

Chapter 7

Modeling Attention in Engineering



Matei Mancias 

7.1 Attention in Computer Science: the Notion of Saliency Map

Studies in neuroscience [7] suggest that human visual attention is enhanced through a process of competing interactions among neurons representing all of the stimuli (or features extracted from the image in computer science) present in the visual field (which depends on the eye fixation). The competition results in the selection of a few attended regions and the suppression of irrelevant material. It means that people and animals are able to spot outstanding patterns in a scene and are drawn in general to anomalous objects (bottom-up attention) and by what is already known (learning) to be interesting for the viewer while performing a visual task (top-down attention). This makes visual attention a vital element in the survival of all creatures that have evolved since the Cambrian explosion [39] when vision first appeared on the Earth.

The ensemble of mechanisms grouped under the term attention swings into action before we are even conscious of anything strange. Indeed there is a pre-attentive period often less than 100 ms during which low-level processes rapidly identify image regions that deserve attention. Pre-attentive processing has been the subject of considerable research. Treisman [50] described experiments that reveal pre-attentive behavior in human vision. She pointed out a “masking effect” that depends upon the presence elsewhere of other elements sharing the local distinctive property. A locally salient feature can be suppressed by either other local or more distant structures in the image.

In computer science, the study of visual attention is based on the notion of “saliency.” A saliency map is a representation of the probability for a human mean

M. Mancias (✉)
Numediart Institute, University of Mons, Mons, Belgium
e-mail: matei.mancias@umons.ac.be

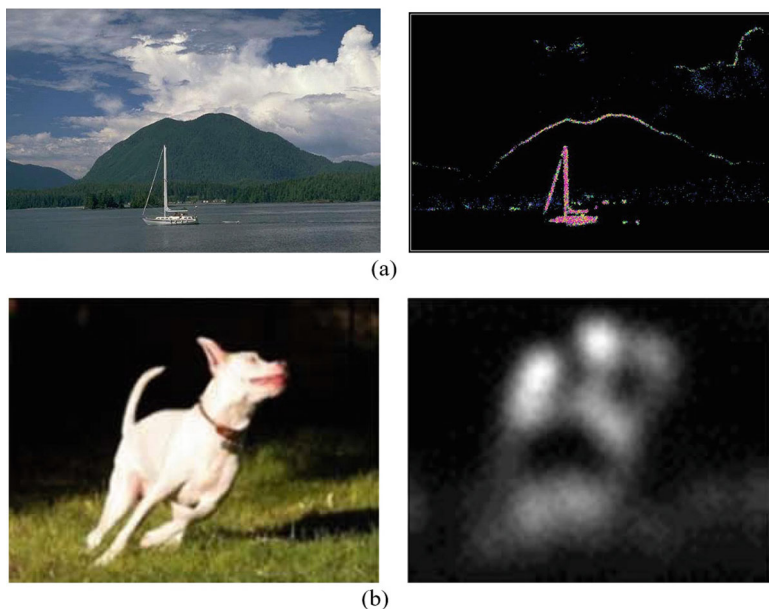


Fig. 7.1 Image and corresponding saliency maps (a) Stentiford [46], (b) Hou and Zhang [11]. The probability of each pixel to be attended given a visual task can be represented by heatmaps. Depending on the visual attention model the heatmap can be fine-grained (a) or very fuzzy (b)



Fig. 7.2 Example of SOD “saliency map” which is a segmentation mask of the most interesting object (at 1) from the background (at 0)

observer to attend each pixel of an image depending on some visual task. Those maps can be represented by heatmaps (Fig. 7.1) which indicate in clear pixels higher attention. Some maps can be more fine-grained, and by taking a look on the saliency map we could recognize the initial image [46] or very fuzzy where it is not possible to recognize the initial image from the shape of the saliency map [11].

In addition to classical saliency maps, an adaptation for detecting salient objects is also sometimes used: salient object detection based (SOD). Those “saliency maps” do not represent the probability for a viewer to attend some pixels, but the most interesting object (Fig. 7.2).

Sometimes there are several objects in the same image, but most of the time there is no hierarchy of the saliency among different objects (being all at 1). The SOD maps are binary maps which are not very rich in terms of saliency information. Ideally those SOD masks should be obtained by an image segmentation model where each object has assigned its average or maximum saliency value (instead of just 1) so that the different objects in the image exhibit a saliency hierarchy in the scene.

In this chapter we will focus only on the saliency maps and not its derived SOD map.

7.1.1 General Framework for Saliency

Based on Torralba et al. [49] and Geisler and Cormack [10] we can express a saliency map S as a function of space (pixel position) x and the position of the eye fixation r_i . Indeed, depending on the eye fixation, due to spatial resolution being greatest at the foveated location (Fig. 7.3), the image resolution can be drastically changed so that the extracted features F in a given position might vary depending on (1) their location in the image and (2) the eye fixation location: $F(x, r_i)$.

Based on that observation, a general formalization of a saliency map is

$$S(x, r_i) = p(X|F(x, r_i), G) \quad (7.1)$$

where $S(x, r_i)$ is the probability to attend to a target object X which depends on the visual task, given both the features which can be extracted from the image at the eye fixation i : $F(x, r_i)$ and the global features of the image (context) G .

In [10] the visual task can be a strong task (such as searching for keys) or a weak task (such as free viewing).

Fig. 7.3 Example of foveated image: The eye fixation is on the cross, and this positioning of the eye has an important influence on the features which can be extracted from the image



Based on a Bayesian framework from Torralba et al. [49] and the dynamic view in [10] we can derive a general framework for saliency in computer science:

$$S(x, r_i) = p(X|F(x, r_i), G) = \frac{p(X = x|G)}{p(F(x, r_i)|G)} p(F(x, r_i)|X = x, G) \quad (7.2)$$

Equation 7.2 exhibits two terms. The first one is

$$S(x, r_i) = \frac{p(X = x|G)}{p(F(x, r_i)|G)} \quad (7.3)$$

This first term 7.3 basically asserts that the more unlikely the encoded features at a location for a given eye fixation, given the type of scene G , then the greater the saliency. This is the definition of the bottom-up attention where “strange” features should attract attention.

$$S(x, r_i) = p(F(x, r_i)|X = x, G) \quad (7.4)$$

The second term 7.4 from Eq. 7.2 describes the probability of the features given the task-related target object X and the image context G . This is the top-down attention where the task and context can orient the attention process.

7.1.2 Static Bottom-up Saliency

In 7.3 we can do a simplification and indeed say that we do not take into account the eye fixation dynamics and use the full image without any foveated preprocessing. In this case saliency does not depend on r_i , which means that there is only one map for the whole image as in Fig. 7.1. This leads to the following equation which describes the static bottom-up saliency:

$$S(x) = \frac{p(X = x|G)}{p(F(x)|G)} \quad (7.5)$$

There are a lot of models in this area, and in Sect. 7.2 we will focus on static bottom-up models and their classification.

7.1.3 Static Top-Down Saliency

In 7.4 we can do the same simplification and arrive to the following equation which describes the static bottom-up saliency:

$$S(x) = p(F(x)|X = x, G) \quad (7.6)$$

Incorporating top-down information allows the statistics of related images to influence the parameters that determine saliency values. In the extreme, top-down attention becomes object recognition when attention is solely directed at a particular class of object and the features characterizing those objects are used as a template in the calculation of the saliency measure.

While introducing top-down information was a challenging task, the deep-learning-based models mainly focus on top-down information. This top-down information will be addressed in Sect. 7.3.

7.1.4 *Dynamic Overt Saliency*

In Sect. 7.4 we will focus on overt attention models and saliency models which try to introduce dynamics in saliency computation based on the eye dynamics. In this case the eye fixation location becomes a fundamental information and r_i comes back into the equations. Those approaches will have as an output a saliency map which can change in time from the first fixation to the last one.

7.2 Saliency: Static Bottom-up Approach

Static bottom-up approaches (see Eq. 7.5) uses features extracted only once from the signal independently from the eye fixations. Extracted features can be various such as low-level features (luminance, color, orientation, texture, ...), position features (objects relative position, neighborhoods, or patches from the signal) but also “deep features” (extracted by a convolutional network encoder). Once those features are extracted, all the existing methods are essentially based on the same principle: looking for high contrast, rare, surprising, novel, worthy to learn, less compressible, or information maximizing areas. All those definitions are actually synonyms as they all amount to searching for some unusual features in a given context which can be spatial and/or temporal. In the following, we provide examples of contexts used for different kind of signals for visual attention on images.

The literature is very active concerning still image saliency models. While some years ago only a few labs in the world were working on this topic, nowadays hundreds of different models are available. Those models have various implementations and technical approaches even if initially they all derive from the same idea.

It is thus very hard to find a simple taxonomy which classifies all the methods. Some attempts of taxonomies proposed an opposition between “biologically-driven” and “mathematically-based” methods with a third class including “top-down information.” This approach implies that only some methods can handle top-down information, while all bottom-up methods could use top-down information more or less naturally. Another difficult point is to judge the biological plausibility which can be obvious for some methods but much less for the others. Another criterion

is the computational time or the algorithm complexity, but it is very difficult to make this comparison as all the existing models do not provide cues about their complexity. Finally, a classification of methods based on center-surround contrast compared to information theory-based methods does not take into account different approaches as the spectral residual one for example. Other taxonomies will also be introduced in the next chapters as for example the dependence on image features or the “classical” approach versus “deep learning” approach. Here, we show a taxonomy of the bottom-up saliency methods which is based on the context that those methods take into account to exhibit signal novelty. In this framework, there are three classes of methods.

The first one is pixel’s surroundings: Here a pixel, a group of pixels, or a patch is compared with its surroundings at one or several scales.

A second class of methods uses as a context the entire image and compares pixels or patches of pixels with other pixels or patches from other locations in the image but not necessarily in the surroundings of the initial patch. Some models even use more than one image as a context: An entire dataset can be used here.

Finally, the third class takes into account a context which is based on a model of what the normality should be.

In the following sections, these three classes of models are illustrated.

7.2.1 Context: Pixel’s Surroundings

This approach is initially based on a biological motivation. Its origins come from the work of Koch and Ullman [16] on attention modeling. The main idea is to compute visual features at several scales in parallel, to apply center-surround inhibition, combination into conspicuity maps (one per feature) and finally to fuse them into a single saliency map. There are a lot of models derived from this approach which mainly use center-surround contrast as a local measure of novelty. A good example of this family of approaches is Itti’s model (Fig. 7.4) [13] which is the first implementation of the Koch and Ullman model. It is composed of three main steps. First, three types of static visual features are selected (color, intensity, and orientation) at several scales. The second step is the center-surround inhibition which provides high responses in case of high contrast, while it has low response in case of low contrast. This step results in a set of feature maps for each scale. The third step consists of an across-scale combination, followed by normalization to form “conspicuity” maps which are single multiscale contrast maps for each feature. Finally, a linear combination is made to achieve inter-feature fusion. Itti proposed several combination strategies: A simple and efficient one is to provide higher weights to conspicuity maps which have global peaks much bigger than their mean. This is an interesting step which integrates global information in addition to the local multiscale contrast information.

This implementation proved to be the first successful approach of attention computation by providing better predictions of the human gaze than chance or simple

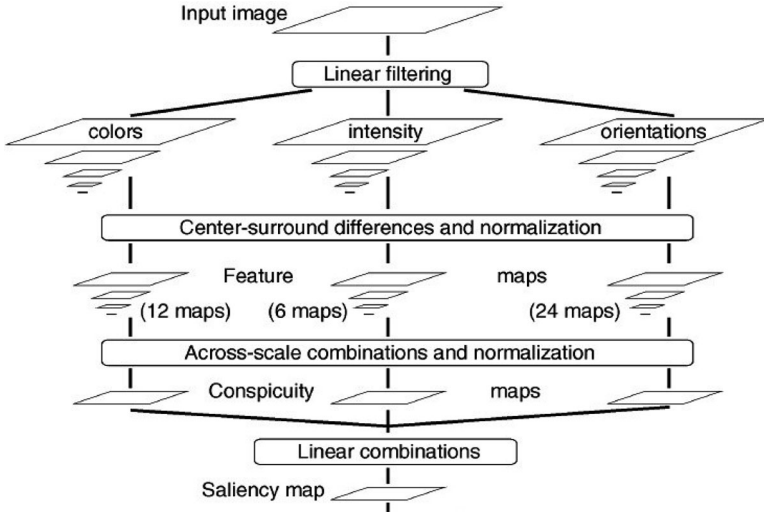


Fig. 7.4 Model of Itti et al. [13]. Three stages: center-surround differences, conspicuity maps, inter-feature fusion into saliency map. (Adapted from Itti et al. [13])

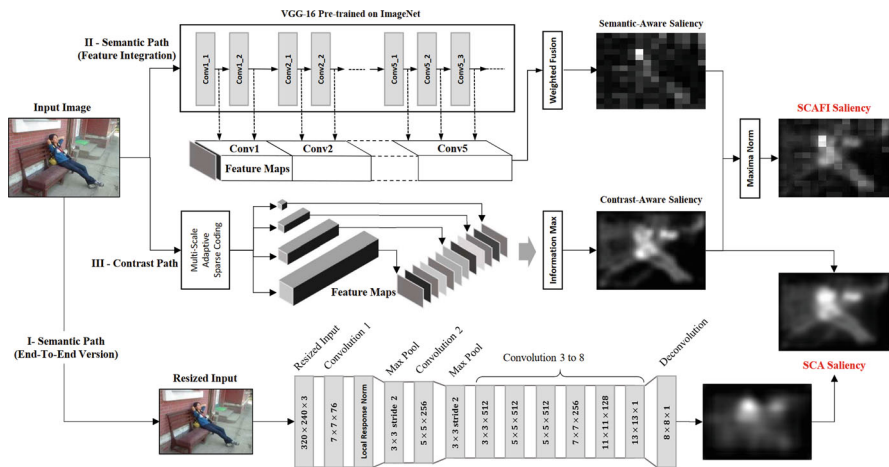


Fig. 7.5 SCAFI model. Path I is mainly top-down, path II mainly based on a full image context, and path III is based on center-surround contrast. (Adapted from Sun [47])

descriptors like entropy. Following this success, most of the computational models of bottom-up attention use the comparison of a central patch to its surroundings as a novelty indicator.

Following advances in deep learning, some models tried to use the idea of center-surround but based on deep features. As an example, in the SCAFI model (Fig. 7.5), Sun [47] proposed three different paths to be mixed at the end in a unique saliency

map. Path I is mainly a top-down end-to-end convolution/deconvolution approach as we will show in Sect. 7.3. Path II is more about global context and could fit into Sect. 7.2.2. Path III is specifically a multiscale approach which will provide contrast information based mainly on local context due to the different scales interactions. Another approach called DeepFeat [26] is based on a center-surround difference of the convolutional layers of a fine versus coarse encoder which is then concatenated into a bottom-up saliency. This bottom-up saliency is then mixed with an object class activation top-down map and a centred gaussian to provide the final saliency map.

7.2.2 Context: The Whole Image or a Dataset of Images

In this approach, the context which is used to provide a degree of novelty or rarity to image patches is not necessarily the surroundings of the patch but can be other patches in its neighborhood or even anywhere in the image or an image database. The idea can be divided in two steps. First, local features are computed in parallel from a given image. The second step measures the likeness of a pixel or a neighborhood of pixels to other pixels or neighborhoods within the image. This kind of visual saliency is called “self-resemblance.” A good example is shown in Fig. 7.6. The model has two steps. First it proposes to use local regression kernels as features. Second it proposes to use a nonparametric kernel density estimation for such features, which results in a saliency map consisting of local “self-resemblance” measure, indicating likelihood of saliency [44].

Mancas [28] and Riche et al. [41] focus on the entire image. These models are designed to detect saliency in the areas which are globally rare and locally high contrast. After a feature extraction step, both local contrast and global rarity of pixels are taken into account to compute a saliency map. An example of the difference between locally contrasted features and globally rare is given in Fig. 7.7. On the left there is the initial image of an apple with a defect in red, the second image shows the fixations predicted by Itti et al. [13] where the locally contrasted apple edges are well detected, while its less contrasted but rare defect is not. The third image shows [30] which detected the apple edges, but also the defect. Finally the rightmost is the

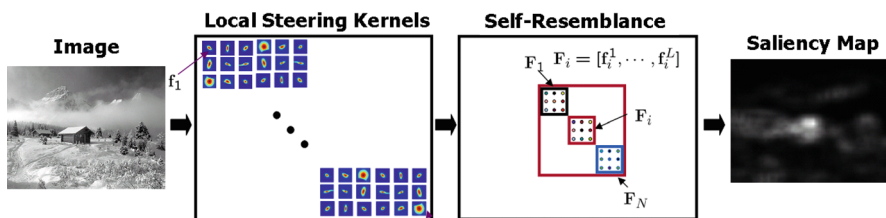


Fig. 7.6 Model of Seo and Milanfar [44]. Patches at different locations in the image are compared. (Adapted from Seo and Milanfar [44])

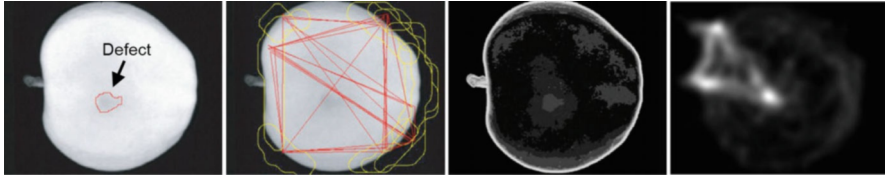


Fig. 7.7 Difference between locally contrasted and globally rare features. Left image: an apple with a defect in red, Second Image: Itti et al. [13], Third image: Mancas et al. [30], Right image: mouse tracking (ground truth)

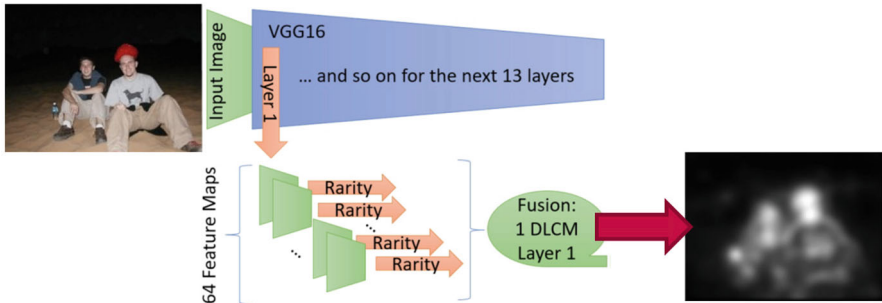


Fig. 7.8 Deep rare idea: applying global rarity on all convolutional feature maps of an encoder (which can be trained or not for saliency). Then all those maps are fused in a final bottom-up global saliency map. (Image adapted from Kong et al. [17])

mouse-tracking result for more than 30 users. Boiman and Irani [4] look for similar patches and relative positions of these patches in an image database which provide more cues about what should be normal. The use of a database might be viewed as an introduction of top-down information.

In this domain, deep features can also be used globally. As we saw in the previous section, a path of the SCAFI Sun [47] model already uses all the features by making a weighted fusion of the convolutional layers of a VGG-16 encoder trained on Image-Net general dataset Deng et al. [6]. The weighted sum of the different layers provides a bottom-up map based on the global context of the image.

In addition, other models like Mancas et al. [31] and its extended version Kong et al. [17] can be included in the global context category. Indeed, the Rare 2012 model Riche et al. [41] was applied to deep features instead of hand-made features (Fig. 7.8). Each convolutional feature map was extracted from each layer, and rarity was applied to those maps which were fused together. The conspicuity maps of the different layers are then fused together in a final saliency map.

Rarity can also be thresholded before the fusion to avoid the noise, and it can be compared on different scales depending on how the encoder layers are reconstructed (see Fig. 7.9). The threshold T is null on the right which can be seen on the different more noisy conspicuity maps at different levels. It is increasing when going to the left leading to less noisy conspicuity maps. Depending on the image, higher or lower

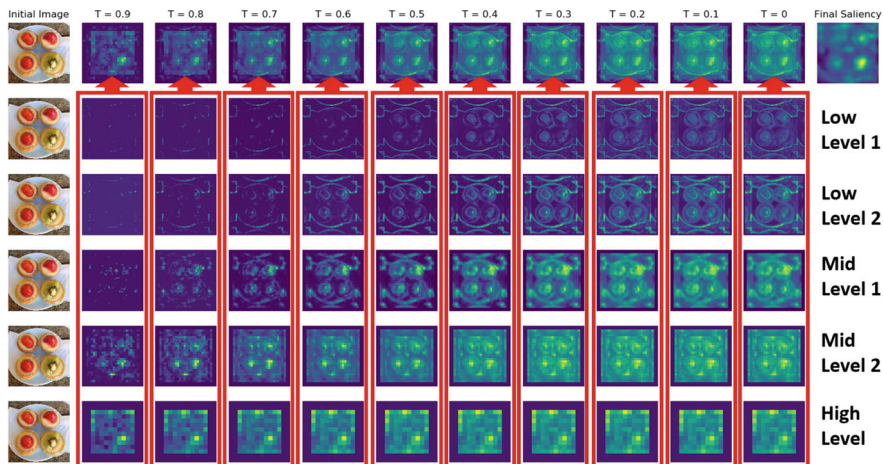


Fig. 7.9 From down to top higher to lower level features conspicuity maps and the final saliency map on top. From left to right : lower and lower thresholds on rarity before the fusion leading to more and more noise in the maps. (Image adapted from Kong et al. [17])

level rarity maps are more or less effective in capturing the objects rarity. On this image, as the 4th more different cake is quite big, its rarity will be captured on high-level features only (last 2 lines and especially the last line).

7.2.2.1 Context: A Model of Normality

This approach is probably less biologically motivated than most of the other implementations. The context which is used here is a model of what the image should be: If things are not like they should be, this can be surprising and thus attracts the observer’s attention. Achanta et al. [1] proposed a very simple attention model (Fig. 7.10): First, the color space is converted from RGB into Lab, and second the Euclidean distance is computed between a Gaussian filtered version of the input image and the average Lab vector of the input image. The mean image used is a kind of model of the image statistics: Pixels which are far from those statistics are more salient. This model is mainly useful in salient objects detection.

Another approach to “normality” can be found in [12], where the authors proposed a spectral model that is independent of any features. As it is known that natural images have a $\frac{1}{f}$ decreasing Fourier log-spectrum, the difference between the log-spectrum of the image and its smoothed log-spectrum (spectral residual) is reconstructed into a saliency map. Indeed, a smoothed version of the log-spectrum is closer to a $\frac{1}{f}$ decreasing log-spectrum as small variations are removed. This approach is almost as simple as Achanta et al. [1] but much more efficient in predicting eye fixations.

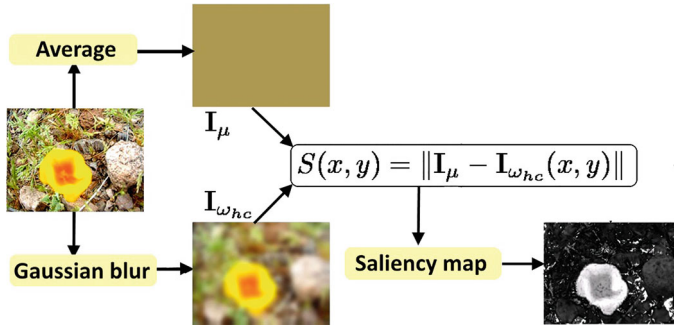


Fig. 7.10 Achanta et al. [1] use a model of the mean image. (Adapted from Achanta et al. [1])

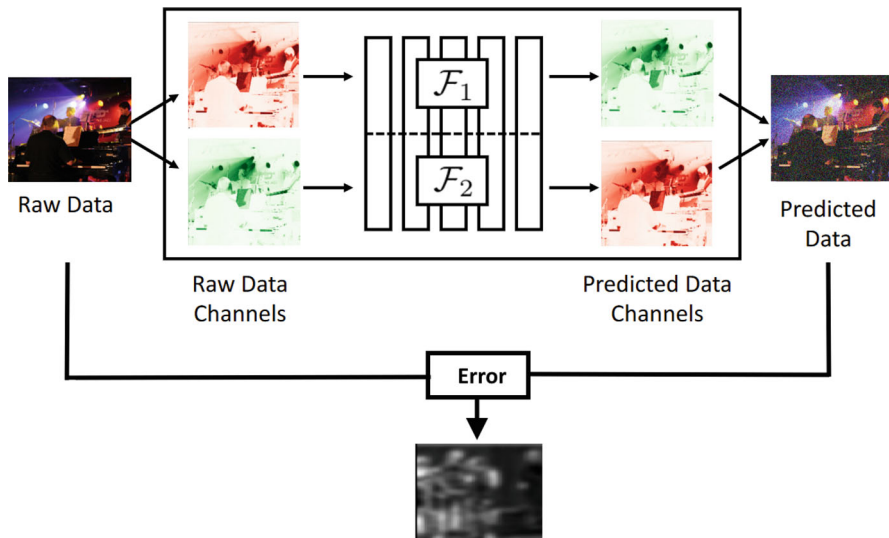


Fig. 7.11 Millidge and Shillcock [33] use an autoencoder architecture to predict what the normality of the image should be and the error between the image and this prediction is considered to be a saliency map. (Adapted from Zhang et al. [54])

With the arrival of deep learning, the normality modeling can also be done using specific deep architectures such as autoencoders. In [33], the authors build on split-brain autoencoders Zhang et al. [54] which aim to predict the initial image from a latent space of reduced dimensionality based on data splits. Figure 7.11 shows how the model works. An image is split into different channels used to predict the other channels and fused together. The error between the initial image and the predicted image is thus the final saliency map.

7.3 Saliency Models: Including Top-Down Information

While bottom-up information basically arises from the exogenously acquired signal as a pop-out region/object, top-down represents endogenous information and comes from the inner world (information from memory, their related emotional level, and also the task-related information). The separation between bottom-up and top-down information is far from being clear. Depending on the viewpoint and the definitions, some notions can be considered as either bottom-up or top-down.

One can say that top-down is not involved if memory and learning are not involved. In this case all the hard-wired features which might be low level (luminance, color, orientation, motion direction), mid-level (object basic properties as the size, centered-Gaussian as a default context), or high level (face detection, people detection), which involve specific brain areas but do not need memory and learning are bottom-up.

Top-down involves learning and memory and will deal with specific contexts (e.g., websites, ads, ...), objects (face recognition, people recognition, specific animal, or object), or a given task coming from inner needs (looking for the keys, ...).

It is thus interesting that face detection might be considered as bottom-up (face feature detection does not necessarily need memory and might be located in a specific brain area, the fusiform gyrus McCarthy et al. [32]), while face recognition is clearly top-down as it directly uses memory to remember a specific person.

In practice, two main families of top-down information can be added to bottom-up attention models.

The first one mainly deals with learned normality in a given context which can come from the experience of the current signal if it is time varying, or from previous experience (tests, databases) for still images.

The second approach is about task modeling which can either use object recognition-related techniques or which can model the usual location of those objects of interest or which can model observer behavior given a specific task.

Those two approaches perfectly fit to 7.6 from Eq. 7.2 after static simplification which describes the probability of the features given the task-related target object X and the image context G . This equation represents top-down attention where the task and context can orient the attention process.

7.3.1 Top-Down as Context: Learned Normality

Concerning still images, the “normal” gaze behavior can be learned from the “mean observer” in a given context. Eye-tracking techniques can be used on several users, and the average of their gaze on a set of natural images can be computed. This was achieved by several authors as it can be seen in Fig. 7.12. Judd et al. [15] used eye trackers, while Mancas [27] used mouse-tracking techniques to compute this mean

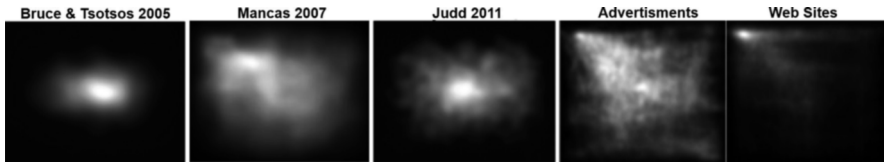


Fig. 7.12 Three models of the mean observer for natural images on the left. The two right images: Model of the mean observer on a set of advertising and websites images

observer. In all cases, it seems clear that, for natural images, the eye gaze is attracted by the center of the images.

This centered distribution seems logical as natural images are taken using cameras, and the photographer will naturally tend to locate the objects of interest in the center of the picture. Another point is that the objects in the center of the visual field are the ones one might interact with, they are then more important than the others.

This observation for natural images is very different from more specific images which use a priori knowledge and which are top-down. In [8], the author shows that the centered distribution mainly follows an horizontal axis for landscapes, while it follows both horizontal and vertical directions for images of interiors. Mancas [28] showed using mouse tracking that gaze density is very different on a set of advertisements and on a set of websites as displayed in Fig. 7.12 on the two right images. This is partly due to a priori knowledge that people have about those images. For example, when viewing a website, the upper part has high chance to contain the logo and title, while the left part should contain the menu. During images or video viewing, the default template is the one of natural images with a high weight on the center of the image. If supplemental knowledge is known about the image, the top-down information will modify the mean behavior toward the optimized gaze density. Those top-down maps can highly influence the bottom-up saliency map, but this influence is variable. In [28] it appears that top-down information seems more important in the case of websites, than advertisements and natural images. Other kinds of models can be learned from videos, especially if the camera is still. It is possible to accumulate motion patterns for each extracted feature, which provides a model of normality. As an example, after a given period of observation, one can say: Here moving objects are generally fast (first feature: speed) and going from left to right (second feature: direction). If an object, at the same location, is slow and/or going from right to left, this is surprising given what was previously learned from the scene, and thus attention will be directed to this object. This kind of consideration can be found in [29]. It is possible to go further and to have different cyclic models in time. In a metro station, for example, normal behavior when a train arrives in the station is different from the one during the waiting period in terms of people's direction, speed, and density ... In the literature (mainly in video surveillance) the variations in time of the normality models are learned through HMMs (Hidden Markov Models) [14].

In order to find abnormality, a more general approach is to learn the features of “normal” situations in images or videos and then compare to the current feature vectors. This approach works well when the backgrounds are stable such as for still cameras of video surveillance always focusing on the same area. Indeed, in that case it is possible to learn the background and “normal” actions happening in the area. A fast Recurrent Neural Network (RNN) with an attention-based Long-Short Term Memory (LSTM) module has been proposed in [45]. Other models rely on Convolutional Neural Networks (CNNs) Sabokrou et al. [43] for anomalous action detection. In general, it has been observed that 3D-CNNs (convolutional networks using 3D convolutions) learn better spatiotemporal features in image sequences compared to standard CNNs as the 3D convolutions capture also the movement features. A 3D autoencoder has also been examined for the task of anomaly detection Sabokrou et al. [42]. Finally we can cite two papers which both called their model “AnomalyNet” in Zhou et al. [55] and Mumtaz et al. [35] with different architectures but both using spatiotemporal data to code normality and check the discrepancy between the current frame features and the learnt normal features.

For 3D signals, another source of information is the proximity of objects. For natural images, centered objects also attract our attention because they might be the ones we will interact with as they are in the center of the visual field. In the same way, a close object is more likely to attract attention as it is more likely to be the first that we will have to interact with. In the real world the default context is a mix between a centered Gaussian and proximity value: Centered close objects are the most important, while far objects on the sides have lower priority.

7.3.2 Top-Down as a Task: Attending to Objects or Actions

While the previous section dealt with attention attracted by events which lead to situations which are not consistent with the knowledge acquired about the scene, here we focus on a second main top-down cue which is a visual task such as “Find the keys!” or “Find the objects which usually attract attention!”. These tasks will also have a huge influence on the way the image is attended, and it implies object recognition (“Recognize the keys”) and knowledge of an object’s usual location (“they could be on the floor, but never on the ceiling”).

7.3.2.1 Object Recognition

Object recognition can be achieved through classical methods or using points of interest (like SIFT, SURF ... [3]) which are somehow related to saliency. Some authors have integrated the notion of object recognition into the architecture of their model like Navalpakkam and Itti [37]. They extract the same features as for the bottom-up model, from the object, and learn them. This learning step will provide

weight modification for the fusion of the conspicuity maps which will lead to the detection of the areas that contain the same feature combination as the learned object. More recently, Kong et al. [18] showed that a classical bottom-up model where different object detectors are added to detect faces, text, people, animals or cars will provide much better results than the bottom-up model alone. This result shows the importance of specific objects in attention detection.

7.3.2.2 Object Location

In addition to object features detection, another approach is to assign greater weight to the areas of the image which have a higher probability to contain the searched for object. Several authors such as Oliva et al. [38] developed methods to learn objects' probable location. Vectors of features are extracted from the images, and their dimension is reduced by using Principal Component Analysis (PCA). Those vectors are then compared to the ones from a database of images containing the given object. Figure 7.13 shows the probable locations of people that have been extracted from the image. This information, combined with bottom-up saliency, leads to the selection of a person sitting down on the left part of the image.

7.3.2.3 Task, Context, and Learning

Recently, learning the salient features becomes more and more popular: The idea here is not to find the rare regions as in bottom-up models, but to find an optimal description of regions which are already known from eye-tracking or mouse-tracking ground truth to be "salient." The learning is based on deep neural networks, sparse coding, and pooling based on large images datasets where the regions of interest are known. The most attended regions based on eye-tracking results are used to train those models which, once trained, will find in novel images similar regions which are likely to be salient.



Fig. 7.13 Bottom-up saliency model inhibited by top-down information to select only salient people. (Adapted from Oliva et al. [38])

After the arrival of DNN-based models for attention modeling, most researchers have switched their research direction to focus more on obtaining an end-to-end DNN saliency model which naturally integrates top-down information. Since 2014, DNNs have changed the saliency paradigm. The deep features were first used in the eDN model Vig et al. [51]. Then, the DeepGaze1 model Kümmerer et al. [22] showed that the DNN features trained on object recognition were very useful for saliency detection. This finding seems logical as objects apparently represent the regions of interest in images. Since then, a variety of models used fine-tuned mixes of features from several deep learning models which naturally incorporated top-down information (i.e., faces, people, texts, ...) during the learning process. However, in Kümmerer et al. [23], the authors showed that bottom-up attention was underestimated by DNN-based models. In their experiments, a simple bottom-up model could outperform a state-of-the-art DNN model when the images contained less top-down information. This demonstrated that DNNs too much neglected the bottom-up aspect of visual attention, and they were mostly trained to detect the attractive top-down objects rather than detect bottom-up saliency itself. Moreover, they could not easily adapt to images in a different context from their training set. In Kong et al. [18], the authors showed that, compared to old detectors which were not accurate enough, current detectors (i.e., face detectors), when mixed to bottom-up saliency maps, provide significantly better visual attention results. It is therefore possible to integrate the top-down information into classical bottom-up attention models in a hand-crafted way. The same is shown in Kong et al. [18] where the authors wonder if DNN-based models truly model bottom-up saliency.

In Kong et al. [17] the tests made on datasets such as O^3 and P^3 datasets Kotseruba et al. [19] show very poor results for DNN-based models. The P^3 dataset evaluates the ability of saliency algorithms to find singleton targets that are defined by color, orientation, and size (without a center bias). The O^3 dataset depicts a scene with multiple objects similar to each other in appearance (distractors) and a singleton (target) defined by color, shape, and size (with center bias). This experiment shows again that DNN-based attention models poorly explain bottom-up attention, but they capture very well top-down information about objects that are usually salient, while some bottom-up information is often present in the training set such as contrast as an eye-catching feature that can be learned. Thus, while some bottom-up saliency can be detected by those models, most of the information which is captured during training concern top-down information from faces, people, animals, text, objects of usual interest, ... This fact seems obvious as rare information is by definition not exhibited enough in the training set, and thus it will not be learned (Fig. 7.14).

DNN-based models are all based on the idea of learning to imitate human eye tracking or mouse tracking on different datasets. One or several encoders are used to extract the most effective features to define the salient regions and one or several decoders to obtain at the end a saliency map as an image. There are dozens of models listed in the MIT/Tubingen saliency benchmark [21] and, for all validation metrics,



Fig. 7.14 Easy-to-learn versus hard-to-learn data

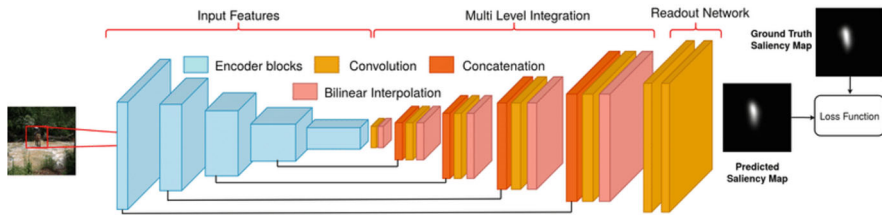


Fig. 7.15 Typical DNN-based architecture, here SimpleNet. (Adapted from Reddy et al. [40])

the best models are based on deep learning architectures. Figure 7.15 shows a basic archetype of DNN-based models including an encoder and decoder with additional links providing better multiresolution information following a U-Net architecture Reddy et al. [40]. A loss function is calculated between the model output and the eye-tracking saliency maps from the training database.

An interesting point with this kind of top-down approaches is that they can be tailored (fine-tuned) to (1) datasets with specific contexts (like outdoor pictures, shopping mall, ...), (2) specific tasks (like looking for wild animals, searching to buy a product, ...), or (3) specific population (like where do women/specialists/learners ... look given a visual task). In that case the initial feature learning phase could exhibit features which are more related to the precise context, task, or population present in the training set.

7.4 Dynamic Overt Saliency

While in the previous sections we simplified Eq. 7.2 by stating that we exclude the fixation dynamics for both bottom-up and top-down approaches, some models intend to take them into account. In this case the output will not only be a saliency map, but a set of fixations and/or a sequence of saliency maps changing over time.

Compared to other Bayesian frameworks, like the one of Oliva et al. [38], overt or visibility models have a dynamic saliency map even for static images, as the map depends on the eye fixations. Indeed, given the spatial resolution drop-off from the fixation point to the periphery (Fig. 7.3), it is clear that some features are well identified in some eye fixation, while less identifiable, or even not visible, during other eye fixations. In contrast to static saliency models, visibility overt models make explicit the variation in resolution across the retina: In that way an attention map is “re-computed” at each new fixation, as the feature visibility changes with every eye movement.

Najemnik and Geisler [36] found that an ideal observer based on a Bayesian framework can predict eye search patterns including the number of saccades needed to find a target, the amount of time needed, as well as the spatial distribution of saccades (Fig. 7.16).

Other authors like Legge et al. [24] proposed a visibility model capable of predicting eye fixations during the task of reading. In the same way, Reninger used similar approaches for the task of shape recognition. Tatler [48] introduces a tendency of the eye gaze to stay in the middle of the scene to maximize the visibility over the image (which reminds the centered preference for natural images or centered Gaussian bias illustrated in Fig. 7.12).

Those visibility models are more often used in the case of strong tasks, and few of them are applied to free viewing which is considered as a weak task [9].

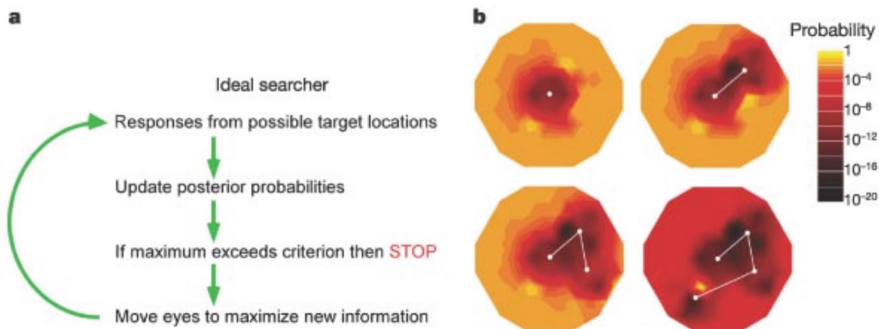


Fig. 7.16 Optimal search pipeline and saliency map evolution with superimposed saccade sequence. (Adapted from Najemnik and Geisler [36])

In parallel with the visibility overt models, some saliency models are adapted to provide dynamic scan paths (and/or priority maps which are saliency maps relevant given the current fixation r_i (see Eq. 7.2). A state of the art of different models can be found in [20]. The author provides a classification of those models into several classes:

- “biologically inspired” with models trying to mimic biological mechanisms such as fixations, inhibition of return on saliency maps, or working memory
- “statistically inspired” with models focusing on reproducing statistical properties of scan paths such as their density distribution or the fact that patches with features similar to already visited patches have smaller occurrence probabilities
- “cognitively inspired” with models which will focus on regions having a semantic meaning
- “engineered models” where data is taken into account only by learning. This class of models uses learning.

The author also shows that almost any existing static saliency model can be adapted to also produce a scan path as there is an internal model state which is built over the previous fixations in almost all models. It is therefore possible to implement existing models to provide priority maps (saliency map based on previous fixation) and use for example the maximum of this priority map as the next fixation predictor.

With the arrival of deep learning, some models indeed can provide scan paths based only on free-viewing datasets with no priors on how biology works, or scan-paths statistics such as ScanPath [2] which uses a Generative Adversarial Network (GAN) to synthesize the scan path.

Other models intend to integrate the task into the scan path to include additional top-down information. Yang et al. [52], Chen et al. [5] and Yang et al. [53] are all based on the idea of encoding specific search targets. Two drawbacks can be found here: (1) the limitation to the relatively small number of encoded target objects (which need segmentation) and (2) the need of a laborious collection of large-scale datasets of search behavior for each target category. Those two important issues do not allow those types of models to work in real life. However, new work in this area has been done with architectures like GazeFormer Mondal et al. [34] where a language model is mixed to an image model. In that case it is sufficient to provide a prompt about the target which is needed to obtain the scan path (position and duration for each fixation). The image is encoded via a Transformer encoder, while the text prompt is encoded via RoBERTa language model Liu et al. [25]. The two embedding vectors (of the image patch and text prompt) are then concatenated and used by a Transformer decoder to output a scan path. Figure 7.17 shows the results of different scan paths from GazeFormer Mondal et al. [34], IRL Yang et al. [52], Chen et al. [5], and FFM Yang et al. [53] where the fixation order is written in the fixation circles and the size of the circles is proportional to the fixation duration.

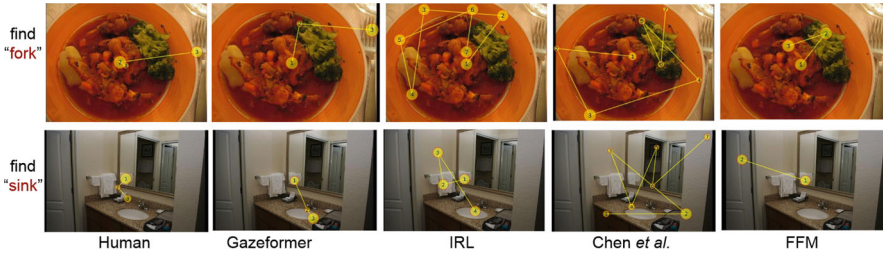


Fig. 7.17 Example of task-based search and provided saccades. The models compared are GazeFormer Mondal et al. [34], IRL Yang et al. [52], Chen et al. [5], and FFM Yang et al. [53]. (Adapted from Mondal et al. [34])

7.5 Modeling Attention in Computer Science

In computer science, visual attention is represented by saliency maps, salient object detection (SOD) maps, or scan paths.

In this chapter, we presented a general saliency framework including (1) bottom-up and top-down approaches and (2) static and dynamic attention.

In the first part we focused on the idea of bottom-up attention and the idea of rarity or surprise of a region in a given context.

We then dealt with top-down attention which can come from the context, tasks, or semantic information.

We described dynamic models of saliency which can provide scan paths or priority maps (temporal saliency maps depending on the previously predicted eye fixation).

The deep-learning-based models begin to be able to provide interesting static and dynamic information about human attention. However, for the moment they (1) do not include enough a priori knowledge about scan-paths statistics and they (2) are not able to capture surprising bottom-up information correctly. One of the next steps might be to bring back into the game this bottom-up rare information and scan-paths statistics.

An important future point is the generalization of text prompts for inducing adaptive contexts or tasks to modulate attention. Visual search is an important future trend as its industrial applications are numerous and its relation to biological attention is high.

Finally, another important topic is the use of multimodal data which are increasingly available on new sensors. Chaps. 10 and 11 will focus on multimodal attention and the data necessary for training which is more and more essential in deep neural networks architectures.

Acknowledgments Section 7.1 includes a short part of the work of the late Dr. Fred Stentiford in the previous book edition.

References

1. Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. <http://www.cvpr2009.org/>
2. Assens, M., Giro-i Nieto, X., McGuinness, K., & O'Connor, N. E. (2018). PathGAN: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*
3. Bay, H., Ess, A., Tuytelaars, T., & Gool, L. V. (2008). SURF: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 110(3), 346–359.
4. Boiman, O., & Irani, M. (2007). Detecting irregularities in images and in video. *International Journal of Computer Vision*, 74(1), 17–31.
5. Chen, X., Jiang, M., & Zhao, Q. (2021). Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10876–10885).
6. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE.
7. Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1373), 1245–1255.
8. Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, 8(2), 6.
9. Geisler, W. S., & Cormack, L. (2011). *Chapter 24: Models of overt attention, in the Oxford handbook of eye movements*. Oxford University Press.
10. Geisler, W. S., & Cormack, L. K. (2011). Models of overt attention. In *The Oxford handbook of eye movements* (pp. 439–454).
11. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. *2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). IEEE.
12. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR '07* (pp. 1–8).
13. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), 1254–1259.
14. Jouneau, E., & Carincotte, C. (2011). Particle-based tracking model for automatic anomaly detection. In *IEEE Int. Conference on Image Processing (ICIP)*.
15. Judd, T., Ehinger, K., Durand & Torralba, A. (2009). Learning to predict where humans look. *IEEE International Conference on Computer Vision (ICCV)* (pp. 2376–2383).
16. Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4), 219–227.
17. Kong, P., Mancas, M., Gosselin, B., & Po, K. (2022). DeepRare: Generic unsupervised visual attention models. *Electronics*, 11(11), 1696.
18. Kong, P., Mancas, M., Thuon, N., Kheang, S., & Gosselin, B. (2018). Do deep-learning saliency models really model saliency? *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 2331–2335). IEEE
19. Kotscheruba, I., Wloka, C., Rasouli, A., & Tsotsos, J. K. (2020). Do saliency models detect odd-one-out targets? New datasets and evaluations. arXiv preprint arXiv:2005.06583.
20. Kümmerer, M., & Bethge, M. (2021). State-of-the-art in human scanpath prediction. arXiv preprint arXiv:2102.12239.
21. Kümmerer, M., Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A., & Bethge, M. (n.d.). Mit/tübingen saliency benchmark. <https://saliency.tuebingen.ai/>

22. Kümmerer, M., Theis, L., & Bethge, M. (2014). Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet. arXiv preprint arXiv:1411.1045.
23. Kümmerer, M., Wallis, T. S., Gatys, L. A., & Bethge, M. (2017). Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4789–4798).
24. Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. S., & Tjan, B. S. (2002). Mr. Chips 2002: New insights from an ideal observer model of reading. *Vision Research*, 42(18), 2219–2234.
25. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
26. Mahdi, A., & Qin, J. (2019). DeepFeat: A bottom up and top down saliency model based on deep features of convolutional neural nets, *IEEE Transactions on Cognitive and Developmental Systems*, 12(1), 54–63.
27. Mancas, M. (2007). *Computational attention towards attentive computers*. Presses Universitaires de Louvain.
28. Mancas, M. (2009). Relative influence of bottom-up and top-down attention. In *Attention in Cognitive Systems*, Vol. 5395 of *Lecture Notes in Computer Science*. Springer
29. Mancas, M., & Gosselin, B. (2010). Dense crowd analysis through bottom-up and top-down attention. In *Processing of the Brain Inspired Cognitive Systems (BICS 2019)*.
30. Mancas, M., Gosselin, B., & Macq, B. (2007). Perceptual image representation. *Journal on Image and Video Processing*. <https://doi.org/10.1155/2007/98181>
31. Mancas, M., Kong, P., & Gosselin, B. (2020). Visual attention: Deep rare features. In *2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)* (pp. 1–6). IEEE
32. McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus, *Journal of Cognitive Neuroscience*, 9(5), 605–610.
33. Millidge, B., & Shillcock, R. (2018). A predictive processing account of bottom-up visual saliency using cross-predicting autoencoders.
34. Mondal, S., Yang, Z., Ahn, S., Samaras, D., Zelinsky, G., & Hoai, M. (2023). Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1441–1450).
35. Mumtaz, A., Sargano, A. B., & Habib, Z. (2024). AnomalyNet: A spatiotemporal motion-aware CNN approach for detecting anomalies in real-world autonomous surveillance. *The Visual Computer*, 40(11), 7823–7844.
36. Najemnik, J., & Geisler, W. (2005). Optimal eye movement strategies in visual search. *Nature*, 434(7031), 387–391.
37. Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, 45(2), 205–231.
38. Oliva, A., Torralba, A., Castelhano, M., & Henderson, J. (2003). Top-down control of visual attention in object detection. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on* (Vol. 1, pp. 1–253–6).
39. Parker, A. (2016). *In the blink of an eye: How vision kick-started the big bang of evolution*. Natural History Museum.
40. Reddy, N., Jain, S., Yarlagadda, P., & Gandhi, V. (2020). Tidying deep saliency prediction architectures. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 10241–10247). IEEE
41. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B. & Dutoit, T. (2013). Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6), 642–658.
42. Sabokrou, M., Fayyaz, M., Fathy, M., & Klette, R. (2017). Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, 26(4), 1992–2004.

43. Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., & Klette, R. (2018). Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, 172, 88–97.
44. Seo, H. J., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12), 15–15. <http://www.journalofvision.org/content/9/12/15.abstract>
45. Shah, A. P., Lamare, J.-B., Nguyen-Anh, T., & Hauptmann, A. (2018). CADP: A novel dataset for CCTV traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1–9). IEEE
46. Stentiford, F. (2001). An estimator for visual attention through competitive novelty with application to image compression. In *Picture Coding Symposium* (pp. 25–27).
47. Sun, X. (2018). Semantic and contrast-aware saliency. arXiv preprint arXiv:1811.03736.
48. Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7, 4.
49. Torralba, A., Oliva, A., Castelano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113(4), 766.
50. Treisman, A. (1985). Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31(2), 156–177.
51. Vig, E., Dorr, M., & Cox, D. (2014). Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2798–2805).
52. Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D. & Hoai, M. (2020). Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 193–202).
53. Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., & Samaras, D. (2022). Target-absent human attention. In *European Conference on Computer Vision* (pp. 52–68). Springer.
54. Zhang, R., Isola, P., & Efros, A. A. (2017). Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1058–1067).
55. Zhou, J. T., Du, J., Zhu, H., Peng, X., Liu, Y., & Goh, R. S. M. (2019). AnomalyNet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, 14(10), 2537–2550.