UDC 004.932
doi:10.31799/1684-8853-2019-3-10-36

Surveys

Saliency detection in deep learning era: trends of development

M. N. Favorskaya[a], *Dr. Sc., Tech., Professor, orcid.org/0000-0002-2181-0454, favorskaya@sibsau.ru*
L. C. Jain[b,c,d], *PhD, Professor, orcid.org/0000-0001-6176-3739*
[a]*Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Ave., 660037 Krasnoyarsk, Russian Federation*
[b]*University of Canberra, 11 Kirinari St., Bruce ACT 2617, Canberra, Australia*
[c]*Liverpool Hope University, Hope Park, Liverpool L16 9JD, UK*
[d]*University of Technology Sydney, PO Box 123, Broadway NSW 2007, Sydney, Australia*

Introduction: *Saliency detection is a fundamental task of computer vision. Its ultimate aim is to localize the objects of interest that grab human visual attention with respect to the rest of the image. A great variety of saliency models based on different approaches was developed since 1990s. In recent years, the saliency detection has become one of actively studied topic in the theory of Convolutional Neural Network (CNN). Many original decisions using CNNs were proposed for salient object detection and, even, event detection.* ***Purpose:*** *A detailed survey of saliency detection methods in deep learning era allows to understand the current possibilities of CNN approach for visual analysis conducted by the human eyes' tracking and digital image processing.* ***Results:*** *A survey reflects the recent advances in saliency detection using CNNs. Different models available in literature, such as static and dynamic 2D CNNs for salient object detection and 3D CNNs for salient event detection are discussed in the chronological order. It is worth noting that automatic salient event detection in durable videos became possible using the recently appeared 3D CNN combining with 2D CNN for salient audio detection. Also in this article, we have presented a short description of public image and video datasets with annotated salient objects or events, as well as the often used metrics for the results' evaluation.* ***Practical relevance:*** *This survey is considered as a contribution in the study of rapidly developed deep learning methods with respect to the saliency detection in the images and videos.*

Keywords *— salient region detection, salient object detection, salient event detection, deep learning, convolutional neural network, feature extraction.*

For citation: Favorskaya M. N., Jain L. C. Saliency detection in deep learning era: trends of development. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2019, no. 3, pp. 10–36. doi:10.31799/1684-8853-2019-3-10-36

Introduction

Saliency detection is a fundamental task in computer vision and includes three aspects, such as the region-based detection, object-based detection, and event-based detection. Salient regions are identified for integrating the entire segments into salient objects. If the goal of object detection is to find and identify a visual object, then the salient object detection means to retrieve an object, which attracts a human attention. The main proposition of these algorithms is in that a human vision focuses on the most distinctive parts of image or video sequence without any prior knowledge. Recently, a salient event detection is applied as a promising technique in video annotation.

The first saliency detection algorithms were proposed three decades ago [1], and since that time this topic attracts many researchers, who suggested numerous modifications of saliency detection for various computer vision tasks, such as image classification [2], person re-identification [3], image resizing [4], image inpainting [5], image cropping [6], image search [7], robot vision [8], and video summarization [9].

Saliency detection algorithms can be classified as two opposite types: top-down (faster, subconscious, and data-driven saliency extraction) and bottom-up (slower, task-related, and knowledge-driven saliency extraction). Bottom-up models have been widely studied in cognitive fields. Note that with the emergence of Convolutional Neural Networks (CNN), a data-driven model becomes more high-level data model [10].

Generally speaking, top-down saliency detection is applied to goal-oriented detection, when prior knowledge regarding the object characteristics is known. According to the task goal, the top-down saliency algorithms design distinctive features, which are defined manually [11, 12] or automatically, e.g. using networks [13]. The low-level features easily extracted, such as edges, straight lines, and corners, allow to construct the salient regions. Thus, in [14] it was introduced five image objectness cues (multi-scale saliency of a window, color contrast, edge density, superpixels straddling, and location and size of a window), which localized the salient objects with a good visibility in a Bayesian framework. The famous histograms of oriented gradients for a human detection were presented in [15]. This method combined different features, such as fine-scale gradients, fine orientation binning, coarse spatial binning, and local-contrast normalization of overlapping blocks, that allowed to achieve the effective results of human recognition.

Bottom-up saliency detection focus on visual stimuli from the image or video scene, and the main incentives are the contrast or movement, respectively. Bottom-up saliency methods are not relevant to the task and more flexible. The majority of these methods were developed during the last few decades. Thus, the saliency score of a certain region was calculated as the sum of contrast values between one region and all the others in [16]. The color contrast of larger color regions was estimated based on the saliency scores, spatial weighting strength, and spatial distance between regions in the Euclidean metric. Some saliency algorithms exploit the idea that the regions with similar colors distributed more widely do not attract the human vision attention [17]. The idea of saliency filters and descriptors of saliency detection features has been explored since 2010s [18]. Extraction of saliency detection features had been remained manual low-level sub-task until novel machine learning techniques, such as CNNs, did not emerge.

Generations of saliency detection methods

Saliency detection means a detection of the most attractive for human vision part of image or frame. Usually, saliency detection is interpreted as the salient object detection. However, the term "saliency detection" is widely used and includes the salient region detection as the first attempts of promoting this technique. At the same time, the salient event detection reflects the recent advantages in video analysis.

The *first generation of saliency models* was based on the multiple disciplines including, first of all, cognitive psychology and neuroscience, and then computer vision. Some fundamental investigations in the cognitive and psychological theories of bottom-up attention [19–21] influenced strongly on the development of the earliest saliency algorithms. Even first saliency models, which used multi-scale color contrast, intensity contrast, and orientation contrast maps processed by dynamic neural network [1], were able to detect conspicuous locations in scenes. At that time, evaluation of saliency detection methods was implemented by subsequent behavioral and computational investigations [22, 23].

The *second generation* is referred to 2000s, when a saliency detection was developed as a binary segmentation problem [24, 25]. This concept was inspired by salient regions and proto-objects detection [26, 27] and had led to tremendous publications in this scope. At present, such representation of saliency detection remains the main type of representation. The crucial issue is how saliency detection relates to such popular tasks of computer vision as image segmentation, object detection, and object generation.

The *third generation of saliency models* deals with CNNs propagation [28–30]. In contradistinction to classic methods based on contrast cues [16, 18, 31–33], the CNNs-based methods eliminate a necessity of handcrafted features extraction and facilitate a dependency on center bias knowledge. Usually CNN contains million of tunable parameters distributed in raw layers, which extract low-level features, and fine layers, which provide high-level features. Therefore, global and local information highlighting salient regions and their boundaries can be obtained. Since 2012, it is considered that CNN models perform in accuracy parameters the handcrafted feature-based models for pattern recognition, and saliency detection models are not exclusion. Thus, the mainstream direction in saliency detection is becoming CNN models [34]. More, this concept is propagated on video saliency detection, when the researchers are passing from the contrast analysis [35] to CNN-based models [36, 37].

Generally speaking, any saliency detection model should meet the following three criteria:

– saliency detection with low values of errors including missing salient regions and falsely marking of non-salient regions;

– high resolution of saliency maps for accurate salient objects localization;

– computational efficiency, especially fast salient regions detection.

At present, none of existing saliency detection algorithms satisfies to these criteria fully that causes a necessity to continue investigations in this scope.

Development of saliency detection methods

Region-based saliency detection methods

The first classical approaches were oriented on the pixel-based saliency models based on the local center-surround differences due to low computational cost [1, 38]. Such models employed the sliding integration windows to estimate the center and surrounding appearances. The global compactness of color was another statement in saliency detection [18, 39]. This approach can be concerned to the region-based saliency detection, when a content-aware segmentation was a preliminary step before saliency detection in each extracted segment instead of each pixel [32, 40].

The invention of algorithm contributed significantly in region-based saliency detection [41]. The followers used this algorithm in many tasks of computer vision, including saliency detection task [42, 43].

Thus, in [43] a segment-based saliency detection method was based on the superpixels computed in multiple scales. The difference between superpixels was measured with the Wasserstein distance on L_2 norm (W_2-distance). First, the simple linear iterative clustering algorithm extracted the superpixels

of more than 64 pixels at the finest scale because too small regions make the appearance distribution estimation less meaningful. Second, this procedure was repeated for three scales starting from the finest scale by decreasing the number of demanded superpixels by a factor of two.

Object-based saliency detection methods

Object-based detection models generate a salient object bounding box through segmenting the salient object based on the saliency maps [44]. Salient object detection has a significant meaning in many practical applications, such as image cropping [45], adaptive image display on mobile devices [46], extracting dominant colors on the object of interest for web image filter [47], image segmentation [48], visual tracking [49], among others.

Usually, the main intensity, color, contrast (including the color contrast, texture energy contrast, and texture gradient contrast) normalized maps, as well as the additional edge, angle, and symmetry saliency maps, are built with further fusion of all types of maps [50]. Such algorithms differ in details of implementation.

However, a localizing the salient objects is always a very challenging problem that could not be solved sufficiently many years ago. We can refer to the challenges, such as various visual characteristics of objects, cluttered background, sometimes low resolution of images, and blurring. The CNNs application allows to overcome the main problems.

Rapid development of deep learning techniques led to the emergence of CNNs with different architectures tuned for practical applications [51–56]. Due to their properties, CNNs provide richer features, which allow to detect the salient object simultaneously on the lower and higher levels using the extracted low-level and high-level features, respectively. A family of CNN-based salient object detection methods includes many interesting decisions with the outstanding results. Let us consider some of them.

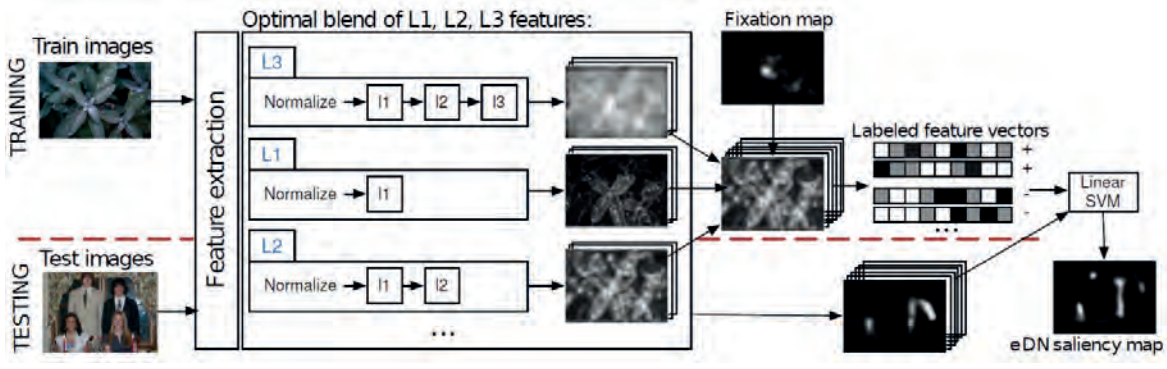
Tables 1 and 2 provide a description of static and dynamic CNN-based saliency models, respectively. Each model is explained by its architecture.

■ ***Table 1***. Static CNN-based saliency models

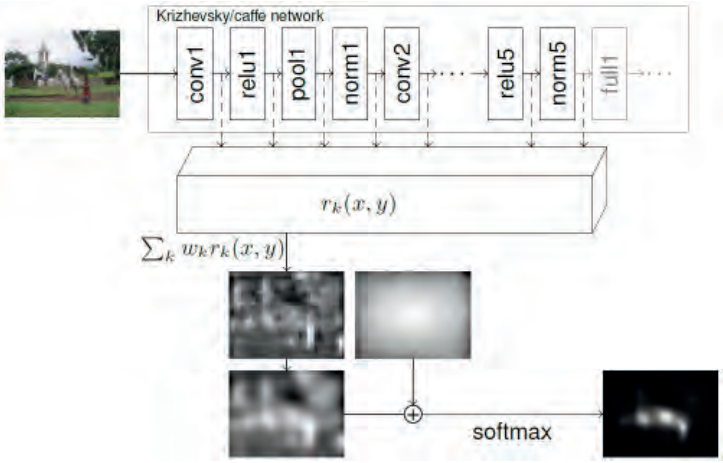
Caption	Description
eDN (ensembles of Deep Networks), 2014	The eDN is the first attempt to apply a prototype of CNN for image saliency prediction (Fig. 1). First, a large number of 1–3 layered networks using biology inspired hierarchical features are constructed. Second, the independent models have been searched by hyper-parameter optimization. Third, the independent models combine into a single model by training a linear SVM [57]
DeepGaze I and DeepGaze II, 2014 and 2017	DeepGaze I is a relatively deeper CNN pre-trained by AlexNet and involved five layers [58]. After convolutional layers, a linear model computes an image salience (Fig. 2). Hereinafter, DeepGaze II [59] built upon DeepGaze I was implemented. Both networks explore the unique contributions between the low-level and high-level features towards a fixation prediction
Mr-CNN (Multi-resolution CNN), 2015	The multi-resolution three-layered Mr-CNN implements the automated learning of early features, bottom-up saliency, top-down factors, and their integration simultaneously using an eye-tracking mechanism [60]. Mr-CNN are learnt both low-level features related to bottom-up saliency and high-level features related to top-down factors in order to improve eye fixation prediction. The fixation and non-fixation image regions are extracted for training Mr-CNN (Fig. 3)
SALICON (SALiency In CONtext), 2015	SALICON uses the elements of AlexNet, VGG-16, and GoogLeNet architectures in order to provide a narrow semantic gap between the predicting eye fixations and strong semantic content [61]. It combines information of high-level and coarse-level semantics encoded in deep neural network pretrained in ImageNet for object recognition. Then both branches are concatenated to produce the final saliency map (Fig. 4)
ML-Net (Multi-Level Network), 2016	ML-Net combines the features extracted from different CNN levels [62]. It is composed of three main blocks: feature extraction CNN, feature encoding network that weights the low and high level feature maps, and prior learning network (Fig. 5)
JuntingNet and SalNet (Junting is the name of the main author and SALiency Network), 2016	This approach proposes two different architectures: a shallow CNN (JuntingNet), trained from scratch, and a deep CNN (SalNet) that reuses parameters from the bottom three layer of a network previously trained for classification [63]. The JuntingNet is inspired by the AlexNet and uses three convolutional and two fully connected layers, which are all randomly initialized. The SalNet contains eight convolutional layers with the first three being initialized from the VGG network (Fig. 6)
PDP (Probability Distribution Prediction), 2016	PDP CNN employs a new saliency map model, which formulates a map as a generalized Bernoulli distribution [64]. PDP CNN is trained using a novel loss functions, which pair the softmax activation function with measures designed to compute distances between probability distributions (Fig. 7). Experiments showed that new loss functions are more efficient than traditional loss functions

■ *Table 1 (completed)*

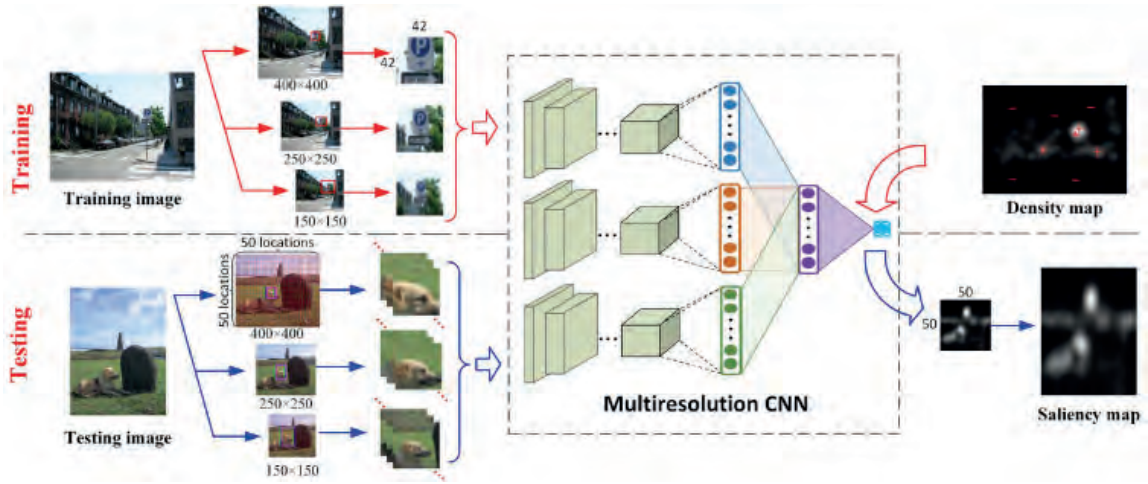
Caption	Description
DSCLRCN (Deep Spatial Contextual Long-term Recurrent Convolutional Neural network), 2016	DSCLRCN, first, learns local saliency of small image regions using CNN [65]. Then, it scans the image both horizontally and vertically using a deep spatial long short-term memory model to capture a global context. These two operations allow DSCLRCN to incorporate simultaneously and effectively the local and global contexts to infer an image saliency (Fig. 8)
FUCOS (FUlly COnvolutional Saliency), 2016	FUCOS (Fig. 9) is applied to either gaze, or salient object prediction [66]. It integrates pre-trained layers from large-scale CNN models and is then fine-tuned on PASCAL-Context dataset [67]
SAM Net (Saliency Attentive Models), 2016	The core of SAM Net is Attentive Convolutional Long Short-Term Memory network (Attentive ConvLSTM) that focuses on the most salient regions of the input image to iteratively refine the predicted saliency map [68]. SAM Net combines a fully convolutional network with a recurrent convolutional network, endowed with a spatial attentive mechanism (Fig. 10)
ELM (Extreme Learning Machines), 2016	Ensemble of ELM [69] is based on a saliency model based on inter-image similarities and ensemble of extreme learning machine [70]. Firstly, a set of images similar to a given image is retrieved. A saliency predictor is then learned on this set using ELM and forming an ensemble. Finally, the saliency maps provided by the ensemble's members are averaged in order to construct the final map (Fig. 11)
DeepFix (Deep Fixation), 2017	DeepFix is a first fully CNN for accurate saliency prediction, which captures semantics at multiple scales with very large receptive fields [71]. It also incorporates Gaussian priors to further improve the learned weights (Fig. 12). Fully convolutional nets are spatially invariant that prevents them from modeling the location dependent patterns (e. g. centre-bias)
SalGAN (Saliency Generative Adversarial Network), 2017	SalGAN model [72] is the extended version of GANs [73]. It includes two networks: generator and discriminator. The generator is trained via back-propagation using a binary cross entropy (adversarial) loss on existing saliency maps. Then the result is passed to the discriminator that is trained to identify whether a saliency map was synthesized by the generator or built using a ground truth (Fig. 13)
DVA (Deep Visual Attention), 2017	In DVA model, an encoder-decoder architecture is trained over multiple scales to predict pixel-wise saliency [74]. The encoder network is topologically identical to the first 13 convolutional layers in the VGG-16 network and decoder network is used to map the low resolution encoder feature maps into dense full-input-resolution feature maps. DVA captures hierarchical global and local saliency information. It is based on a skip-layer network structure. Final multi-level saliency prediction is achieved via a combination of the global and local predictions (Fig. 14)
Attentional Push, 2017	Attention Push model based on shared attention considers a viewer of a scene actor and uses it to augment image salience [75]. It contains two pathways: an Attentional Push pathway, which learns the gaze location of the scene actors, and a saliency pathway. However, the limitation is that it is required to find the actor' head location respect to the camera (Fig. 15)
EML-NET (Expandable Multi-Layer NETwork), 2018	EML-NET is a scalable model, in which the encoder and decoder components are separately trained [76]. The encoder can contain more than one CNN model to extract features, and these models can have different architectures or be pre-trained on different datasets (Fig. 16)



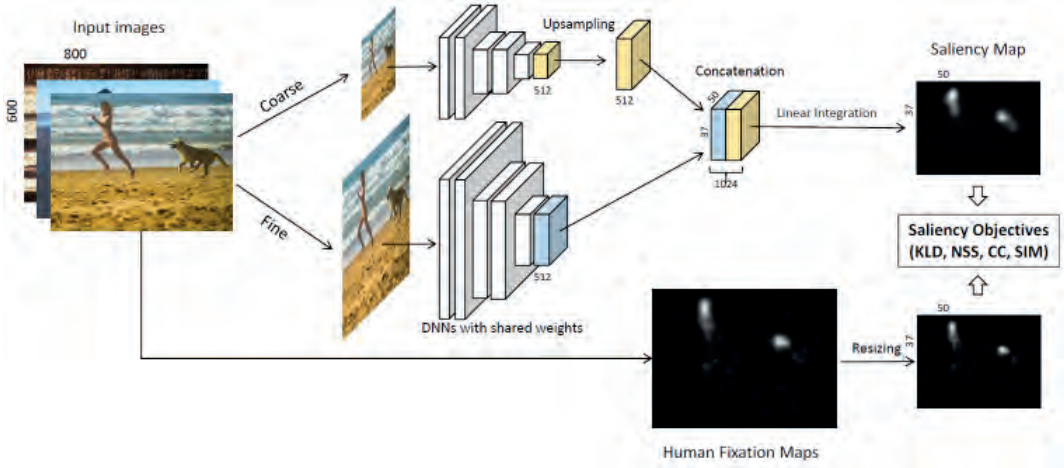
■ ***Fig. 1.*** The eDN architecture. Good multilayer feature extractors are found by a guided hyperparameter search (not shown) and combined into an optimal blend. Resulting feature vectors are labeled with empirical gaze data and fed into a linear SVM [57]



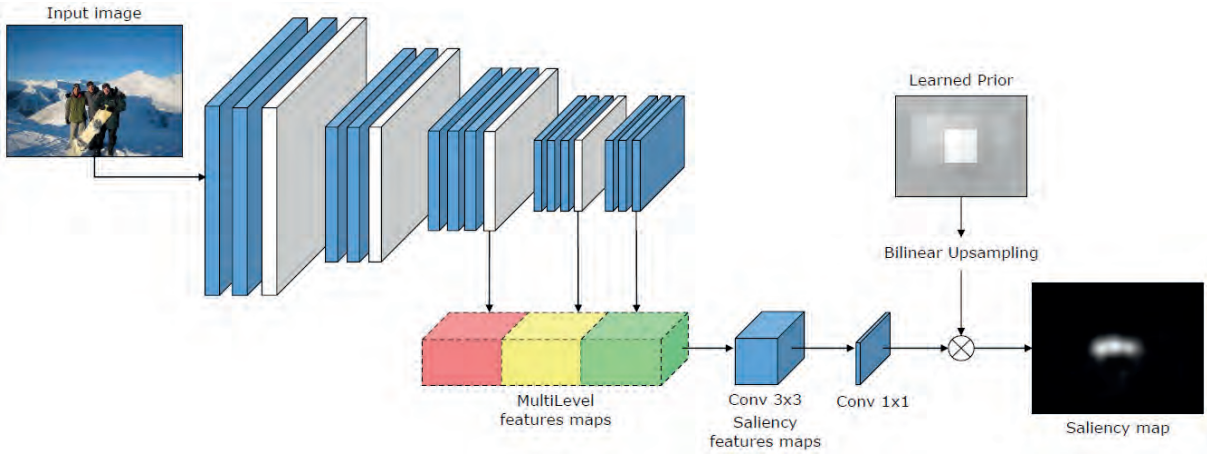
■ ***Fig. 2.*** The DeepGaze I architecture. The image is first downsampled and preprocessed with the Krizhevsky network. The responses of the layers are scaled up to the size of the largest network layer and normalized to have unit standard deviation. This list of maps is linearly combined and blurred with a Gaussian kernel. The model output is fed through a softmax rectification, yielding 2D probability distribution, which is used to compensate for the central fixation bias [58]



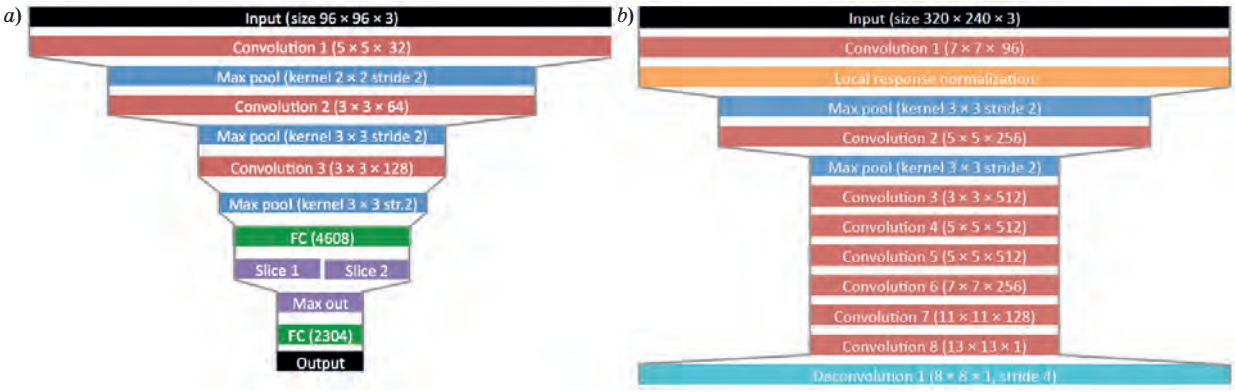
■ ***Fig. 3.*** The Mr-CNN architecture. The original image is rescaled to three scales (150 × 150, 250 × 250, and 400 × 400). The extracted 42 × 42 sized image regions with the same center locations are inputs to Mr-CNN. During testing, 50 × 50 sized samples are used to estimate their saliency values in order to reduce computation cost. The obtained down-sampled saliency map is rescaled to the original size [60]



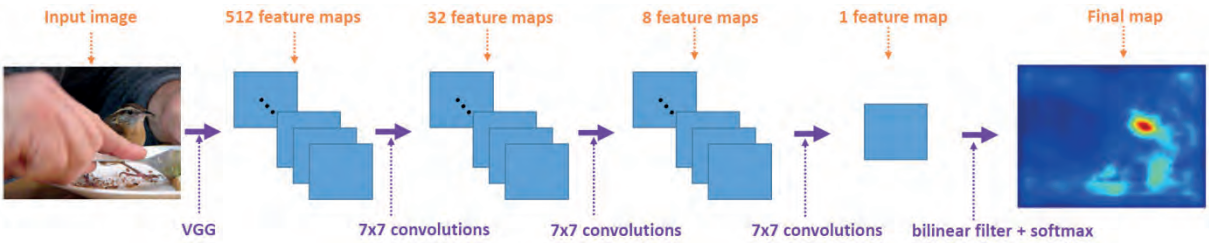
■ ***Fig. 4.*** The SALICON architecture consists of deep neural network applied at two different image scales. The last convolutional layer in the pretrained network feeds a randomly initialized convolutional layer with one filter that detects the salient regions. The parameters are learnt end-to-end with back-propagation [61]



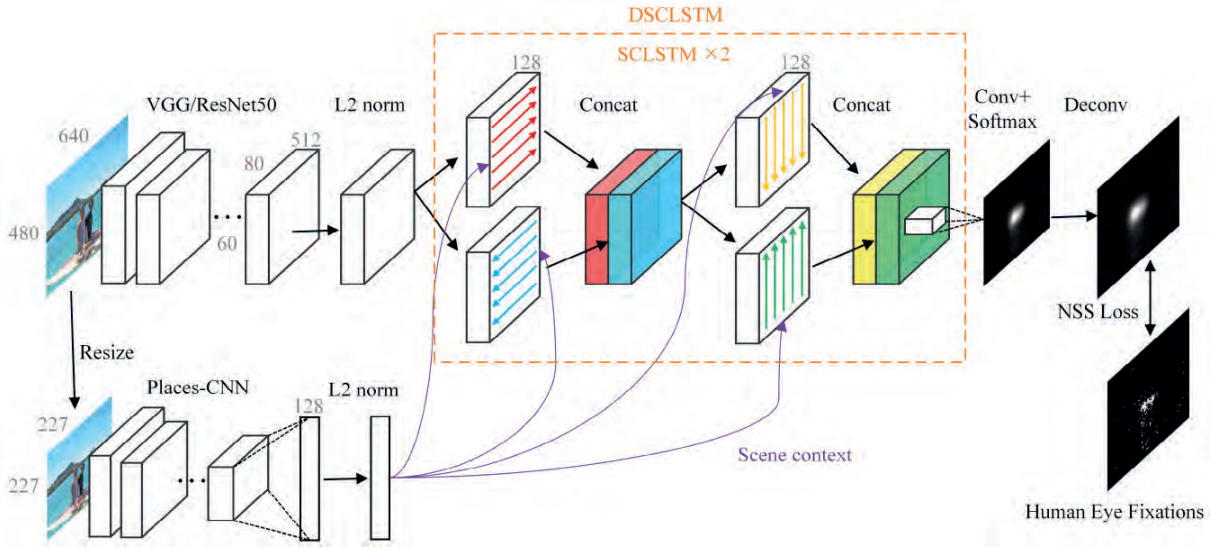
■ ***Fig. 5***. The ML-Net architecture computes the low and high level features from the input image. Extracted features maps are then fed to an Encoding network, which learns a feature weighting function to generate saliency-specific feature maps. A prior image is also learned and applied to the predicted saliency map [62]



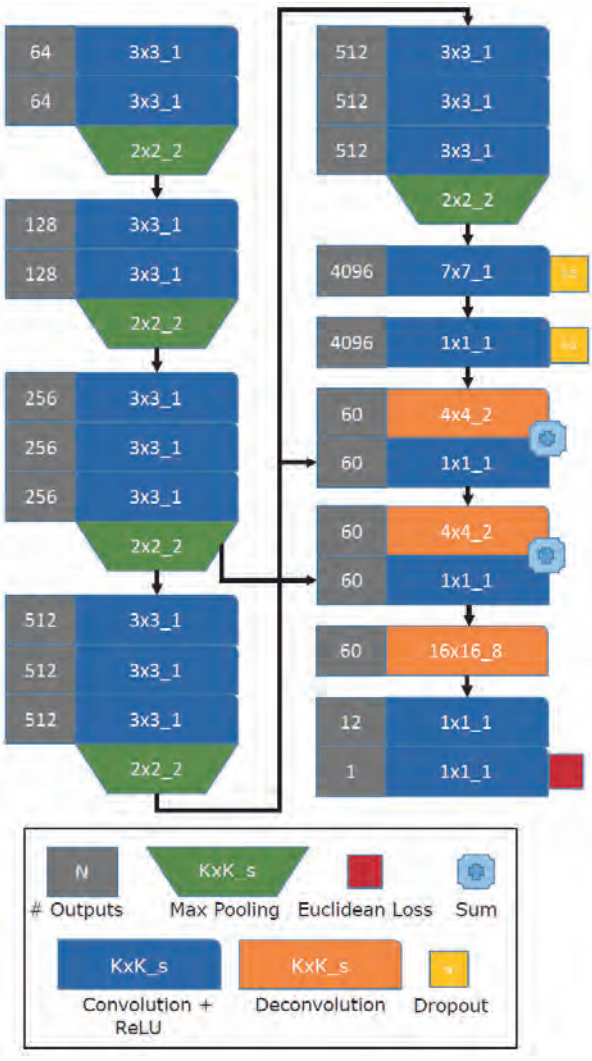
■ ***Fig. 6***. The JuntingNet and SalNet architectures: *a* — JuntingNet has to a total of 64.4 million free parameters and uses lesser number of convolutional layers regarding AlexNet and VGG-16. The input images are resized to 96 × 96. The three max pooling layers reduce the initial 96 × 96 feature maps down to 10 × 10 by the last of the three pooling layers; *b* — SalNet is composed of 10 weight layers and a total of 25.8 million parameters. The architecture of the first three weight layers is compatible with the VGG layers [63]



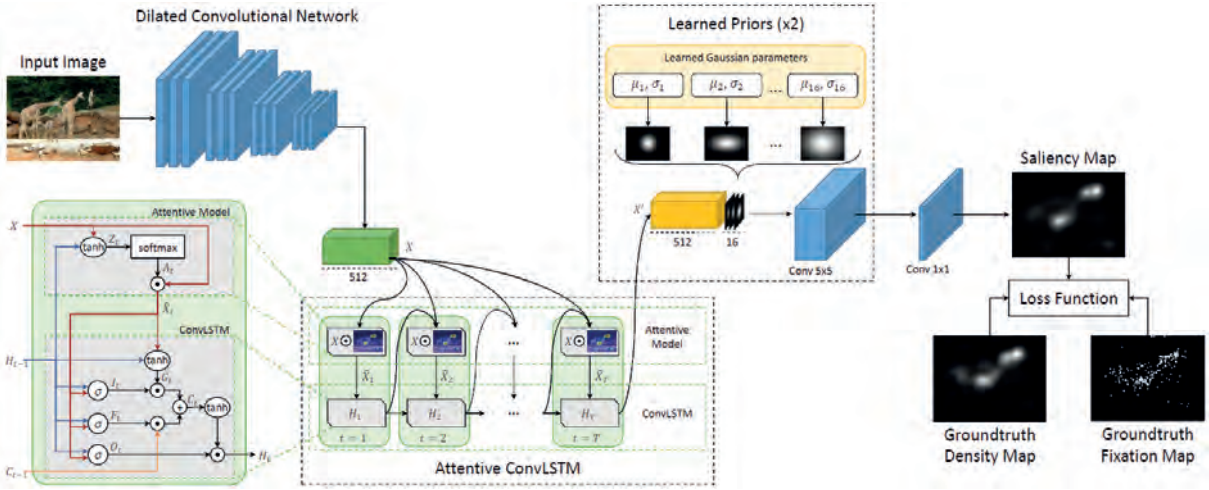
■ ***Fig. 7***. The PDP CNN architecture. The input image is introduced into a CNN similar to VGGNet. Additional convolutional layers are then applied, resulting in a single response map which is upsampled and softmax-normalized to produce a final saliency map [64]



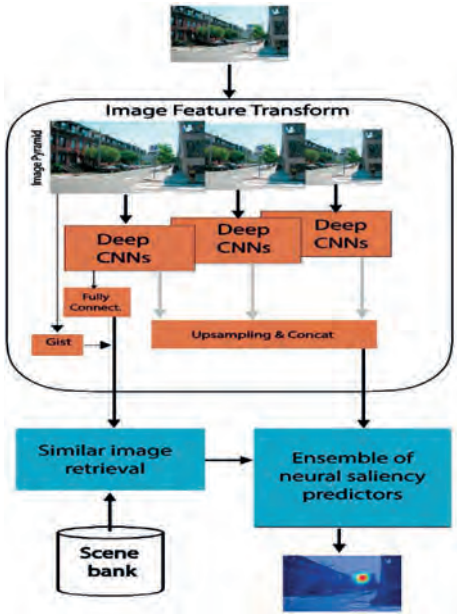
■ ***Fig. 8***. The DSCLRCN architecture. First, local feature map and scene feature are extracted using pretrained CNNs. Then, a DSCLSTM model is adopted to simultaneously incorporate global context and scene context. Finally, saliency map is generated and upsampled [65]



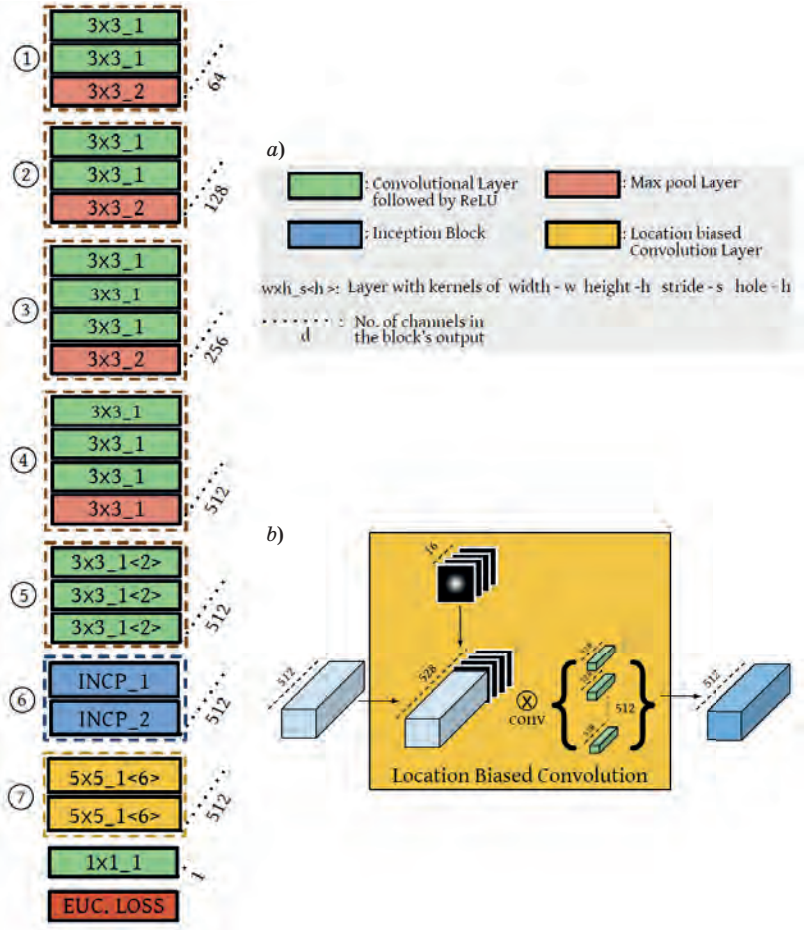
■ ***Fig. 9***. The FUCOS architecture: K corresponds to kernel dimensions; s — to stride; N — to number of outputs; % — to the percent of dropout units [66]



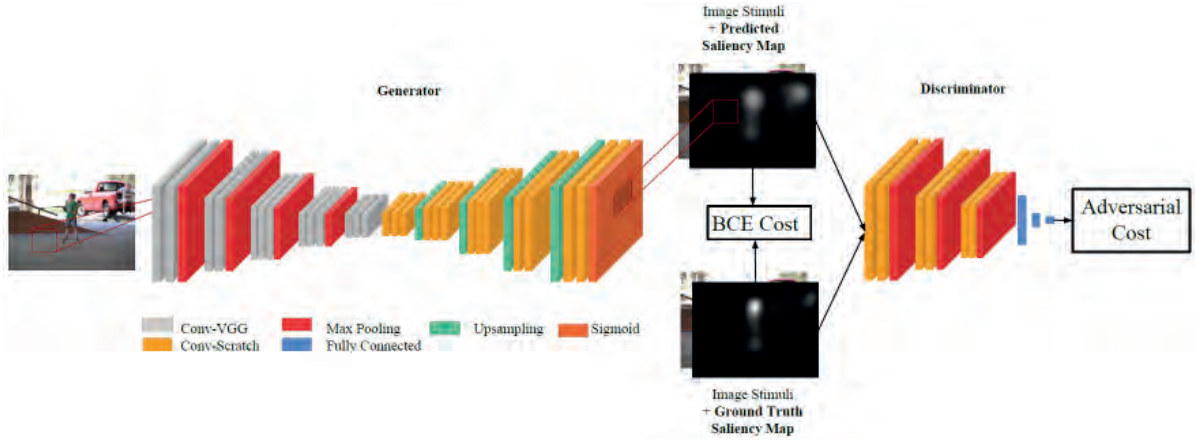
■ ***Fig. 10***. The SAM Net architecture. After computing a set of feature maps on the input image through architecture called Dilated Convolutional Network, an Attentive Convolutional LSTM sequentially enhances saliency features thanks to an attentive recurrent mechanism. Predictions are then combined with multiple learned priors to model the tendency of humans to fix the center region of the image [68]



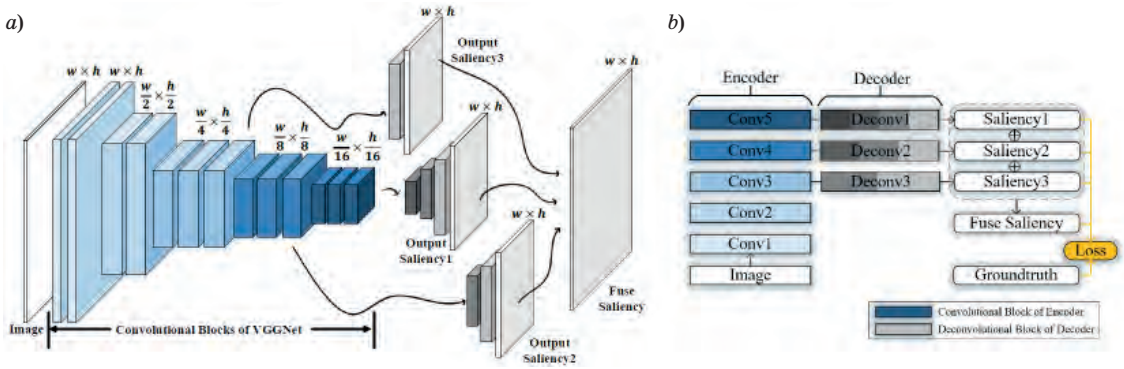
■ ***Fig. 11***. The ELM architecture. The image feature transform performs produces a pool of features. The similar image retrieval finds the top most similar images, stored in the scene bank. Then Ensemble of neural saliency predictors forms a prediction saliency map [70]



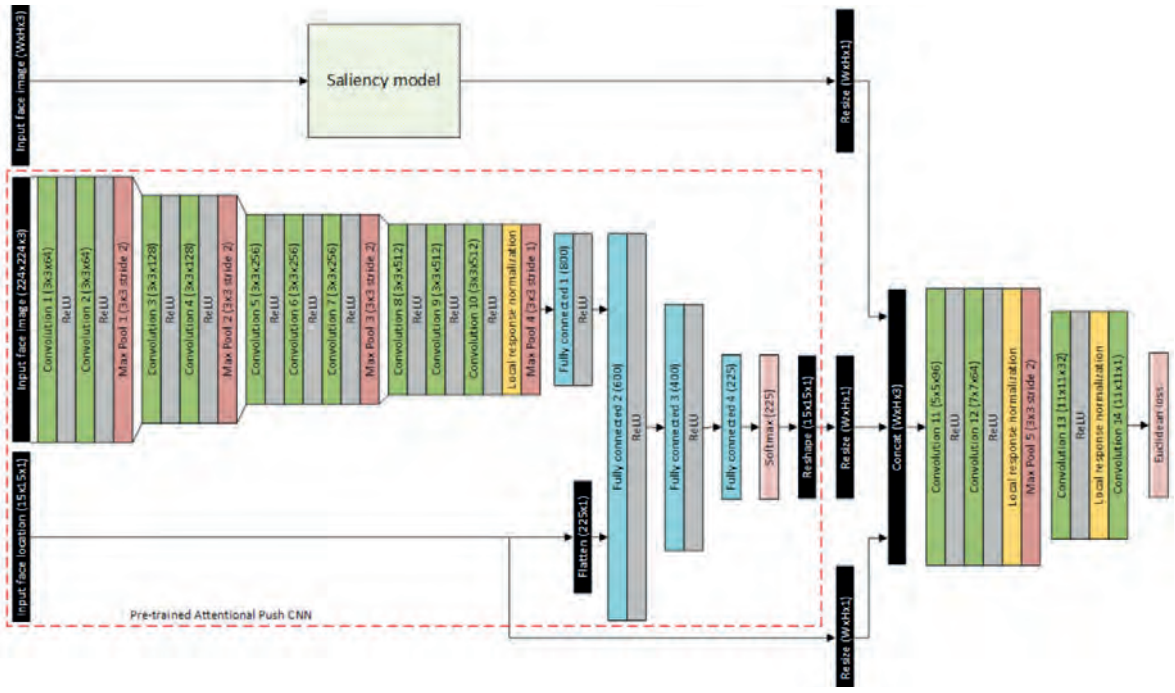
■ ***Fig. 12***. The DeepFix architecture: *a* — starting from the first convolutional block *1*, the number of channels in the outputs of successive blocks gradually increase as 64, 128, 256, 512 that enables the net to progressively learn richer semantic representations of the input image; *b* — additional location biased convolution filter for learning location dependent patterns in data (centre-bias present in the eye-fixations) [71]



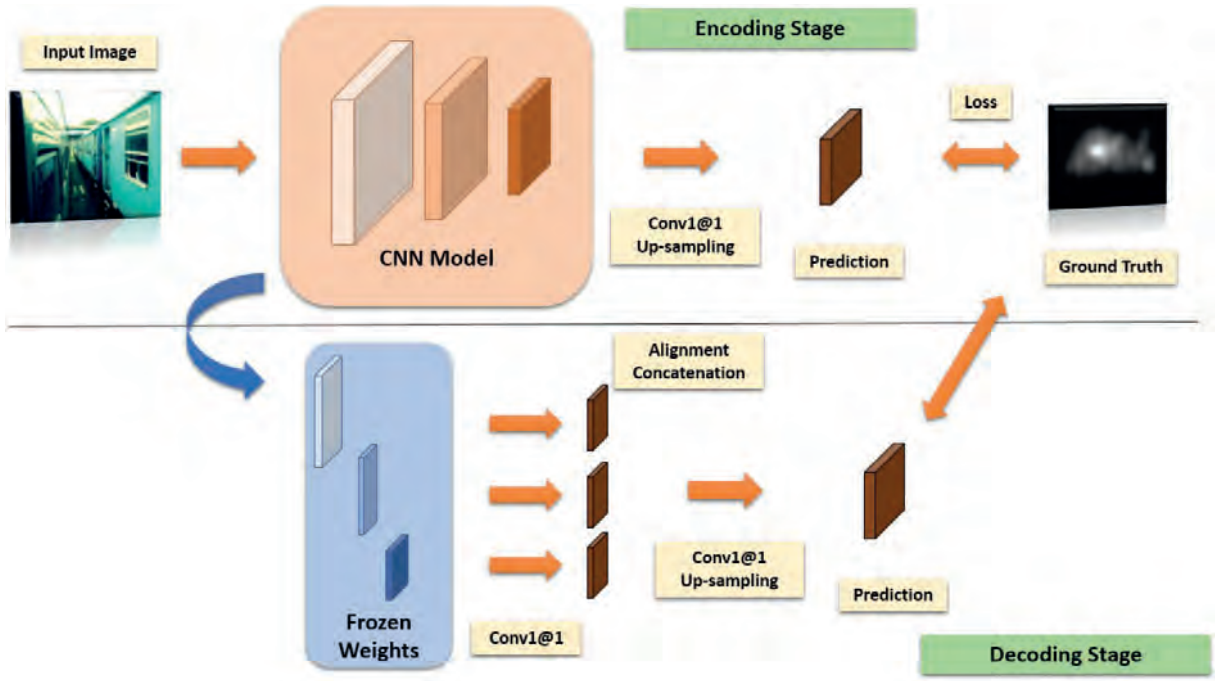
■ ***Fig. 13.*** The SalGAN architecture detects a salient object as real or fake. Generator network produces the predicted saliency map, which with ground truth saliency map is feed into the discriminator network [72]



■ ***Fig. 14.*** The DVA architecture: *a* — attention model learns to combine multi-level saliency information from different layers with various receptive field sizes; *b* — deep visual attention network adopts the encoder-decoder architecture. The supervision is directly fed into hidden layers, encouraging the model to learn robust features and generate multi-scale saliency estimates [74]



■ ***Fig. 15.*** The Attention push architecture (augmented saliency network). Data conditioning layers are depicted in black. The attentional push network is indicated by the red dashed line [75]



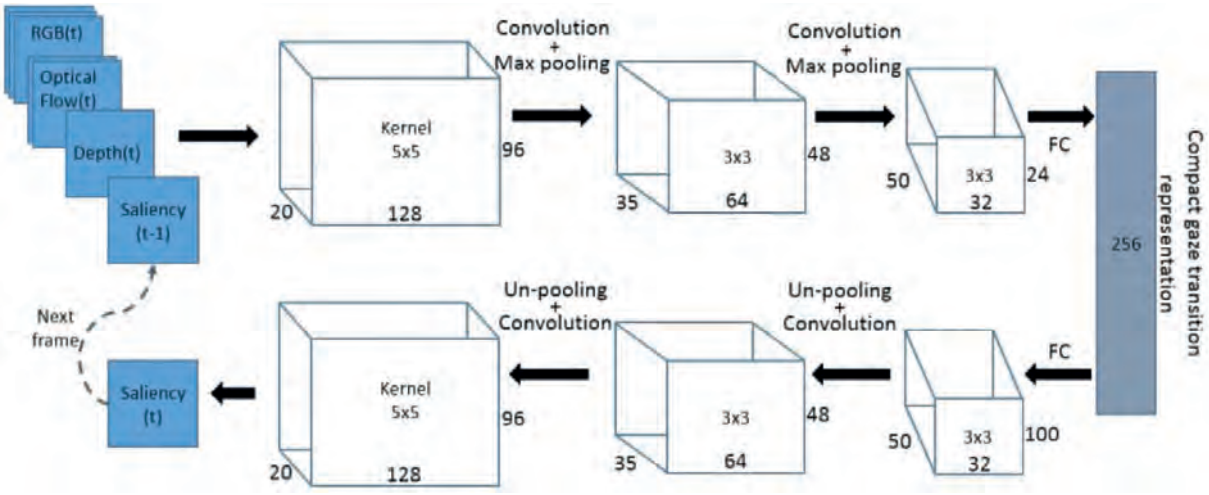
■ ***Fig. 16***. The EML-NET architecture. During training a decoder (in order to combine the multi-level features), the weights of EML-NET model are frozen so that the size of the models can be halved due to no gradients being required [76]

■ ***Table 2***. Dynamic CNN-based saliency models

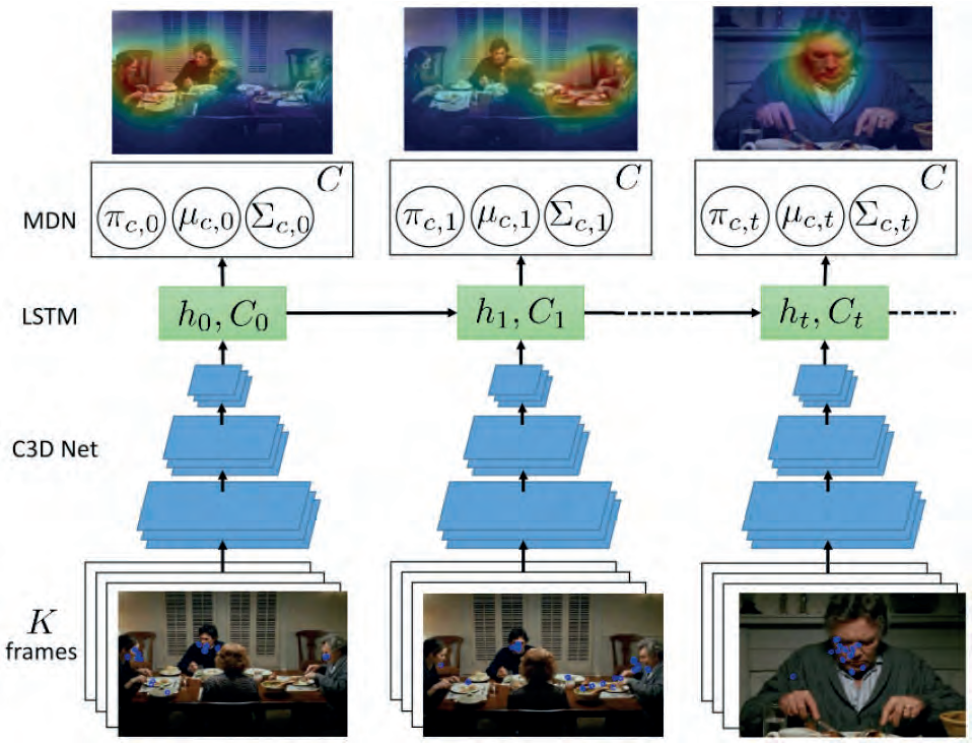
Caption	Description
RGBD generative CNN, 2016	RGBD generative CNN predicts a saliency map for a frame, given the fixation map of the previous frame [77]. Two principals were proposed for saliency detection. First, the gaze slightly varies between frames, and when it does change significantly, it is constrained to a limited number of foci of attention. Second, an actor usually follows the action by shifting their gaze to a new interesting location. Due to these common principals, a sparse candidate set of salient locations are considered and transitions between them over time are predicted (Fig. 17). This means that depth perception has an impact on human attention
RMDN (Recurrent Mixture Density Network), 2016	RMDN for saliency prediction has three levels [78]. The input clip of 16 frames is fed to a 3D CNN, whose output becomes the input to a LSTM. Then a linear layer projects the LSTM representation to a Gaussian mixture model, which describes the saliency map (Fig. 18). Finally, C3D model was connected to the recurrent network in order to perform a temporal aggregation of past clip-level signals. In a similar manner, Liu et al. [79] applied LSTMs to predict video saliency maps, relying on both short- and long-term memory of attention deployment
Deep CNN, 2016	Deep CNN ensures the learning of salient areas in order to predict the saliency maps in videos [80]. First, extraction of salient and non-salient patches in video frames is implemented. Then on the basis of these classifications, a visual fixation map is predicted (Fig. 19)
OM-CNN (Object-to-Motion CNN), 2017	OM-CNN (Fig. 20) predicts saliency of intra-frame, which integrates both objectness and object motion in a uniform deep structure [81]. The objectness and object motion information are used to predict the intraframe saliency of videos. Inter-frame saliency is computed by means of a structure-sensitive [82]

■ *Table 2 (completed)*

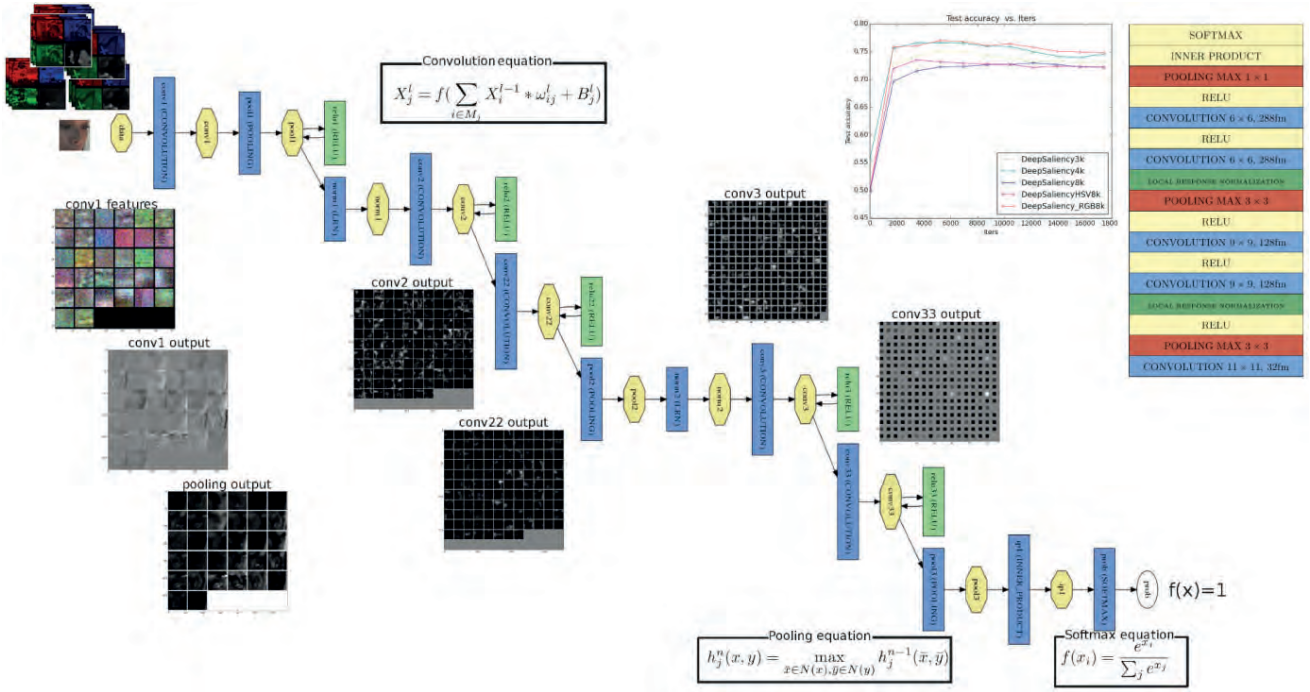
Caption	Description
ConvLSTM (Convolutional Long Short-Term Memory network), 2018	Multi-stream ConvLSTM augments the state-of-the-art static saliency models with dynamic attentional push (shared attention) [82]. This network contains a saliency pathway and three push pathways (Fig. 21). The multi-pathway structure is followed by an augmenting ConvNet by minimizing the relative entropy between the augmented saliency and viewers fixation patterns on videos
The SSNet (Spatial Saliency Network), TSNet (Temporal Saliency Network), STSMaxNet (Spatio-Temporal Max Fusion Network), and STSConvNet (Spatio-Temporal Convolution Fusion Network) architectures, 2018	This is a family of CNNs proposed for predicting saliency from RGB dynamic scenes [83]. SSNet model employs a static saliency model for dynamic saliency prediction by simply ignoring temporal information and using the input video frame alone (Fig. 22). TSNet model is a single stream network contributing a temporal information to the saliency prediction. STSMaxNet model accepts both video frame and the corresponding optical flow image as the inputs and merges together the spatial and temporal single stream networks via an element-wise max fusion. STSConvNet model integrates the spatial and temporal streams by applying a convolutional fusion. The last two models apply two-stream CNN architecture for a video saliency prediction
ACL (Attentive CNN-LSTM), 2018	Attentive CNN-LSTM architecture is based on a video saliency model with a supervised attention mechanism [84]. CNN layers are utilized for extracting the static features within the input frames, while convolutional LSTM is utilized for sequential fixation prediction over successive frames (Fig. 23). An attention module is applied to enhance spatially informative features. The spatial and temporal factors of dynamic attention allow ConvLSTM to learn the temporal saliency representations efficiently
SG-FCN (Spatial Gained Fully Convolutional Network), 2018	SG-FCN is a robust deep model that utilizes the memory and motion information to capture the salient points across the successive frames [85]. The inputs of SGF model are the current frame, the saliency maps in previous frame, and the moving object boundary map, while the output is a spatiotemporal prediction that ensures the time and space consistency (Fig. 24)



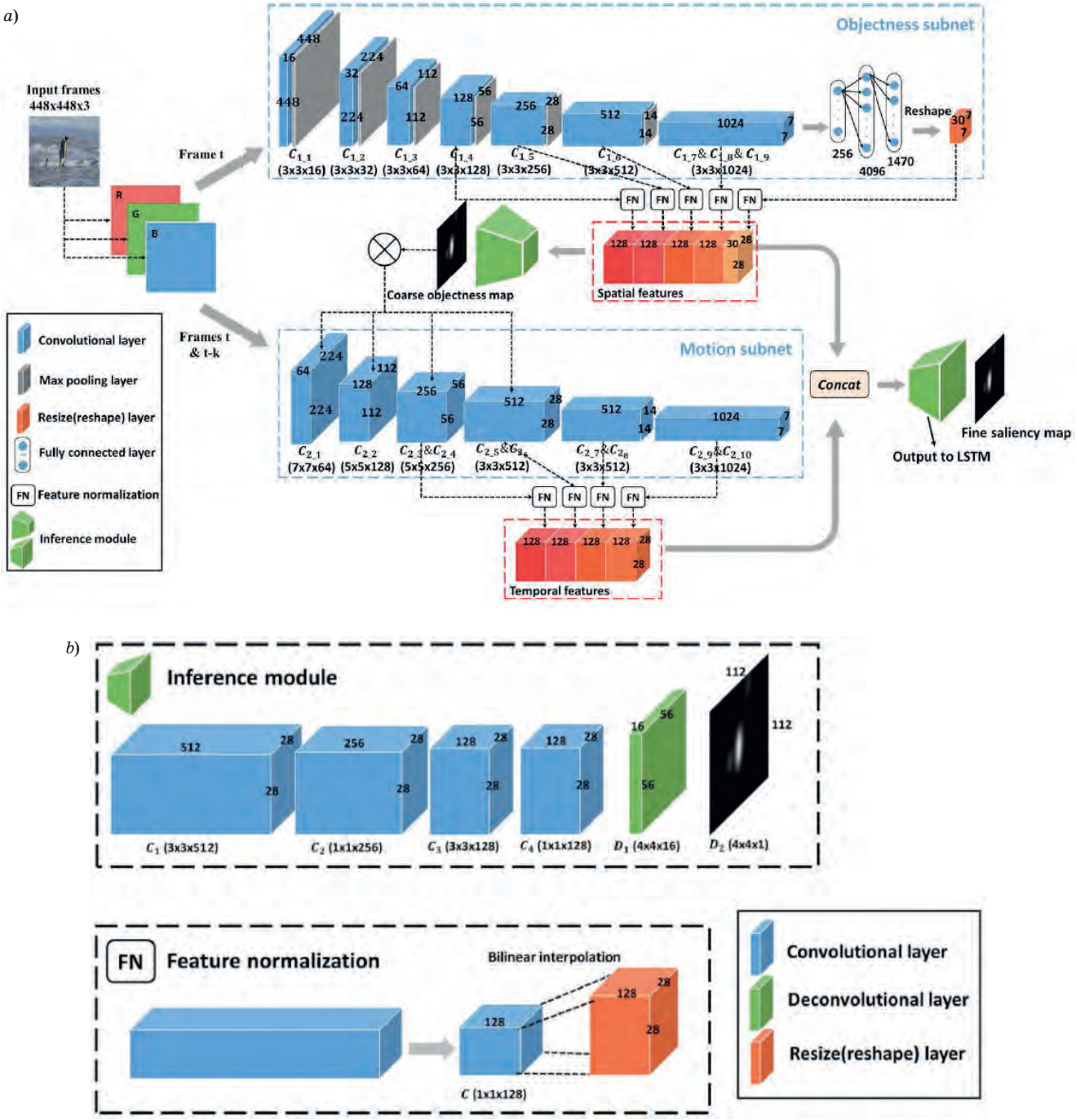
■ *Fig. 17*. The RGBD generated CNN architecture supports a saliency reconstruction using a generative CNN. The input is the saliency calculated for the previous frame and additional information from the current frame. Then the data is encoded, and only the saliency of the current frame is reconstructed [77]



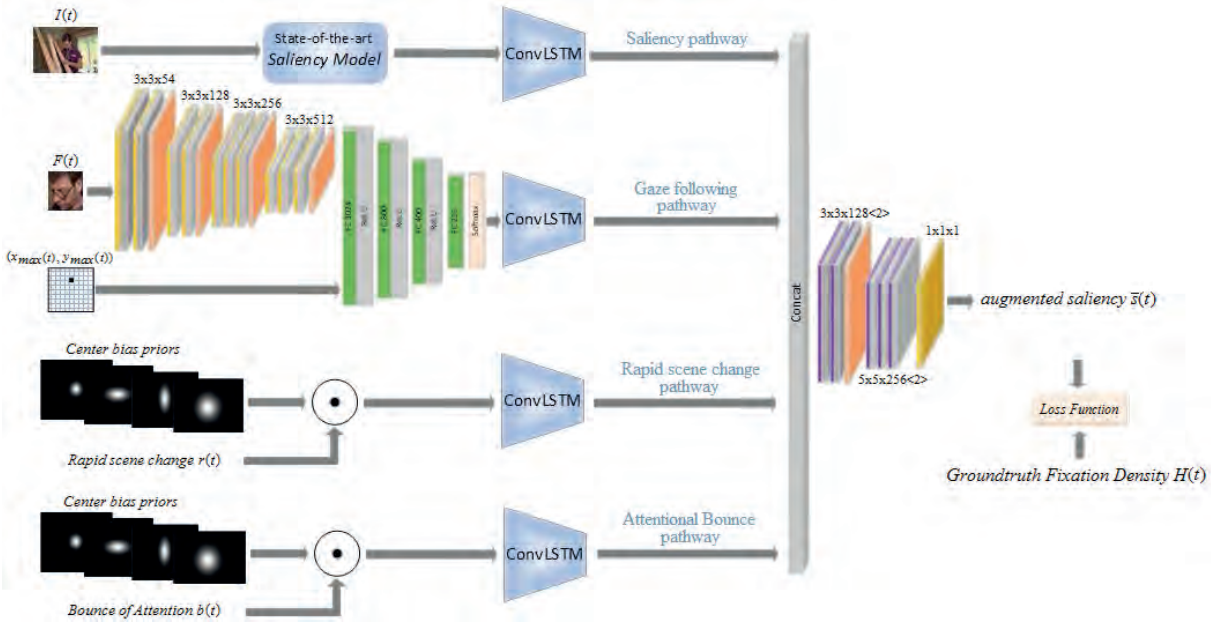
■ ***Fig. 18***. The RMDN architecture. The input clip of K frames is fed into 3D CNN, whose output becomes the input of LSTM network. Finally, a linear layer projects the LSTM representation to the parameters of a Gaussian mixture model, which describes the saliency map [78]



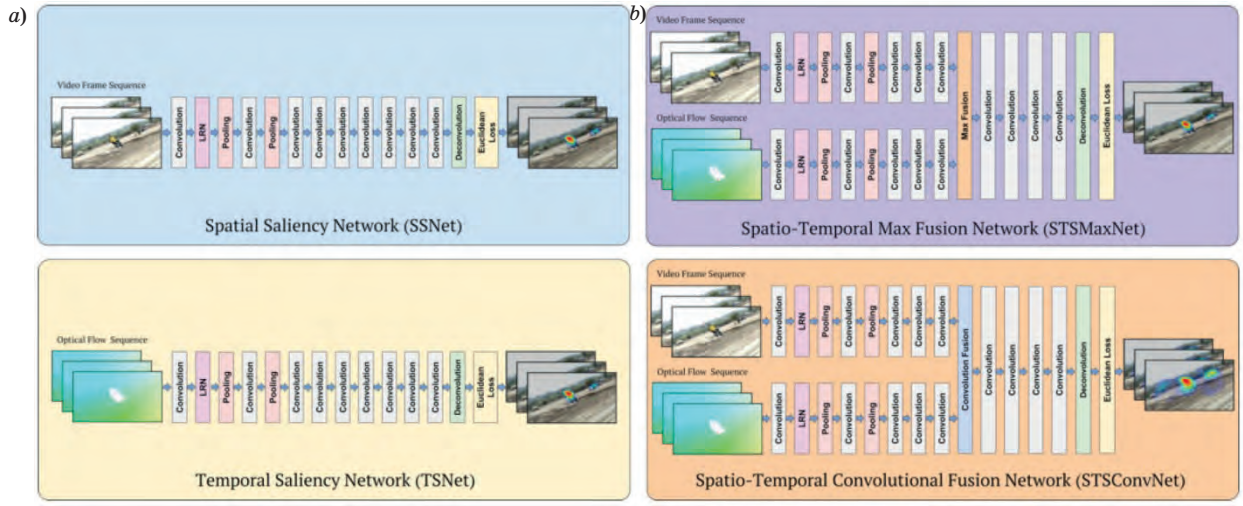
■ ***Fig. 19***. The Deep CNN architecture includes five layers of convolution, three layers of pooling, five layers of rectified linear units, two normalization layers, and one layer of Inner product followed by a loss layer [80]



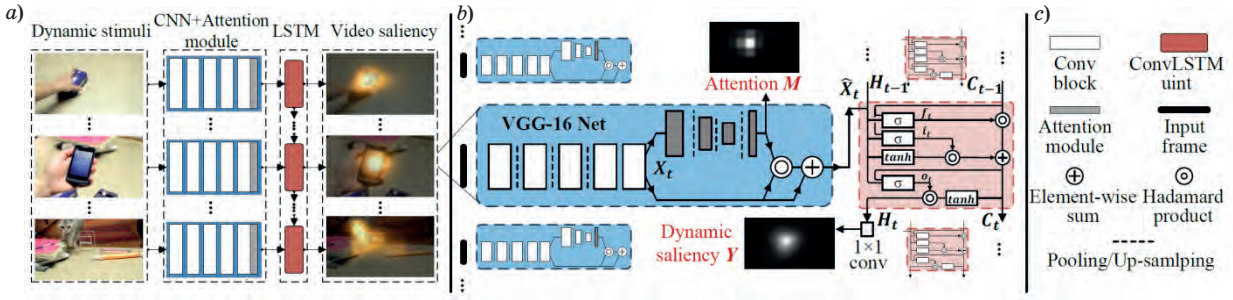
■ ***Fig. 20.*** The OM-CNN architecture for predicting a video saliency of intra-frame: *a* — the overall architecture of OM-CNN; *b* — the details for sub-modules of inference module and feature normalization [81]



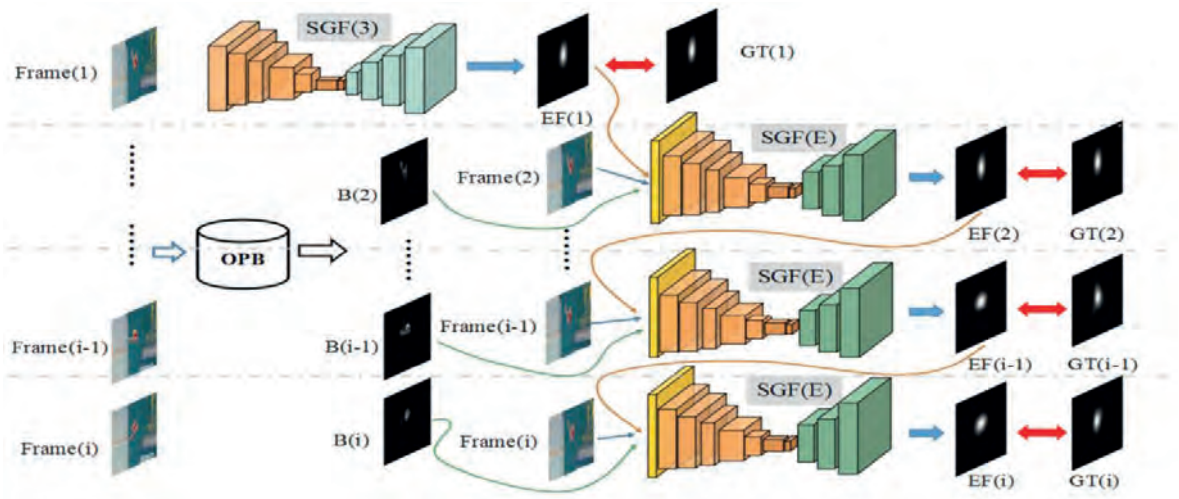
■ ***Fig. 21.*** The Attentional Push architecture includes a Saliency pathway and three Attentional Push pathways, gaze following, rapid scene changes and attentional bounce [82]



■ ***Fig. 22.*** The SSNet, TSNet, STSMaxNet, and STSConvNet architectures. While SSNet utilizes only spatial (appearance) information and accepts still video frames, TSNet exploits only temporal information, whose input is given in the form of optical flow images. STSMaxNet performs fusion by using the element-wise max fusion, whereas STSConvNet employs convolutional fusion after the fifth convolution layers: *a* — single stream saliency networks; *b* — two-stream saliency networks [83]



■ ***Fig. 23.*** The ACLNet architecture: *a* — attentive CNN-LSTM architecture; *b* — CNN layers with attention module are used for learning intra-frame static features, where the attention module is learned with the supervision from static saliency data; *c* — ConvLSTM used for learning sequential saliency representations [84]



■ ***Fig. 24***. The SG-FCN architecture: SGF(3) is used to handle the first frame because neither motion nor temporal information is available. From the next frame onward, the SGF(E) model takes EF(1) from SGF(3), a fast moving object edge map B(2) from the OPB algorithm, and the current frame(2) as the input, and directly outputs the spatiotemporal prediction EF(2) [85]

Thus, the CNN approach for salient object detection in images had been developed intensively since 2014, while a deep learning for salient event detection in videos was activated since 2016. At present, this is a mainstream of investigations in this scope.

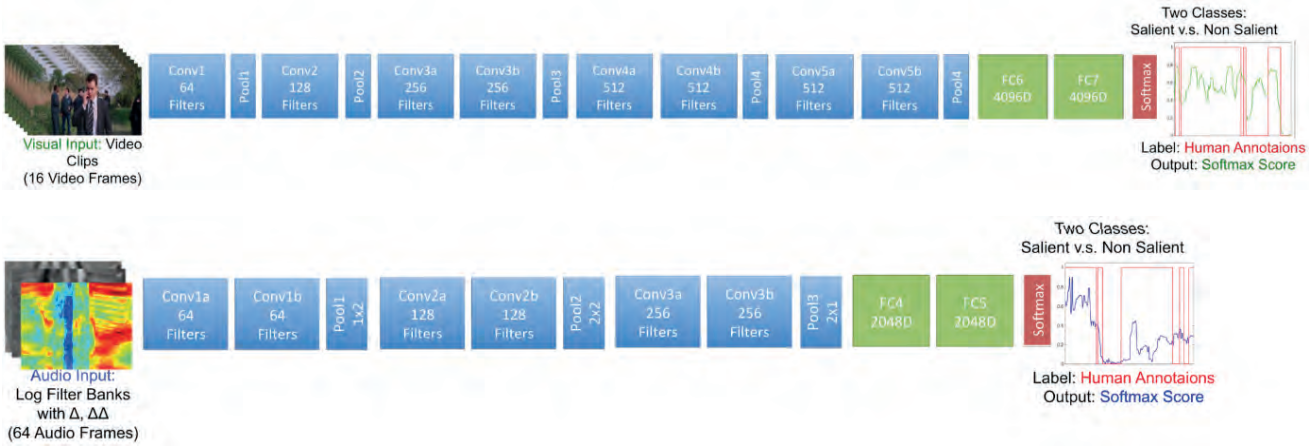
Event-based saliency detection methods

Event detection for the task of video summarization and abstraction appeared after millennium and was oriented on key-frames' extraction, especially in the most descriptive and informative video shorts. The traditional approach is to find the hand-crafted features, which ought to be generic, compact, efficient to compute, and simple to implement [86]. Nowadays, CNN approach prevails in the event-based saliency detection also.

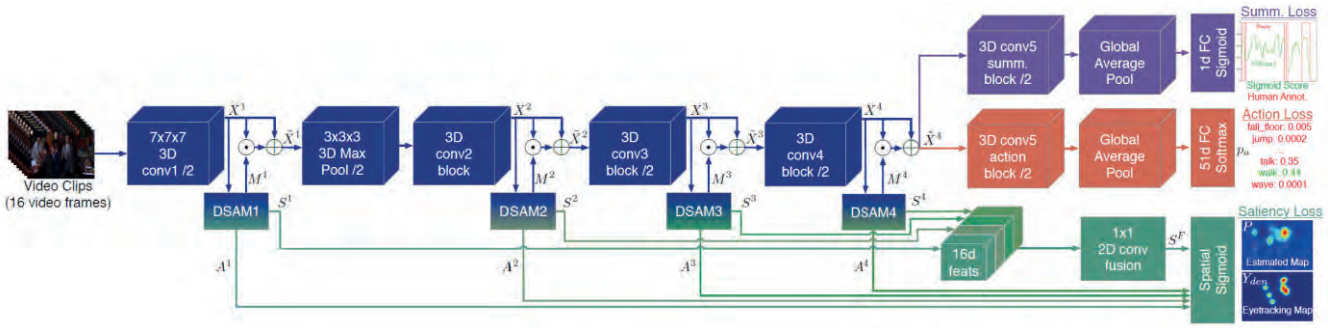
The CNN approaches for video content analysis are classified into two main categories: the learning local spatiotemporal filters (so-called C3D method) and the incorporating optical flow using two-stream CNNs.

In [37], C3D network was employed for video stream, while for the audio stream 2D CNN similar to VGG network was applied for a salient event detection in movies. The architectures of these CNNs are depicted in Fig. 25.

The 3D CNN can model successfully the temporal information due to the convolutions and pooling operations are applied inside spatio-temporal cuboids, while the classic CNNs work only in the spatial domain. The dimension of the feature maps in each convolutional layer of C3D is $n \times t \times h \times w$, where n is the number of filters in each layer; t is the number of video frames; w and h are the width and height of each frame. Videos are split into non-overlapping 16-frame RGB clips, which are used as in-



■ ***Fig. 25***. CNN architectures: *a* — saliency detection using C3D; *b* — audio saliency detection using 2D CNN [37]



■ ***Fig. 26.*** SUSiNet architecture. The multi-task spatio-temporal network is based on the ResNet architecture and has three different branches associated with the different spatio-temporal tasks [87]

put to the networks. The proposed C3D network has eight convolutional layers (with kernels $3 \times 3 \times 3$ and the stride of all these kernels are 1 in both spatial and temporal domain), five max-pooling layers (with kernels $2 \times 2 \times 2$ except for the first one), and two fully connected layers, followed by a softmax output layer.

For the audio stream, 2D CNN was employed for acoustic event detection. The raw audio signal was represented in 2D time-frequency domain and preserve locality in both axes. Note that conventional mel-frequency cepstral coefficients cannot maintain locality to the frequency axis due to the discrete cosine transform projection.

Another example is a multi-task spatio-temporal network called SUSiNet (See, Understand and Summarize it Network) that can execute the saliency estimation, visual concept understanding, and video summarization [87]. The SUSiNet, which architecture is depicted in Fig. 26, is a single network that is jointly end-to-end trained for all three mentioned above tasks.

Implementation of SUSiNet is very similar to 3D ResNet-50 architecture [88], which has showed competitive performance and computational budget for the task of action recognition. As starting point, the weights from the pretrained model in the Kinetics 400 database are used.

The input samples in the network consist of 16-frames RGB video clips spatially resized at 112×112 pixels. Also data augmentation for random generation of training samples is utilized. For saliency estimation, spatial transformations to the 16 frames of the video clip had been done. The eye-tracking based saliency maps is extracted from the median frame, which has been considered as the ground truth map of the whole clip.

Saliency datasets

Validation of saliency algorithms has been done using the public datasets. Visual material from these datasets is marked by different ways. Traditional saliency datasets are annotated using information about eye movements of humans watching the images or videos. Recent datasets follow two trends: increasing visual material and introducing new saliency measures based on contextual annotations (e.g. image categories). One of the last trends for large scale data annotation is the application of crowdsourcing schemes, such as gaze tracking using webcams [89] or mouse movements [90, 91] instead of the lab-based eye trackers.

Let us consider the recent image and video datasets.

Image datasets

The work for creation of salient object detection in the images was initialized since 2012 and nowadays continues to evolve new versions of public datasets.

MIT Saliency Benchmark dataset (MIT300) includes 300 natural images [92, 93]. This dataset was the first dataset with held-out human eye movements using eyetracker ETL 400 ISCAN (240 Hz). In MIT300 dataset, the eye fixations of 39 observers are available per an image, more than in other datasets of similar size. Eye movements were collected under different conditions, such as free viewing, visual search, and so on. The robustness of the data depends on the eye tracking setup (participant distance to the eye tracker, calibration error, and image size) and number of eye fixations collected.

CAT2000 contains two sets of images: train and test images [94]. The train images (100 from each category) and fixations of 18 observers are shared but six observers are held-out. Test images are available but fixations of all 24 observers are held out. The eyetracker: EyeLink1000 (1000 Hz) was employed for CAT2000 dataset collection.

SALICON is the largest crowd-sourced saliency dataset [95]. The images were imported from Microsoft COCO dataset and contain MS COCO's pixelwise semantic annotations. The SALICON con-

tains 10 000 training images, 5000 validation images, and 5000 test images (Fig. 27). The mouse-contingent saliency was stored using Amazon mechanical turk. Minor discrepancies between the eye movements and mouse movements led to that SALICON dataset is used for rough training, and then the deep saliency models are fine tuned on MIT1000 or CAT2000 datasets, which contain information of predicting fixations.

MSRA10K is formally named as THUS10000 [96]. It contains 195 MB of images and binary masks. Pixel accurate salient object labeling was implemented for 10 000 images from MSRA dataset. **MSRA-B** dataset involves 5000 images from hundreds of different categories. Because of its diversity and large quantity, MSRA-B has been one of the most widely used datasets in salient object detection literature. Most images in this dataset have only one salient object, and, hence, this dataset becomes a standard dataset for evaluating the capability of processing simple scenes. The ground truth of MSRA-B is represented in a form of the labeled rectangles, which were drawn by nine participants. Thus, the objects are segmented into rectangles in order to obtain the binary masks as the pixelwise annotations.

ECSSD is an extension of Complex Scene Saliency Dataset (CSSD) [97]. The matter is that the

■ *Fig. 27.* Samples of salient detection images from SALICON dataset [95]

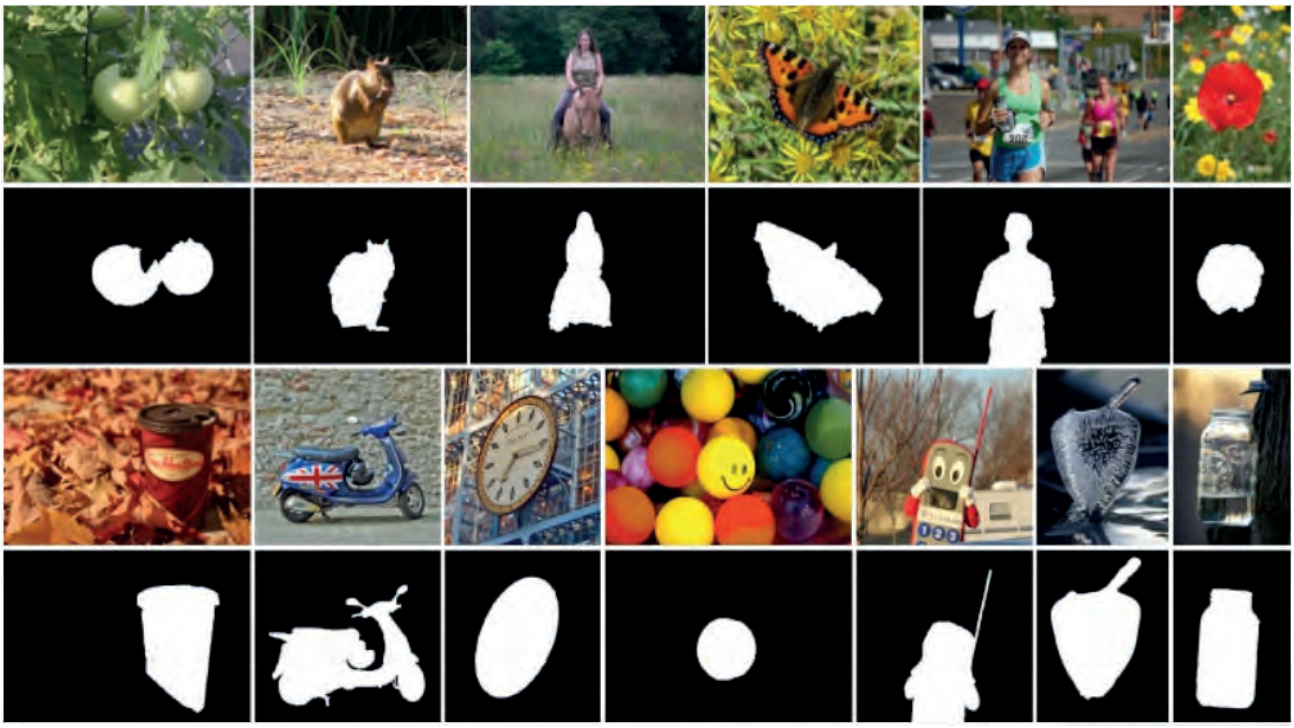
images from MSRA-B dataset have a primarily simple and smooth background. In the contradiction, ECSSD dataset was created using structurally complex images with their ground truth binary masks, which were made by five participants. ECSSD dataset contains 1000 semantically meaningful but structurally complex natural images. Samples are depicted in Fig. 28.

HKU-IS (University of Hong Kong) is a large-scale dataset that contains more than 4400 challenging images with the salient objects annotated as binary masks, where 50.34% images have the multiple salient objects, and 21% have the salient regions touching the boundary. Most of images in this dataset have low contrast with more than one salient object [98]. In order to remedy the weakness of dataset images containing one salient object and 98% of the pixels in the border belonging to the background, the HKU-IS dataset provides a more challenging dataset. The HKU-IS dataset is divided into three parts: 2500 images for training, 500 images for validation and the remaining 1447 images for testing.

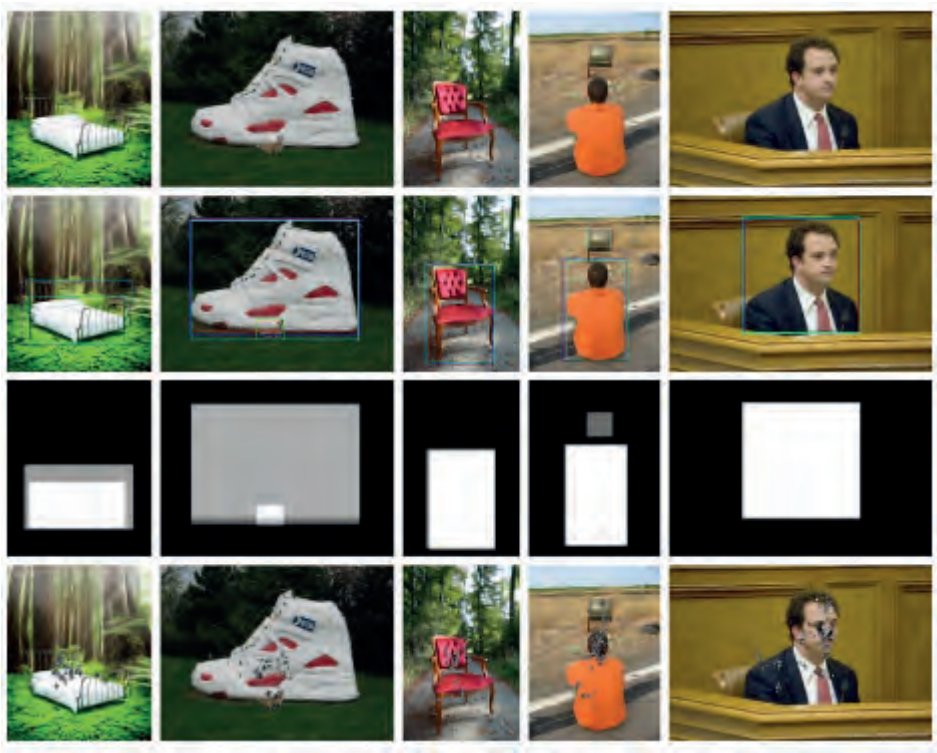
PASCAL has a goal to recognize objects from a number of visual object classes in realistic scenes [99]. The 20 object classes images categorized into "Person" (person), "Animal" (bird, cat, cow, dog, horse, sheep), "Vehicle" (aeroplane, bicycle, boat, bus, car, motorbike, train), and "Indoor" (bottle, chair, dining table, potted plant, sofa, tv/monitor). The main purposes are the classification, detection, and segmentation with additional tasks, such as the person layout, action classification, and ImageNet large scale recognition. The train/validation data has 10 103 images containing 23 374 annotated objects (regions of interests) and 4203 segmentations. The saliency detection function does not support directly. However, some images can be chosen as a saliency detection subset [100].

SOD (Salient Object Detection) is a collection of salient object boundaries based on Berkeley Segmentation Dataset (BSD) [101]. It contains 300 images, most of which possess the multiple salient objects. All of these datasets consist of the ground truth human annotations. Seven objects are asked to choose the salient object(s) in each image used in BSD. Each subject is shown randomly as a subset of the Berkeley segmentation dataset with boundaries overlapped on the corresponding images. Participant can then choose which regions or segments correspond to salient objects by clicking on them.

DUT-OMRON dataset includes the nature images for the research of more applicable and robust methods in both salient object detection and eye fixation prediction [102]. The DUT-OMRON dataset consists of 5168 high quality images manually selected from more than 140,000 images. The images of DUT-OMRON database have one or more salient objects and a relatively complex background. The pixel-wise ground truth, bounding box ground truth, and eye-fixation ground truth in large scaled images were constructed (Fig. 29).



■ *Fig. 28*. Samples of original and binary ground-truth masked images from ECSSD [97]



■ *Fig. 29*. Samples from DUT-OMRON dataset. From top to bottom: original image; bounding box ground truth, pixel-wise ground truth; average of the five binary masks; and eye-fixation ground truth [102]

Saliency video datasets are considered in the following Section.

Video datasets

A spectrum of saliency video datasets is consistency extended. Often saliency video datasets are built using public video datasets that were constructed for other purposes.

DAVIS dataset is a special dataset containing a ground-truth of human attention in RGBD video sequences [103]. The videos from DAVIS dataset represent the scenarios, where a depth-aware saliency is beneficial [77]. The RGBD videos were acquired by built in the phone/tablet/laptop depth/stereo cameras or 3D sensors, such as Kinect or LiDAR. Video sequences contain the static and dynamic indoors and outdoors scenes, such as video conference, surveillance, tracking, and obstacle avoidance. Nearly 54 videos with varying durations ranging from 25 to 200 s were chosen from public datasets. The videos were converted to a 30 frame-rate, resulting in approximately 100K frames across all videos. Gazepoint GP3 Eye Tracker with the Gazepoint Analysis Standard software was applied for the eye movements' monitoring of 91 participants.

LEDOV (Large-scale Eye-tracking Database of Videos) dataset involves 538 videos, in total 179 336 frames and 6431 s, equally divided into six non-overlapping groups with similar numbers of videos in content (i. e., human, animal and man-made object) [104]. Videos were collected according to the following four criteria [81]: the diverse video content (daily blogs, documentaries, movies, sport casts, TV shows, etc.) including at least one object, high quality video (high quality of videos with at least 720 p resolution and 24 Hz frame rate), and the stable shots (212 videos were obtained with stable camera motion and 316 videos were received without any camera motion). For monitoring the binocular eye movements, an eye tracker Tobii TX300 was used in carefully conducted experiments.

DIEM (visualizing Dynamic Images and Eye Movements with a tool called Computational Algorithms for Representation and Processing of Eye-movements (CARPE)) dataset contains 85 high-definition natural videos including movie trailers, advertisements, and so on. Each video sequence has the eye fixation data collected from approximately 50 different human subjects. [105]. The DIEM project is an investigation of how people look and see. The applied CARPE technique allows one to begin visualizing eye-movement data in a number of ways. The project includes a number of different visualization options: the low level visual features that process the input video to show flicker or edges, the heat-maps that show where people are looking, the clustered heat-maps that use pattern recognition to define the best model of fixations for

each frame, and the peek-through, which uses the heat-map information to only show parts of the video where people are looking.

UCFSports dataset is collected from broadcast television channels, such as the BBC and ESPN, and a wide range of websites, which consists of a set of sport actions [106]. The UCFSports dataset contains 150 video sequences with 720 × 480 resolution and cover a range of scene and viewpoints. The dataset includes 10 actions, such as diving (14 videos), golf swing (18 videos), kicking (20 videos), lifting (6 videos), riding horse (12 videos), running (13 videos), skateboarding (12 videos), swing-bench (20 videos), swing-side (13 videos), and walking (22 videos), for recognition purpose. Recently, additional human gaze annotations were collected in [107]. These fixations were collected over 16 human subjects under the task specific and task independent free viewing conditions.

COGNIMUSE is a recent multi-modal video database annotated with the saliency, events, semantics, and emotion with application to summarization [108]. The COGNIMUSE database includes data collection, data conversion, and annotation in different phases [109]. The dataset consists of half-hour continuous segments (with the final shot/scene included) from seven Hollywood movies (three and a half hours in total), five travel documentaries (20 min long each), and a full-length annotated movie, namely "Gone with the Wind" (the first part with a total duration 104 min). All database videos have been annotated with the sensory and semantic saliency, audio-visual events and emotion. The structure of COGNIMUSE project is depicted in Fig. 30.

First, the movie clips are manually segmented (cut or fade), and scenes defined as a complete, continuous chain of actions (shots). The average shot and scene duration for the movies are 3.5 s–2.3 min, while for the travel documentaries, the respective duration is 3–40 s. Second, the sensory and semantic saliency content annotation (segments that captured the viewer's attention with respect to the following layers) is performed.

Let us hope that the saliency image and video datasets will be developed in future providing more complex and diverse wildlife visual content like the COGNIMUSE database.

Evaluation metrics

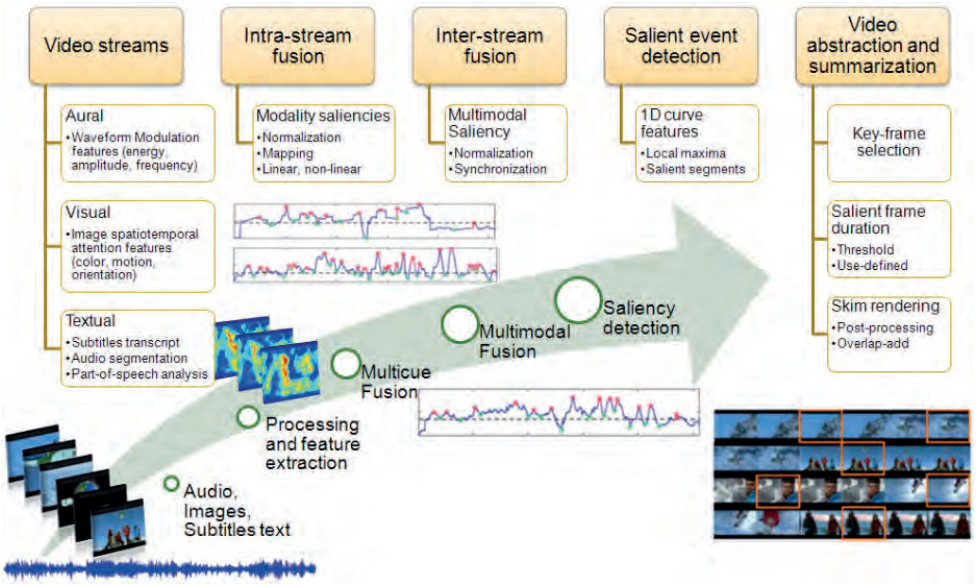
Saliency models are usually evaluated by comparing their predicting maps to the human fixation maps. Generally, the evaluation metrics fall into two categories: location-based (computing some statistics at fixated locations) and distribution-based (comparing smoothed prediction and fixation maps).

Location-based metrics

Receiver Operating Characteristic ROC. The ROC is a binary classification measure of the intersected area between the predicted saliency and human fixations. At various thresholds, the tradeoff between True Positive Rates (TPR) and False Positive Rates (FPR) is plotted:

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}, \tag{1}$$

where TP and FP are the truth positive and truth negative, respectively; TN and FN are the truth negative and false negative, respectively.



■ *Fig. 30.* Structure of COGNIMUSE project [108]

The ROCs are computed by two ways. The first way is to measure the intersection between a saliency map and a ground-truth distribution of human fixation. The second way uses a uniform random sample of image pixels as negatives and the saliency map values are defined above threshold at these pixels as false positives.

Area Under ROC Curve (AUC). The AUC is an integration of the spatial area under the ROC curve such that the random guessing score is 0.5. A score above 0.5 indicates that predictions are above random guessing. The AUC curves have modifications, such as AUC-Judd, AUC-Borji, and Shuffled AUC (sAUC) [110].

Precision-recall and F-measure. The estimates Precision and Recall are defined according to equation

$$\text{Presicion} = \frac{TP}{TP+FP}; \qquad \text{Recall} = \frac{TP}{TP+FN}. \quad (2)$$

F-measure determines a successfulness of salient object detection respect to the chosen binary threshold provided by equation

$$F(T) = \frac{\left(1+\beta^2\right)\text{Precision}(T)\times\text{Recall}(T)}{\beta^2\times\text{Precision}(T)+\text{Recall}(T)}, \quad (3)$$

where β is an empirical coefficient.

Normalized Scanpath Saliency (NSS). The NSS is a measure of the normalized saliency at fixations. Unlike in the AUC, the absolute saliency values are part of the normalization calculation. Thus, NSS is sensitive to false positives, relative differences in saliency across the image, and general monotonic transformations. However, due to the mean saliency value subtraction during a computation, NSS is invariant to the linear transformations like the contrast offsets (given a saliency map P and a binary map of fixation locations Q^B):

$$NSS\left(P,Q^B\right) = \frac{1}{N}\sum_i \bar{P}\times Q_i^B, \quad (4)$$

where

$$N = \sum_i Q_i^B \quad \text{and} \quad \bar{P} = \frac{P-\mu(P)}{\sigma(P)},$$

where N is the total number of fixated pixels.

Positive NSS indicates correspondence between the maps above chance (chance is at 0) and negative NSS indicates the anti-correspondence. For instance, a unity score corresponds to fixations falling on portions of the saliency map with a saliency value one standard deviation above average.

Information Gain (IG). The IG was proposed in [111] as an information-theoretic metric that measures saliency model performance beyond systematic bias (e. g., a center prior baseline). Given a binary map of fixations Q^B, a saliency map P, and a baseline map B, information gain is computed as

$$IG\left(P,Q^B\right) = \frac{1}{N}\sum_i Q_i^B\left(\log_2\left(\varepsilon+P_i\right)-\log_2\left(\varepsilon+B_i\right)\right), \quad (5)$$

where ε is the regularization parameter.

The IG is measured in bits per fixation. This metric measures the average information gain of the saliency map over the center prior baseline at fixated locations (i. e., where $Q^B = 1$). The IG assumes that the input saliency maps are probabilistic, properly regularized, and optimized to include a center prior. As it was mentioned in [111], a score above zero indicates that the saliency map predicts the fixated locations better than the center prior baseline.

Distributed-based metrics

Similarity (or histogram) Intersection Metric (SIM). The SIM measures the similarity between two distributions, viewed as histograms. The earliest version of SIM was interpreted as a metric for color-based and content-based image matching. For saliency task, the SIM is introduced as a simple comparison between pairs of saliency maps. The SIM is computed as the sum of the minimum values at each pixel, after normalizing the input maps (given a saliency map P and a continuous fixation map Q^D):

$$SIM\left(P,Q^D\right) = \sum_i \min\left(P_i,Q_i^D\right), \quad (6)$$

where

$$\sum_i P_i = \sum_i Q_i^D = 1.$$

The SIM of 1 indicates the distributions are the same, while the SIM of 0 indicates no overlap. Note that the model with the sparser saliency map has a lower histogram intersection with the ground truth map. Also, the SIM is sensitive to the missing values and penalizes predictions that fail to account of all ground truth data. The SIM is good for evaluating the partial matches, when a subset of the saliency map reflects well the ground truth fixation map. As a side-effect, false positives tend to be penalized lesser than false negatives.

Pearson's Correlation Coefficient (CC). The CC, also called linear correlation coefficient, specifies the statistical relationship between the predicted saliency map and human ground-truth. The saliency map and human ground-truth are treated as random variables, and the strength and direction between the two variables are measured by CC estimate:

$$CC\left(P,Q^D\right)=\frac{\operatorname{cov}\left(P,Q^D\right)}{\sigma(P)\times\sigma\left(Q^D\right)}, \quad (7)$$

where cov(*S*, *F*) denotes the covariance between the saliency map *P* and fixation map Q^D. High positive CC values occur at locations, where both the saliency map and ground truth fixation map have values of similar magnitudes. A score of zero indicates that two maps are not correlated.

For visualizing CC each pixel *i* has value

$$V_i=\frac{P_i\times Q_i^D}{\sqrt{\sum_j\left(P_j^2+\left(Q_j^D\right)^2\right)}}.$$

Due to its symmetric computation, the CC cannot distinguish whether differences between the maps are due to false positives or false negatives.

Kullback — Leibler (KL) distance. The KL is a general information-theoretic measure of the difference between two probability distributions. In saliency detection, the KL show how the saliency predictions and ground truth fixations are interpreted as distributions. The KL metric takes as input a saliency map *P* and a ground truth fixation map Q^D, and evaluates the loss of information, when *P* is used to approximate Q^D:

$$KL\left(P,Q^D\right)=\sum_i Q_i^D\log\left(\varepsilon+\frac{Q_i^D}{\varepsilon+P}\right), \quad (8)$$

where ε is a regularization constant. A score of 0 indicates that two maps are identical. A positive score indicates the divergence between two maps.

Earth Mover's Distance (EMD). The EMD incorporates spatial distance into evaluation. It was introduced as a spatially robust metric for image matching. The linear time variant of EMD has a view [112]

$$\widehat{EMD}\left(P,Q^D\right)=$$

$$=\min_{\{f_{ij}\}}\sum_{i,j}f_{ij}d_{ij}+\left|\sum_i P_i-\sum_j Q_j^D\right|\times\max_{i,j}d_{ij}, \quad (9)$$

under the constrains

$$f_{ij}\ge 0\ \sum_j f_{ij}\le P_i\ \ \sum_j f_{ij}\le Q_j^D\ \ \sum_j f_{ij}=\min\left(\sum_i P_i,\sum_j Q_j^D\right),$$

where each f_{ij} represents the amount of density transported (or the flow) from the *i*th supply to the *j*th demand and d_{ij} is the ground distance between bin *i* and bin *j* in the distribution.

A larger EMD indicates a larger difference between two distributions, while an EMD of 0 indicates that two distributions are the same. Generally, the saliency maps that spread density over a larger area have larger EMD values (worse scores). The EMD penalizes false positives proportionally to the spatial distance they are from the ground truth.

Matching Score (MSc). Due to saliency does not suppose the classification, the matching scores define how relevant the feature map is to the salient object. The matching score *MS* is defined as the sum of the absolute differences between the saliency map *P* and ground truth *F*, expressed as

$$MSc=\sum_i\sum_j\left|P_{i,j}-F_{i,j}\right|, \quad (10)$$

where *i* and *j* denote the row and column matrix indexes, respectively.

The estimates for 3D salient object detection are proposed in [113]. Early saliency models computed a multi-scale representation of a mesh and observed a local vertex property (curvature, surface variation, or normal displacement changes at different scales). The following saliency models achieved robustness and speed by segmenting a mesh into the patches represented by descriptors using a ranking process that specifies patch distinctiveness. Recent saliency models focus on the point sets. Let us consider briefly the saliency metrics focusing on recent saliency models.

Saliency of large point sets (LS). The LS metric was the first one that supports saliency detection on large points sets [114]. Saliency is considered as a combination of point distinctiveness at two scales with point association. The LS assigns higher saliency to regions near foci of attention. Distinctiveness is computed by comparing local neighbourhoods described by the Fast Point Feature Histograms (FPFH) [115], which consists of 33D histograms of angles between oriented points in a local region.

Mesh saliency via spectral processing (MS). The MS metric proposed a spectral-based approach [116]. The MS is more robust metric that analyzes the changes in local vertex properties. The *n* lowest frequencies of log-Laplacian spectrum *L* are applied. The log-Laplacian spectrum amplifies the low-frequency variation of the Laplacian spectrum and detects the most "fundamental" saliencies.

Cluster-based point set saliency (CS). The CS allows to detect a fine-scale saliency with better time complexity [117]. The point sets are segmented into *K* clusters, and a cluster saliency is computed as a sum of cluster distinctiveness and spatial distribution. Cluster distinctiveness is based on the mean FPFH of points belonging to that cluster. The CS metric uses a method similar to [114].

PCA-based saliency (PS). The PS value is computed as the absolute value of the FPFH descriptors

projected onto the largest principal axis after the mean centering.

In [109], guidelines for designing the saliency benchmarks are offered. For example, the KL-divergence and IG metrics are suggested for evaluating probabilistic saliency models. If the saliency models are not probabilistic but capture behavior including the systematic biases, then NSS or Pearson's CC are recommended.

Summery and conclusions

The CNN-based models are trained in a single end-to-end manner, combining feature extraction, feature integration, and saliency value prediction that led to a large gap in performance relative to traditional saliency models. The CNN capability to extract the high-level image features and capture the global context is extremely useful to predict fixation locations and, as a result, saliency detection.

The accuracy and speed of CNN depends on many factors, among which are the following: CNN parameters and setting, errors of models, transfer learning, hardware platform, and routine. Deep learning models have shown the impressive performance in saliency detection. However, they continue to miss the key elements in the images and videos. Partially, this effect is caused by a human annotation of public datasets, when several participants mark close but different salient regions in visual material. One of the ways to avoid errors is to train CNNs on different tasks, to learn to detect gaze and action. Another way is additional information about important regions in image, for instance person in indoor/outdoor environment, animal in the wild, the most informative traffic sign on the road, and forth.

It is well-known that CNN extracts million of features. However, the question, which features are the best for saliency prediction, is unsolved. It is considered that one of the first CNNs, ImageNet involving five layers [118] provides extraction of corners, edges, and colors at layer 2, texture information at level 3, and class-specific features at levels 4 and 5. Some researchers try to analyze information in each layer in order to understand a local effectiveness. In this sense, CNNs transform from "black box" structure to more predicted system.

Combination of several CNNs is becoming a conventional approach for decision related to complex task. Thus, in [30], a combination of 13-layered VGG network [119] pre-trained on the ImageNet dataset and 5-layered convolutional network based on selection (SCnet) was applied for salient object detection. The issues of accuracy and speed estimates for such combined CNNs require future investigations.

However, new deep saliency models still suffer from several shortcomings before they can reach a level of human accuracy. Failure analysis allows to design better optimization models, CNN architecture, datasets, and training and evaluation procedures. To close the gap between the human input/output model and saliency models, it is necessary to understand how attention is deployed in humans. We strongly believe that only common efforts and multidiscipline cooperation will lead to better results in saliency detection and prediction.

References

1. Itti L., Koch C., Niebur E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998, vol. 20, no. 11, pp. 1254–1259.
2. Cooley C., Coleman S., Gardiner B., Scotney B. Saliency detection and object classification. *Proc. 19th Irish Machine Vision and Image Processing Conf. (IMVIP 2017)*, 2017, pp. 84–90.
3. Bi S., Li G., Yu Y. Person re-identification using multiple experts with random subspaces. *Int. J. Image Graph.*, 2014, vol. 2, no. 2, pp. 151–157.
4. Avidan S., Shamir A. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 2007, vol. 26, no. 3, pp. 1–10.
5. Aswathy S., Unnikrishnan A. S., Santhosh B. S. An integrated approach for image inpainting based on saliency detection. *Int. J. Innovative Research in Science, Engineering and Technology*, 2017, vol. 6, no. 6, pp. 104–114.
6. Marchesotti L., Cifarelli C., Csurka G. A framework for visual saliency detection with applications to image thumbnailing. *Proc. IEEE 12th Int. Conf. Computer Vision*, 2009, pp. 2232–2239.
7. Wang P., Wang J., Zeng G., Feng J., Zha H., Li S. Salient object detection for searched web images via global saliency. *Proc. 2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3194–3201.
8. Zhu W., Liang S., Wei Y., Sun J. Saliency optimization from robust background detection. *Proc. 27th IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2014)*, 2014, pp. 2814–2821.
9. Ma Y. F., Lu L., Zhang H. J., Li M. A user attention model for video summarization. *Proc. 10th ACM Int. Conf. Multim.*, 2002, pp. 533–542.
10. Zhang J., Malmberg F., Sclaroff S. *Visual saliency: from pixel-level to object-level analysis*. Springer International Publishing, 2019. 138 p.
11. Zhang, Q., Lin, J., Tao, Y., Li, W., Shi Y. Salient object detection via color and texture cues. *Neurocomputing*, 2017, vol. 243, pp. 35–48.
12. Yang, B., Zhang, X., Chen, L., Yang, H., Gao Z. Edge guided salient object detection. *Neurocomputing*, 2017, vol. 221, pp. 60–71.

13. Hu Y., Chen Z., Chi Z., Fu H. Learning to detect saliency with deep structure. *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2015, pp. 1770–1775.
14. Alexe B., Deselaers T., Ferrari V. Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, vol. 34, no. 11, pp. 2189–2202.
15. Dalal N., Triggs B. Histograms of oriented gradients for human detection. *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2005, vol. 1, pp. 886–893.
16. Cheng M., Mitra N. J., Huang X., Torr P. H., Hu S. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, vol. 37, no. 3, pp. 569–582.
17. Felzenszwalb P. F., Huttenlocher D. P. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 2004, vol. 59, no. 2, pp. 167–181.
18. Perazzi F., Krahenbuhl P., Pritch Y., Hornung A. Saliency filters: Contrast based filtering for salient region detection. *IEEE 2012 Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 733–740.
19. Treisman A. M., Gelade G. A feature-integration theory of attention. *Cognitive Psychology*, 1980, vol. 12, pp. 97–136.
20. Koch C., Ullman S. *Shifts in selective visual attention: towards the underlying neural circuitry*. In: Vaina L. M. (eds.). *Matters of intelligence. Synthese library (Studies in epistemology, logic, methodology, and philosophy of science)*, Springer, 1987, vol. 188, pp. 115–141.
21. Wolfe J. M., Cave K. R., Franzel S. L. Guided search: an alternative to the feature integration model for visual search. *J. Exp. Psychol. Human.*, 1989, vol. 15, no. 3, pp. 419–433.
22. Parkhurst D., Law K., Niebur E. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 2002, vol. 42, no. 1, pp. 107–123.
23. Bruce N. D., Tsotsos J. K. Saliency based on information maximization. *Proc. 18th Int. Conf. "Neural Information Processing Systems"*, 2005, pp. 155–162.
24. Achanta R., Estrada F., Wils P., Susstrunk S. Salient region detection and segmentation. In: Gasteratos A., Vincze M., Tsotsos J. K. (eds.) *Computer Vision Systems, Int. Conf. Computer Vision Systems*, LNCS, 2008, vol. 5008, pp. 66–75.
25. Liu T., Yuan Z., Sun J., Wang J., Zheng N., Tang X., Shum H.-Y. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, vol. 33, no. 2, pp. 353–367.
26. Liu F., Gleicher M. Region enhanced scale-invariant saliency detection. *IEEE Int. Conf. Multimedia and Expo*, 2006, pp. 1477–1480.
27. Walther D., Koch C. Modeling attention to salient proto-objects. *Neural Networks*, 2006, vol. 19, no. 9, pp. 1395–1407.
28. He S., Lau R. W. H., Liu W., Huang Z., Yang Q. SuperCNN: a superpixelwise convolutional neural network for salient object detection. *Int. J. Computer Vision*, 2015, vol. 115, no. 3, pp. 330–344.
29. Li X., Zhao L., Wei L., Yang M. H., Wu F., Zhuang Y., Ling H., Wang J. Deepsaliency: multi-task deep neural network model for salient object detection. *IEEE Trans. Image Process.*, 2016, vol. 25, no. 8, pp. 3919–3930.
30. Cao F., Liu Y., Wang D. Efficient saliency detection using convolutional neural networks with feature selection. *Information Sciences*, 2018, vol. 456, pp. 34–49.
31. Ma Y.-F., Zhang H.-J. Contrast-based image attention analysis by using fuzzy growing. *Proc. 7th ACM Int. Conf. on Multimedia*, 2003, pp. 374–381.
32. Cheng M.-M., Zhang G.-X., Mitra N. J., Huang X., Hu S.-M. Global contrast based salient region detection. *Proc. 2011 IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 409–416.
33. Shi K., Wang K., Lu J., Lin L. PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors. *Proc. 2013 IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 2115–2122.
34. Li G., Yu Y. Deep contrast learning for salient object detection. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 478–487.
35. Bin S., Li Y., Ma L., Wu W., Xie Z. Temporally coherent video saliency using regional dynamic contrast. *IEEE Trans. Circuits and Systems for Video Technology*, 2013, vol. 23, no. 12, pp. 2067–2076.
36. Koutras P., Maragos P. A perceptually based spatio-temporal computational framework for visual saliency estimation. *Signal Processing: Image Communication*, 2015, vol. 38, pp. 15–31.
37. Koutras P., Zlatinsi A., Petros Maragos P. Exploring CNN-based architectures for multimodal salient event detection in videos. *Proc. 13th IEEE Image, Video, and Multidimensional Signal Processing*, 2018, pp. 1–5.
38. Klein D. A., Frintrop S. Center-surround divergence of feature statistics for salient object detection. *Proc. IEEE Int. Conf. Computer Vision*, 2011, pp. 2214–2219.
39. Yeh H.-H., Chu-Song Chen C.-S. From rareness to compactness: Contrast-aware image saliency detection. *Proc. IEEE Int. Conf. Image Process.*, 2012, pp. 1077–1080.
40. Xie Y.-L., Lu H.-C., Yang M.-H. Bayesian saliency via low and mid level cues. *IEEE Trans. Image Process.*, 2013, vol. 22, no. 5, pp. 16809–1698.
41. Achanta R., Shaji A., Smith K., Lucchi A., Fua P., Susstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, vol. 34, no. 11, pp. 2274–82.
42. Ren Z., Hu Y., Ling-Tien Chia L.-T., Rajan D. Improved saliency detection based on superpixel clustering and saliency propagation. *Proc. ACM Int. Conf. Multimedia*, 2010, no. 2, pp. 1099–1102.
43. Zhu L., Klein D. A., Frintrop S., Cao Z., Cremers A. B. Multi-scale region-based saliency detection using W2 distance on N-dimensional normal distributions. *Proc. 2013 IEEE Int. Conf. Image Process.*, 2013, pp. 176–180.

44. Frintrop S. *VOCUS: A visual attention system for object detection and goal-directed search*. LNAI 3899. Springer, 2006. 216 p.
45. Santella A., Agrawala M., DeCarlo D., Salesin D., Cohen M. Gaze-based interaction for semi-automatic photo cropping. *Proc. ACM CHI 2006 Conf. on Human Factors in Computing Systems*, 2006, vol. 1, pp. 771–780.
46. Chen L.-Q., Xie X., Fan X., Ma W.-Y., Zhang H., Zhou H.-Q. A visual attention model for adapting images on small displays. *Multimedia Syst*, 2003, vol. 9, no. 4, pp. 353–364.
47. Wang P., Zhang D., Zeng G., Wang J. Contextual dominant color name extraction for web image search. *Proc. 2012 IEEE Int. Conf. Multimedia and Expo Workshops*, 2012, pp. 319–324.
48. Ko B. C., Nam J.-Y. Object-of-interest image segmentation based on human attention and semantic region clustering. *J. Optical Society of America A: Optics and Image Science, and Vision*, 2006, vol. 23, no. 10, pp. 2462–2470.
49. Garcıa G. M., Klein D. A., Stuckler J., Frintrop S., Cremers A. B. Adaptive multi-cue 3D tracking of arbitrary objects. *DAGM/OAGM 2012: Pattern Recognition*, LNCS, 2012, vol. 7476, pp. 357–366.
50. Favorskaya M., Buryachenko V. *Fast salient object detection in non-stationary video sequences based on spatial saliency maps*. In: De Pietro G., Gallo L., Howlett R. J., Jain L. C. (Eds.) *Intelligent Interactive Multimedia Systems and Services*, SIST, 2016, vol. 55, pp. 121–132.
51. Wang L., Ouyang W., Wang X., Lu H. Visual tracking with fully convolutional networks. *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 3119–3127.
52. Dong C., Loy C. C., He K., Tang X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, vol. 38, no. 2, pp. 295–307.
53. Long J., Shelhamer E., Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, vol. 39, no. 4, pp. 640–651.
54. Patacchiola M., Cangelosi A. Head pose estimation in the wild using convo- lutional neural networks and adaptive gradient methods. *Pattern Recognition*, 2017, vol. 71, pp. 132–143.
55. Nogueira K., Penatti O. A., dos Santos J. A. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, 2017, vol. 61, pp. 539–556.
56. Ren S., He K., Girshick R., Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, vol. 39, no. 6, pp. 1137–1149.
57. Vig E., Dorr M., Cox D. Large-scale optimization of hierarchical features for saliency prediction in natural images. *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 2798–2805.
58. Kummerer M., Theis L., Bethge M. Deep gaze I: Boosting saliency prediction with feature maps trained on ImageNet. 2014. *arXiv preprint arXiv:1411.1045*.
59. Kummerer M., Wallis T. S., Gatys L. A., Bethge M. Understanding low-and high-level contributions to fixation prediction. *Proc. Int. Conf. Computer Vision*, 2017, pp. 4799–4808.
60. Liu N., Han J., Zhang D., Wen S., Liu T. Predicting eye fixations using convolutional neural networks. *Proc. 2015 IEEE Conf. Computer Vision Pattern Recognition*, 2015, pp. 362–370.
61. Huang X., Shen C., Boix X., Zhao Q. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. *Proc. IEEE Intern. Conf. on Computer Vision*, 2015, pp. 262–270.
62. Cornia M., Baraldi L., Serra G., Cucchiara R. A deep multilevel network for saliency prediction. *Proc. Int. Conf. Pattern Recognition*, 2016, pp. 3488–3493
63. Pan J., Sayrol E., Giro-i Nieto X., McGuinness K., O'Connor N. E. Shallow and deep convolutional networks for saliency prediction. *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 598–606.
64. Jetley S., Murray N., Vig E. End-to-end saliency mapping via probability distribution prediction. *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 5753–5761.
65. Liu N., Han J. A deep spatial contextual long-term recurrent convolutional network for saliency detection. 2016. *preprint arXiv:1610.01708*.
66. Bruce N. D., Catton C., Janjic S. A deeper look at saliency: feature contrast, semantics, and beyond. *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 516–524.
67. Mottaghi R., Chen X., Liu X., Cho N.-G., Lee S.-W., Fidler S., Urtasun R., Yuille A. The role of context for object detection and semantic segmentation in the wild. *Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 891–898.
68. Cornia M., Baraldi L., Serra G., Cucchiara R. Predicting human eye fixations via an LSTM-based saliency attentive model. 2016. *arXiv preprint arXiv:1611.09571*.
69. Huang G.-B., Zhu Q.-Y., Siew C.-K. Extreme learning machine: theory and applications. *Neurocomputing*, 2006, vol. 70, no. 1-3, pp. 489–501.
70. Tavakoli H. R., Borji A., Laaksonen J., Rahtu E. Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features. 2016. *arXiv preprint arXiv:1610.06449v1*.
71. Kruthiventi S. S., Ayush K., Babu R. V. DeepFix: A fully convolutional neural network for predicting human eye fixations. *IEEE Trans. Image Processing*, 2017, vol. 26, no. 9, pp. 4446–4456.
72. Pan J., Ferrer C. C., McGuinness K., O'Connor N. E., Torres J., Sayrol E., Giro-i Nieto X. SalGAN: Visual saliency prediction with generative adversarial networks. 2017. *arXiv preprint arXiv:1701.01081*.
73. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.

Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680. *arXiv preprint arXiv:1406.2661*.
74. Wang W., Shen J. Deep visual attention prediction. 2017. *arXiv preprint arXiv:1705.02544*.
75. Gorji S., Clark J. J. Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes. *Proc. 2017 IEEE Computer Vision and Pattern Recognition*, 2017, vol. 2, no. 4, 2017, pp. 2510–2519.
76. Jia S., Bruce N. D. B. EML-NET: An expandable multi-layer network for saliency prediction. 2018. *arXiv preprint arXiv:1805.01047*.
77. Leifman G., Rudoy D., Swedish T., Bayro-Corrochano E., Raskar R. Learning gaze transitions from depth to improve video saliency estimation. *Proc. IEEE Int. Conf. on Computer Vision*, 2017, vol. 3, pp. 1698–1707.
78. Bazzani L., Larochelle H., Torresani L. Recurrent mixture density network for spatiotemporal visual attention. *Proc. Int. Conf. Learning Representations*, 2017, pp. 1–15.
79. Liu Y., Zhang S., Xu M., He X. Predicting salient face in multipleface videos. *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 4420–4428.
80. Souad Chaabouni, Benois-Pineau J., Hadar O., Ben Amar C. Deep learning for saliency prediction in natural video. 2016. *arXiv preprint arXiv:1604.08010*.
81. Jiang L., Xu M., Wang Z. Predicting video saliency with object-to-motion CNN and two-layer convolutional LSTM. 2017. *arXiv preprint arXiv:1709.06316*.
82. Gorji S., Clark J. J. Going from image to video saliency: Augmenting image salience with dynamic attentional push. *Proc. 2018 IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 7501–7511.
83. Bak C., Kocak A., Erdem E., Erdem A. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia*, 2018, vol. 20, no. 7, pp. 1688–1698.
84. Wang W., Shen J., Guo F., Cheng M.-M., Borji A. Revisiting video saliency: A large-scale benchmark and a new model. *Proc. 2018 IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 4894–4903.
85. Sun M., Zhou Z., Hu Q., Wang Z., Jiang J. SG-FCN: A motion and memory-based deep learning model for video saliency detection. *IEEE Transactions on Cybernetics*, 2018, pp. 1–12.
86. Tran D., Bourdev L., Fergus R., Torresani L., Paluri M. Learning spatiotemporal features with 3D convolutional networks. *Proc. 2015 IEEE Int. Conf. Computer Vision*, 2015, pp. 4489–4497.
87. Koutras P., Maragos P. SUSiNet: See, Understand and Summarize it. 2019. *arXiv preprint arXiv:1812.00722v2*.
88. Hara K., Kataoka H., Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2d CNNs and ImageNet. *Proc. 2018 IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.
89. Xu P., Ehinger K. A., Zhang Y., Finkelstein A., Kulkarni S. R., Xiao J. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. 2015. *arXiv preprint arXiv:1504.06755*.
90. Jiang M., Huang S., Duan J., Zhao Q. SALICON: Saliency in context. *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1072–1080.
91. Kim N. W., Bylinskii Z., Borkin M. A., Gajos K. Z., Oliva A., Durand F., Pfister H. BubbleView: an alternative to eyetracking for crowdsourcing image importance. 2017. *arXiv preprint arXiv:1702.05150*.
92. *MIT saliency benchmark*. Available at: http://saliency.mit.edu/datasets.html (accessed 17 April 2019).
93. Judd T., Durand F., Torralba A. A benchmark of computational models of saliency to predict human fixations. 2012. *Technical report MIT-CSAIL-TR-2012-001*.
94. Borji A., Laurent Itti L. CAT2000: A large scale fixation dataset for boosting saliency research. 2015. *arXiv preprint arXiv:1505.03581*.
95. *Saliency in Context*. Available at: http://salicon.net/ (accessed 18 April 2019).
96. *MSRA10K Salient Object Database*. Available at: https://mmcheng.net/msra10k/ (accessed 11 April 2019).
97. *Extended Complex Scene Saliency Dataset (ECSSD)*. Available at: http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/dataset.html (accessed 11 April 2019).
98. Li G., Yu Y. Visual Saliency based on multiscale deep features. *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 5455–5463.
99. *Visual Object Classes Challenge 2010 (VOC2010)*. Available at: http://host.robots.ox.ac.uk/pascal/VOC/voc2010/index.html (accessed 18 April 2019).
100. Everingham M., Van Gool L., Williams C. K. I., Winn J., Zisserman A. The PASCAL visual object classes (VOC) challenge. *Int. J. Computer Vision*, 2010, vol. 88, no. 2, pp. 303–338.
101. *Salient Objects Dataset (SOD)*. Available at: http://elderlab.yorku.ca/SOD/ (accessed 18 April 2019).
102. *The DUT-OMRON Image Dataset*. Available at: http://saliencydetection.net/dut-omron/ (accessed 18 April 2019).
103. *DAVIS: Densely Annotated VIdeo Segmentation*. Available at: https://davischallenge.org/ (accessed 15 April 2019).
104. *The Large-scale Eye-tracking Database of Videos (LEDOV) for video saliency*. Available at: https://github.com/remega/LEDOV-eye-tracking-database (accessed 15 April 2019).
105. *The DIEM Project*. Available at: https://thediemproject.wordpress.com/ (accessed 18 April 2019).
106. *UCF Sports Action Data Set*. Available at: https://www.crcv.ucf.edu/data/UCF_Sports_Action.php (accessed 18 April 2019).
107. Mathe S., Sminchisescu C. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, vol. 37, no. 7, pp. 1408–1424.

108. *COGNIMUSE Database*. Available at: http://cognimuse.cs.ntua.gr/database (accessed 18 April 2019).
109. Zlatintsi A., Koutras P., Evangelopoulos G., Malandrakis N., Efthymiou N., Pastra K., Potamianos A., Maragos P. COGNIMUSE: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing*, 2017, vol. 2017, no. 1, p. 54.
110. Bylinskii Z., Judd T., Oliva A., Torralba A., Durand F. What do different evaluation metrics tell us about saliency models? 2017. *arXiv preprint arXiv:1604.03605*.
111. Kummerer M., Wallis T. S., Bethge M. Information-theoretic model comparison unifies saliency metrics. *Proc. National Academy Sciences*, 2015, vol. 112, no. 52, pp. 16054–16059.
112. Pele O., Werman M. A linear time histogram metric for improved sift matching. *Proc. 10th European Conf. Computer Vision*, 2008, part III, pp. 495–508.
113. Tasse F. P., Kosinka J., Dodgson N. A. Quantitative analysis of saliency models. *Proc. SIGGRAPH ASIA*, 2016, Technical Briefs, pp. 19.1–19.4.
114. Shtrom E., Leifman G., Tal A. Saliency detection in large point sets. *Proc. 2013 IEEE Int. Con. Computer Vision*, 2013, pp. 3591–3598.
115. Rusu R. B., Blodow N., Beetz M. Fast point feature histograms (FPFH) for 3D registration. *Proc. 2009 IEEE Int. Conf. Robotics and Automation*, 2009, pp. 1848–1853.
116. Song R., Liu Y., Martin R. R., Rosin P. L. Mesh saliency via spectral processing. *ACM Trans. Graph.*, 2014, vol. 33, no. 1, pp. 6:1–6:17.
117. Tasse F. P., Kosinka J., Dodgson N. Cluster-based point set saliency. *Proc. 2015 IEEE Int. Conf. Computer Vision*, 2015, pp. 163–171.
118. Krizhevsky A., Sutskever I., Hinton G. E. ImageNet classification with deep convolutional neural networks. *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
119. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. *arXiv preprint arXiv: 1409.1556*.

УДК 004.932
doi:10.31799/1684-8853-2019-3-10-36

Обнаружение значимости в видеоконтенте в эпоху глубокого обучения: тенденции развития

М. Н. Фаворская[а], доктор техн. наук, профессор, orcid.org/0000-0002-2181-0454, favorskaya@sibsau.ru
Л. Ч. Джайн[б, в, г], PhD, профессор, orcid.org/0000-0001-6176-3739
[а]Сибирский государственный университет науки и технологий им. академика М. Ф. Решетнёва, Красноярский рабочий пр., 31, Красноярск, 660037, РФ
[б]Университет Канберры, ул. Киринари, 11, Брюс ACT 2617, Канберра, Австралия
[в]Ливерпульский университет Надежды, Парк надежды, L16 9JD, Ливерпуль, Великобритания
[г]Технологический университет Сиднея, PO Box 123, Бродвей NSW 2007, Сидней, Австралия

Постановка проблемы: обнаружение значимости в видеоконтенте является фундаментальной задачей компьютерного зрения. Конечной целью обнаружения значимости является локализация объектов интереса, которые привлекают внимание человека относительно остальной части изображения. Большое разнообразие моделей значимости, основанных на различных подходах, разработано с 1990-х годов. В последние годы обнаружение значимости стало одной из активно изучаемых разделов в теории сверточных нейронных сетей. Много оригинальных решений на основе сверточных нейронных сетей было предложено для обнаружения значимых объектов и даже событий. **Цель:** подробный обзор методов обнаружения значимости в эпоху глубокого обучения, который позволит понять возможности сверточных нейронных сетей для визуального анализа, проводимого с помощью слежения за глазами человека и цифровой обработки изображений. **Результаты:** обзор отражает последние достижения при решении задачи обнаружения значимости с использованием сверточных нейронных сетей. Различные модели, доступные в литературе, такие как статические и динамические 2D сверточные нейронные сети для обнаружения объектов значимости и 3D сверточные нейронные сети для обнаружения значимых событий, обсуждаются в хронологическом порядке. Стоит отметить, что автоматическое обнаружение значимых событий в продолжительных видеопоследовательностях стало возможным с использованием недавно появившихся 3D сверточных нейронных сетей в сочетании с 2D сверточными нейронными сетями для обнаружения значимых звуковых сигналов. В статье дано краткое описание общедоступных наборов изображений и видеопоследовательностей с аннотированными значимыми объектами или событиями, а также представлены часто используемые метрики для оценки результатов. **Практическая значимость:** данный обзор рассматривается как вклад в изучение быстро развивающихся методов глубокого обучения для задачи обнаружения значимости на изображениях и видеопоследовательностях.

Ключевые слова — обнаружение регионов значимости, обнаружение значимых объектов, обнаружение значимых событий, глубокое обучение, сверточная нейронная сеть, извлечение признаков.

Для цитирования: Favorskaya M. N., Jain L. C. Saliency detection in deep learning era: trends of development. *Информационно-управляющие системы*, 2019, № 3, с. 10–36. doi:10.31799/1684-8853-2019-3-10-36
For citation: Favorskaya M. N., Jain L. C. Saliency detection in deep learning era: trends of development. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2019, no. 3, pp. 10–36. doi:10.31799/1684-8853-2019-3-10-36