher-level understanding in order to accurately learn and predict the salient regions of an image.

Many of the state-of-the-art solutions to these challenges leverage the ability of existing methods to recognize and classify objects. However, this alone is not sufficient to determine which locations in an image are the most attention-grabbing for a visual saliency system. In any given image, there may be hundreds of detected objects, and while relative size could be a signal to lend greater weight to a region's saliency, it does not necessarily accurately represent human visual attention. For example, consider a scene with a repeated pattern of background objects and a major foreground object, such as a woman lying in a flower field in Figure 1. The repeated background objects can be detected, but the ground-truth human visual system would pay very little attention to these objects, if at all. Instead, a human would likely focus most, if not all, of their attention on the woman directly.

Unfortunately, the process of accurately and efficiently training systems for predicting saliency is challenging and tedious compared to more traditional computer vision problems like classification and segmentation. Firstly, training data is difficult and costly to obtain because it requires eye-tracking data from human observers, which is much more involved than simply labelling or segmenting objects in an image. This challenge has recently made progress through effective methods for crowdsourcing data collection, which has led to larger, higher quality, and publicly available datasets for advancing saliency research. Secondly, saliency prediction involves predicting and assigning values for approximately every pixel for a given input image, rather than determining one or several labels globally. Furthermore, the resulting heatmap must be spatially coherent over the image space and have sensible transitions between adjacent pixels. This is overall a more complicated input-output task, but is similar in some ways to the semantic segmentation task.

As crucial as high-level understanding is to the success of a visual saliency system, we will soon see that low-level features must not be thrown out of consideration entirely. Early attempts towards visual saliency prediction began with learning from low-level features, and the information to be gained from these low-level features can still be used in conjunction with a system's learned high-level features to potentially produce even better results.

Altogether, the task of predicting human visual attention presents a relatively new and exciting challenge for computer vision research. Although it shares many similarities with more classic and familiar visual tasks, it continues to demand more advanced cognitive abilities from our computer systems, making this problem an especially worthwhile and interesting endeavor within the computer vision community.

This literature review will continue by discussing the driving motivations and potential applications of saliency research in Section 2. Then, we will briefly discuss some early approaches and models from the beginning of modern saliency research in Section 3. Moving forward, in Section 4, we will survey current methods and the state-of-the-art in saliency research. In Section 5, we will discuss training and evaluation by examining the most relevant datasets, benchmarks, and metrics. Finally, we conclude our survey and discuss future directions in Section 6.

2. Motivations and Applications

Applications of saliency maps widely vary across computer vision tasks. Because of its representation of human visual attention, saliency maps have been applied to attention-based recognition, detection, and segmentation tasks. Understanding how humans naturally allocate their visual processing resources can help with creating more intelligent classification of objects and implicitly provides an underlying ranking of categories.

Saliency information can also be applied towards context-aware image and video compression and resizing, as well as other image processing tasks such as automatic cropping. In a similar vein, saliency prediction has recently been applied to new, emerging visual mediums such as panoramic and 360 degree images, videos, and virtual reality environments. In 360 degree video, saliency maps can be applied to the entire large frame to determine which parts would be the most interesting to a viewer. Using this information, a system can plan a standard field-of-view virtual camera path through the 360 degree video and produce a standard field-of-view camera that can be viewed normally. Similar systems can be applied to panoramic images and videos as well.

Saliency prediction has been further explored in the context of virtual reality environments [18], since users are no longer limited to just eye movement. Instead, users have the freedom to significantly move a combination of their heads and their eyes to explore the environment, which provides new information and ground truth that can be used to learn more accurate predictors.

In the captioning task, saliency predictors can help with clarifying and improving otherwise ambiguous captions, by verifying the existence of the salient objects or actions and by providing an implicit ranking on what parts of the scene are most important to include in the caption itself.

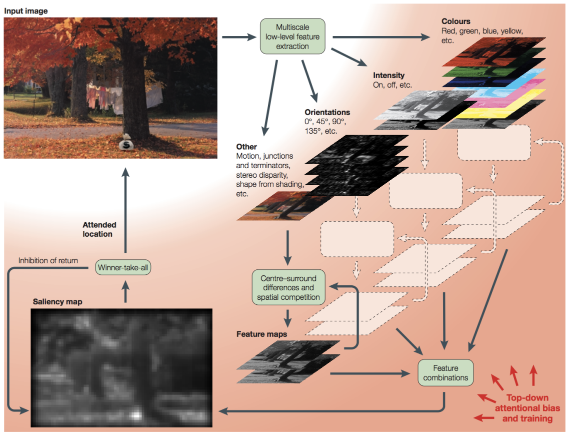

Input image
Multiscale low-level feature extraction
Colours
Red, green, blue, yellow, etc.
Intensity
On, off, etc.
Orientations
0°, 45°, 90°, 135°, etc.
Other
Motion, junctions and terminators, stereo disparity, shape from shading, etc.
Attended location
Inhibition of return
Winner-take-all
Saliency map
Centre–surround differences and spatial competition
Feature maps
Feature combinations
Top-down attentional bias and training

Figure 4. Itti and Koch's classical model for predicting a saliency map

Moving slightly away from computer vision tasks, saliency prediction has also been used to assist with the development of effective user interface designs. For instance, if an effective saliency predictor fails to identify a high level of saliency around a relatively important part of a user interface design, then this might be a hint that the current design would not be very intuitive and user-friendly to a human, suggesting that the designer consider other alternatives.

Even more interestingly, sufficiently accurate and efficient saliency predictors could be used ”full circle” by integrating them back into other computer vision systems to create more complex and powerful pipelines that potentially outperform previously existing systems in their respective domains. For example, a saliency prediction system built upon object classification could be recursively used to create an even more powerful object classifier.

3. Early Methods and Unsupervised Models

Initial approaches towards saliency prediction are motivated by fundamental concepts of how the human brain works. The human visual system processes information in two steps: a pre-attentive process, in which low-level features like edges are ”instantly” observed, followed by a complex attentive process that achieves higher-level understanding of the object(s) in the scene. From this, many traditional models for saliency are fundamentally related to feature and object detection [4, 20]. An overarching theme in these early models was that they were all unsupervised and not at all data-driven, since no real datasets of human fixation data was available at the time. As such, these models attempt to extract features and information from the image directly, sometimes in a hierarchical approach, and then produce a predicted saliency map from only the source image.

Among early works in saliency research, Itti and Koch's [6] work in computational modeling for visual attention in

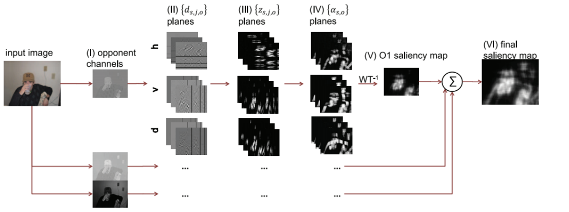

input image
(I) opponent channels
(II) $\{d_{s,j,o}\}$ planes
(III) $\{z_{s,j,o}\}$ planes
(IV) $\{\alpha_{s,o}\}$ planes
h
v
d
...
(V) O1 saliency map
WT-1
Σ
(VI) final saliency map

Figure 5. Murray *et al.*'s spatiochromatic wavelet model, SIM

2001 is perhaps the most influential. Inspired by ideas in neuroscience, Itti and Koch's classical model (illustrated in Figure 4) performed low-level hand-crafted feature extraction for components such as color, intensity, and orientation to create multiple feature maps. Using these feature maps, they create a single unique saliency map, which is then scanned by attention using a winner-takes-all network to determine the region of greatest attention within the image. This relatively low-level approach towards predicting saliency set the standard for much of the later saliency research to come.

In another biologically inspired work, Murray *et al.* [14] propose a saliency model named SIM (saliency by induction mechanisms) based on low-level spatiochromatic information. Their low-level color induction model (shown in Figure 5) decomposes each color channel into wavelet planes, computes contrast planes, which then produces induction weight planes that they combine via an inverse wavelet transform to create a saliency map for that channel. Combining all of the channels produces a final, single saliency map for the image.

However, these low-level models fall significantly short of accurately mimicking human visual attention, due to their inability to capture the high-level nuances of the human visual system. Fortunately, in recent years, convolutional neural networks have rapidly grown in popularity for computer vision tasks due to their widespread success across multiple visual challenges. Their ability to extract high-level information and learn more complex functions such as human attention has contributed to its successful application to saliency prediction as well.

4. Current Methods and the State-of-the-Art

Recent work in saliency prediction has leveraged the capacity of neural networks to achieve higher-level understanding of visual data and overcome the problems that more traditional and lower-level models experienced. In most cases, the networks are trained by formulating saliency as a problem that can be learned via end-to-end regression. Essentially all of the more recent and current methods are data-driven and result in largely supervised models. However, it is important to note that this would not be possible without the availability of sufficiently large enough datasets. Initial data-driven models and approaches work around this constraint by adapting other trained models, and more recent models leverage the introduction of even more recently available and large datasets. We will return to the topic of datasets for deep supervised learning in Section 5.

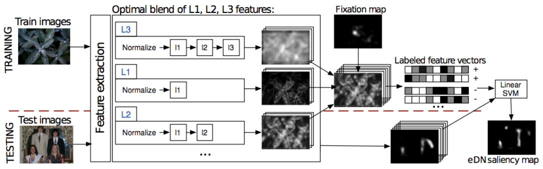


Figure 6. Vig *et al.*'s ensemble of deep networks (eDN)

4.1. Early Supervised Methods Using Neural Networks

In 2014, Vig *et al.* proposed the *eDN* (ensembles of deep networks) model [19]. *eDN* generates a large number of richly-parameterized 1 to 3 layer networks using biology-inspired hierarchical features. They use hyperparameter optimization to search for independent models that are predictive of saliency and combine them into a single model. From the model, they extract feature vectors, label them with a small-scale empirical gaze dataset, and feed them into a linear SVM classifier, which produces the final saliency map. The *eDN* pipeline (shown in Figure 6) achieved state-of-the-art results on the MIT300 saliency benchmark when it was first published in 2014.

From this early application of networks came *Deep Gaze I* [12] also in 2014, which extended the deep AlexNet architecture [9] for image recognition to the saliency prediction task. *Deep Gaze I* is one of the first applications of transfer learning for saliency prediction. It uses pre-trained parameters from AlexNet's architecture on ImageNet [3], and then further fine-tunes the network on the MIT saliency dataset. This instance of transfer learning was primarily motivated by the lack of sufficiently large saliency datasets at the time, so the authors would normally have been constrained to work with relatively shallow neural networks to avoid severe overfitting. To overcome this, they maintained a relatively shallow network with AlexNet pre-trained on ImageNet and further refined the learned weights using a subset of the MIT1003 dataset. They also made a novel observation in that images tended to have a center-bias for salient regions, so they incorporated a center-bias prior into their training. With this approach, *Deep Gaze I* was able to significantly outperform the *eDN* [19] ensemble method on the MIT300 saliency benchmark.

Deep Gaze I showed that using off-the-shelf image recognition features from ImageNet in a convolutional neural network can significantly outperform shallower and simpler, non-data-driven saliency models without being trained explicitly nor end-to-end specifically for the saliency task.

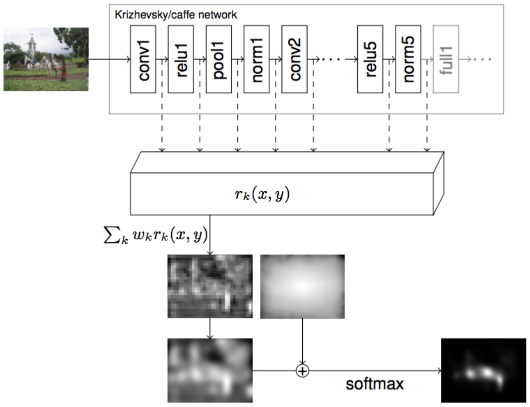


Figure 7. Kmmerer *et al.*'s Deep Gaze I convolutional neural network, pre-trained for object recognition

It was also the first model (shown in Figure 7) to employ a relatively deeper convolutional neural network model for saliency prediction (5 layers compared to *eDN*'s 1-3 layers), although it borrows a significant portion of its weights from AlexNet trained on ImageNet. From here, we move onto more recent methods that approach the saliency prediction task completely end-to-end.

4.2. Deep Supervised Learning

In 2015, *SalDet* [21] attempts to overcome problems with low-level saliency cues in low-contrast backgrounds and confusing overall appearances. They create a deep learning framework focused around consideration of multiple contexts to detect salient objects in images. Their deep model considers both global context to detect saliency in the entire image as well as local context to predict saliency in small, localized regions. Together, the model examines the image at multiple resolutions and both global and local contexts are trained jointly. However, this system only achieves saliency detection, which answers the question of whether or not a detected object is salient, and does not perform saliency prediction, which is to produce a saliency map over the original image.

In the same year, Kruthiventi *et al.* introduce a groundbreaking model named *DeepFix* [10]. *DeepFix* was the first fully-convolutional neural network for saliency prediction of its kind, learning hierarchical features in an end-to-end fashion. It captures semantics at multiple scales while still accounting for global context through the use of a novel Location Based Convolutional (LBC) Layer, which overcomes the otherwise problematic spatially-invariant property of classic fully-convolutional neural networks. *DeepFix* (shown in Figure 8 employs 5 convolution blocks with weights initialized from the VGG-16 network and ends with two of their novel LBC layers before producing a final saliency map. In addition, their convolution blocks leverage Gaussian priors as saliency map priors to further improve their learned weights. This fully-convolutional model significantly outperformed all other models in 2015, and competitively remains near the top of the MIT300 benchmark today.

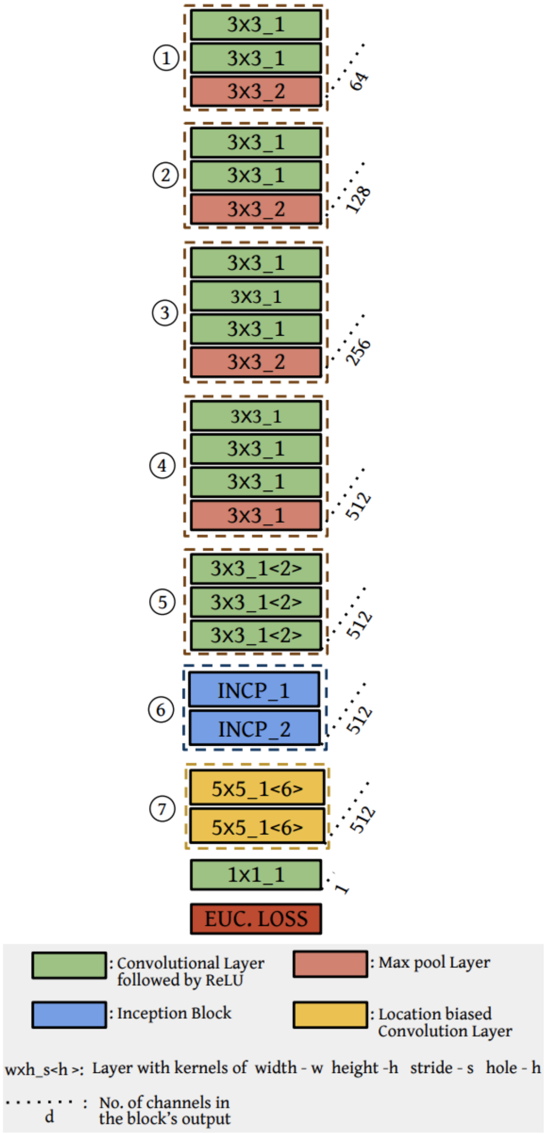

①
3x3_1
3x3_1
3x3_2
64
②
3x3_1
3x3_1
3x3_2
128
③
3x3_1
3X3_1
3x3_1
3x3_2
256
④
3X3_1
3x3_1
3x3_1
3x3_1
512
⑤
3x3_1<2>
3x3_1<2>
3x3_1<2>
512
⑥
INCP_1
INCP_2
512
⑦
5x5_1<6>
5x5_1<6>
512
1x1_1
1
EUC. LOSS
: Convolutional Layer followed by ReLU
: Max pool Layer
: Inception Block
: Location biased Convolution Layer
wxh_s<h >: Layer with kernels of width - w height -h stride - s hole - h
. . . . : No. of channels in the block's output
d

Figure 8. DeepFix: the first fully-convolutional neural network trained end-to-end for saliency prediction

Also in 2015, Huang *et al.* introduce *SALICON* [5], a deep neural network for saliency prediction, as well as the largest publicly available saliency prediction dataset today, also named SALICON. We will visit the dataset in Section 5. Huang *et al.* attribute the large gap in saliency prediction between computers and humans to a so-called "semantic gap", which is the limited capability of computer models to predict eye fixations in scenes with strong semantic content.

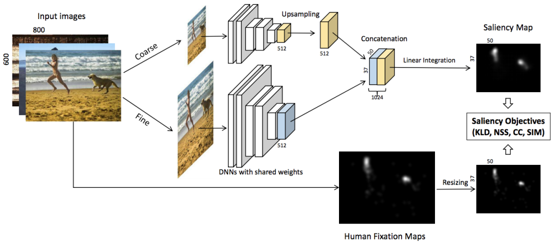


Figure 9. SALICON (Saliency in Context): a deep neural network to reduce the semantic gap via domain adaptation and multi-scale learning

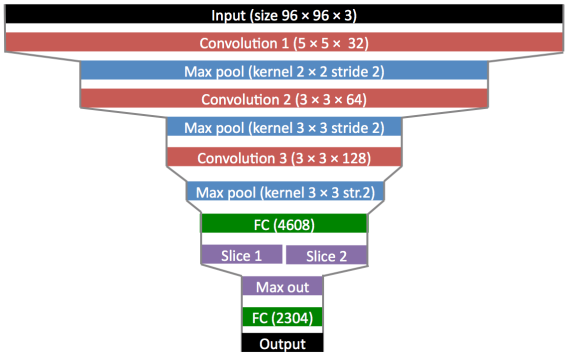

Input (size 96 × 96 × 3)
Convolution 1 (5 × 5 × 32)
Max pool (kernel 2 × 2 stride 2)
Convolution 2 (3 × 3 × 64)
Max pool (kernel 3 × 3 stride 2)
Convolution 3 (3 × 3 × 128)
Max pool (kernel 3 × 3 str.2)
FC (4608)
Slice 1
Slice 2
Max out
FC (2304)
Output

Figure 10. JuntingNet: the first shallow convolutional network trained end-to-end for saliency prediction

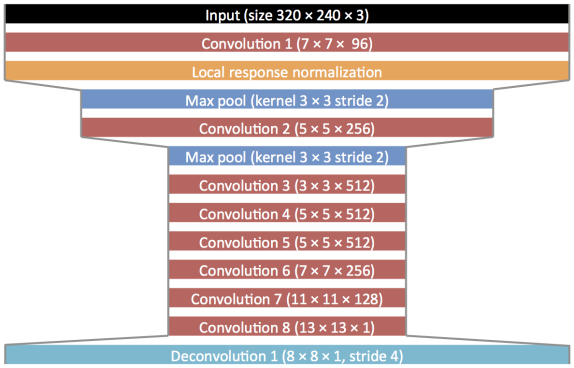

Input (size 320 × 240 × 3)
Convolution 1 (7 × 7 × 96)
Local response normalization
Max pool (kernel 3 × 3 stride 2)
Convolution 2 (5 × 5 × 256)
Max pool (kernel 3 × 3 stride 2)
Convolution 3 (3 × 3 × 512)
Convolution 4 (5 × 5 × 512)
Convolution 5 (5 × 5 × 512)
Convolution 6 (7 × 7 × 256)
Convolution 7 (11 × 11 × 128)
Convolution 8 (13 × 13 × 1)
Deconvolution 1 (8 × 8 × 1, stride 4)

Figure 11. SalNet: the first deep convolutional network trained end-to-end for saliency prediction

To help bridge this semantic gap, they propose a deep neural network in *SALICON*, which leverages information from multiple image scales. Shown in Figure 9, *SALICON* leverages existing deep architectures such as AlexNet, VGG-16, and GoogLeNet, shares weights between them, and is applied at two different image scales, fine and coarse, to obtain a saliency map.

Next, in 2016, Pan et al. [16] present two different end-to-end convolutional neural networks, nicknamed *JuntingNet* and *SalNet* (shown in Figures 10 and 11, respectively). JuntingNet is a shallow network inspired by AlexNet, but with only 3 convolution layers and 2 fully-connected layers. On the other hand, SalNet is a deeper network with 8 convolution layers. Both models learn saliency prediction completely end-to-end by using the convolutional neural networks for regression rather than classification (which would more closely achieve saliency detection). Their shallow network is trained entirely from scratch using the recently available saliency datasets, while their deep network uses transfer learning from a network trained for classification. They make a key observation that the parameters from the lowest levels in networks trained for classification can be transferred effectively for saliency prediction. Using the pre-trained layers, they add on some new layers and train those specifically with saliency prediction in mind.

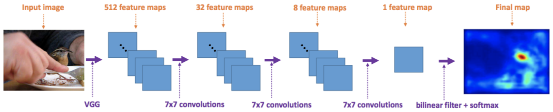


Figure 12. PDP (Probability Distribution Prediction): a probabilistic model that approximates a saliency map as a generalized Bernoulli distribution

Also in 2016, Jetley *et al.* take a very different and probabilistic approach towards the saliency prediction task. Their model, named *PDP* [7] (Probability Distribution Prediction) formulates the saliency map as a generalized Bernoulli distribution and trains a model to learn this distribution. They train a deep neural network completely end-to-end using novel loss functions that pair the classic softmax loss function with functions that compute the distances between different probability distributions. With *PDP* (shown in Figure 12), Jetley *et al.* showed that using their novel loss functions for training deep networks outperforms other traditional loss functions such as Euclidian and Huber loss.

At around this time, recurrent neural networks (RNN) arose as a variant of the usually convolutional neural network that performed especially well on some computer vision tasks that required more internal state. With the introduction of RNNs to the computer vision community, Liu and Han took the concept of the RNN and applied it to the saliency prediction task, producing *DSCLRCN* [13], a deep spatial contextual long-term recurrent convolutional neural network. Shown in Figure 13, *DSCLRCN* first learns local saliency on multiple small regions throughout the image entirely in parallel. Then, unlike many other deep neural networks for saliency prediction, *DSCLRCN* mimics the human visual system's ability to incorporate global context by leveraging a deep spatial long short-term model (LSTM). The model first uses a state-of-the-art CNN model for scene classification to extract scene features, which is then used as contextual information for an internal LSTM, producing a novel deep spatial contetual LSTM. Putting all of these components together, each of which can be trained end-to-

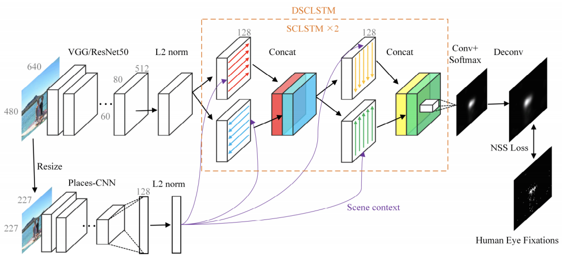


Figure 13. DSCLRCN: A deep spatial contextual long-term recurrent convolutional neural network (recurrent neural network) for saliency prediction

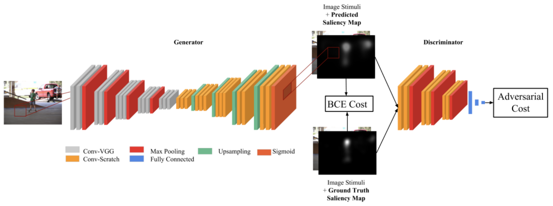


Figure 14. SalGAN: a generative-adversarial network for predicting saliency maps

end, produces the overall *DSCLRCN* model, which outputs a final saliency map. The introduction of the LSTM helps the model incorporate global contextual information, and drastically enhances the accuracy of its predicted saliency maps, allowing it to outperform the state-of-the-art. Today, *DSCLRCN* still remains at the top of the MIT300 benchmark.

In 2017, Pan *et al.* [16] improve upon their work in *JuntingNet* and *SalNet* with the introduction of the *SalGAN* model [15] (Figure 14), a deep convolutional neural network trained with adversarial examples. They create a generator model learned from back-propagation via binary cross entropy loss on existing saliency maps, which is then passed into the discriminator network that is trained to identify whether or not a provided saliency map was generated by the generator, or a ground-truth one captured from a human. *SalGAN* achieved competitive results at the top of the MIT300 benchmark, and still hovers near the top today.

More recently, Kmmerer *et al.* [11] build upon *Deep Gaze* to further explore the unique contributions between low-level and high-level features towards fixation prediction. While high-level features learned by deep networks have proven to be very useful in predicting natural human fixation locations in images, the performance of current saliency prediction systems still fall short of human performance. To this end, they train *Deep Gaze II* to learn high-level features and a separate model to learn low-level features and analyze the two. They show that low-level features are excessively neglected in favor of learning higher-level features with increasingly deeper and deeper networks. This provides some motivation for further experimentation of different feature types within this project.

5. Datasets, Benchmarks, and Evaluation

None of these supervised models could exist without the help of sufficiently large and well-labeled datasets. In this section, we discuss some of the most influential datasets and benchmarks within saliency research that have helped advance the research area in the past several years.

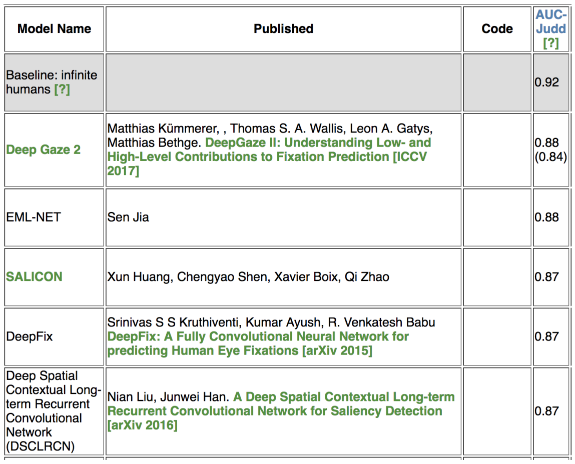
5.1. MIT300 and MIT1003

Released in 2012, the MIT300 benchmark and test dataset [1] is the most well-known and standard benchmark for evaluating saliency predictors. It is captured via eye-tracking devices on human subjects and contains a test set of 300 images of natural indoor and outdoor scenes for evaluation. Because the fixation points in the MIT300 dataset are not publicly available, this dataset is used exclusively for benchmarking. Recent research in saliency prediction all benchmark their models on the MIT300 benchmark as the primary way to compare their models with other models.

The MIT1003 dataset contains 1003 natural indoor and outdoor scenes commonly used as training data before evaluation via the MIT300 bechmark above. Like the MIT300 benchmark, all 1003 images are captured using an eye-tracker on humans.

It is worth noting that these datasets are relatively small. As a result, there is a real danger for potential overfitting when training and testing these models. Despite this, the MIT300 benchmark is currently the gold standard for evaluating and comparing saliency models. It supports multiple different metrics for comparison; at this time, models are primarily ranked using AUC-Judd (a varient of the standard area-under-curve metric), but will change to using NSS (normalized scanpath saliency) very soon in light of recent research towards finding and using better metrics for saliency evaluation [2]. NSS is the normalized scanpath saliency between two different saliency maps. It is measured as the mean value of the normalized saliency map at fixation locations, and should provide a more meaningful evaluation and ranking of saliency models compared to the current AUC-Judd metric.

Figure 15 shows the top of the standings for the MIT300 benchmark. Currently, *Deep Gaze II*, *SALICON*, *DeepFix*, and *DSLCRCN* are very closely tied with each other at the top of the MIT300 benchmark when ranked using AUC-Judd. However, the differences become much clearer once the rankings shift to using NSS, as *DSLCRCN* becomes a much clearer victor.



Model Name	Published	Code	AUC-Judd [?]
Baseline: infinite humans [?]			0.92
Deep Gaze 2	Matthias Kümmerer, , Thomas S. A. Wallis, Leon A. Gatys, Matthias Bethge. DeepGaze II: Understanding Low- and High-Level Contributions to Fixation Prediction [ICCV 2017]		0.88 (0.84)
EML-NET	Sen Jia		0.88
SALICON	Xun Huang, Chengyao Shen, Xavier Boix, Qi Zhao		0.87
DeepFix	Srinivas S S Kruthiventi, Kumar Ayush, R. Venkatesh Babu DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations [arXiv 2015]		0.87
Deep Spatial Contextual Long-term Recurrent Convolutional Network (DSCLRCN)	Nian Liu, Junwei Han. A Deep Spatial Contextual Long-term Recurrent Convolutional Network for Saliency Detection [arXiv 2016]		0.87

Figure 15. The MIT300 saliency benchmark, ranked by AUC-JUDD metric

5.2. CAT2000

Released in 2015, the CAT2000 dataset is a relatively larger dataset consisting of 2000 training images and 2000 test images that span 20 different categories. The training set contains 100 images from each of the categories and have fixation annotations from 18 different observers each. Similar to the MIT300 and MIT1003 datasets, the CAT2000 dataset is obtained using the EyeLink1000 eye-tracking device.

5.3. SALICON

Introduced in 2015, the SALICON dataset [8] is the largest dataset available for saliency prediction, initially released with 10000 training images, 5000 validation images, and 5000 test images. Most saliency prediction research use SALICON to train their models. SALICON is a subset of the Microsoft CoCo dataset and uses MS COCO's pixel-wise semantic annotations to create a large set of saliency annotations on a subset of the original MS COCO images. In order to achieve such a large dataset for saliency prediction, SALICON foregoes the current standard of using eye-tracking devices to gather ground-truth data and instead uses a more scalable, crowd-sourced method involving a mouse. Although this data collection method may affect the accuracy and quality of the dataset, SALICON nevertheless introduces an acceptable and scalable method for the collection of further data for saliency research.

5.4. Further Crowdsourced Methods

Following in the footsteps of SALICON, other researchers sought to solve the problem of saliency datasets being orders-of-magnitude smaller than typical datasets for other visual recognition tasks. One such example proposes *TurkerGaze* [17], a method for crowdsourcing saliency datasets with webcam-based eye tracking using Amazon Mechanical Turk (AMT) workers. Their method uses a carefully designed web-based game as an eye-tracking experiment to effectively perform facial landmark tracking and then learns to predict the AMT worker's gaze. From then, the model uses its gaze prediction model to collect gaze data from presenting novel images to the AMT worker. Using this method, it is possible to use AMT workers to quickly and efficiently collect novel saliency data for further saliency research.

6. Future Directions and Conclusion

Saliency research has advanced significantly in recent years, in large part due to the use of deep supervised learning, made possible by recently available datasets like SALICON. Now, there are multiple variations of deep learning models, including but not limited to shallow as well as deep convolutional networks, multi-resolution networks, recurrent neural networks, fully convolutional networks, generative-adversarial networks, and even networks for predicting probability distributions. However, the saliency task is far from solved.

Many of the current methods compete closely with one another at the top of the MIT300 saliency benchmark, but progress has somewhat slowed down, partially due to a lack of good saliency evaluation methods and partially due to a lack of sufficiently large datasets like those typically available to other visual recognition tasks. These constraints limit how we train models and urge continued research on both fronts in order to support the development of even better models.

Even so, current state-of-the-art methods still fall short from human levels of performance. Although effective for general scenes, we still find that saliency predictors still cannot fully understand the high-level semantics in semantically rich scenes. In other words, the ”semantic gap” has closed slightly with recent models, but still remains an outstanding research problem today. For example, even the best saliency predictors today tend to place a disproportionate amount of importance on text and humans, even when they are not necessarily the most semantically interesting parts of the image. Further, when presented with multiple pieces of text, some text may be more semantically salient than others, but today's models are incapable of detecting the difference and treats them equally.

As it stands, there is still room for improvement in both the models themselves as well as how we approach collecting data and evaluating our results. The enormous potential of accurate and efficient saliency predictors is an exciting prospect for the advancement of technology and should prove to be a promising visual recognition research area for years to come.

References

[1] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark.

[2] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *CoRR*, abs/1604.03605, 2016.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[5] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 262–270, Dec 2015.

[6] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2, 2001.

[7] S. Jetley, N. Murray, and E. Vig. End-to-end saliency mapping via probability distribution prediction. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5753–5761, June 2016.

[8] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[10] S. S. Kruthiventi, K. Ayush, and R. V. Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *CoRR*, abs/1510.02927, 2015.

[11] M. Kummerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge. Understanding low- and high-level contributions to fixation prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] M. Kmmerer, L. Theis, and M. Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. In *ICLR Workshop*, May 2015.

[13] N. Liu and J. Han. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *CoRR*, abs/1610.01708, 2016.

[14] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga. Low-level spatiochromatic grouping for saliency estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2810–2816, Nov 2013.

[15] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. a. Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. In *arXiv*, January 2017.

[16] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] Y. Z. A. F. S. R. K. J. X. Pingmei Xu, Krista A Ehinger. Rich feature hierarchies for accurate object detection and semantic segmentation. 2015.

[18] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? PP, 12 2016.

[19] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2798–2805, June 2014.

[20] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1162, June 2013.

[21] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.