Received January 25, 2022, accepted February 11, 2022, date of publication February 16, 2022, date of current version March 1, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3152189

SATSal: A Multi-Level Self-Attention Based Architecture for Visual Saliency Prediction

MAROUANE TLIBA[1], MOHAMED A. KERKOURI[1], BASHIR GHARIBA[2], ALADINE CHETOUANI[1], (Member, IEEE), ARZU ÇÖLTEKIN[3], MOHAMED SHEHATA[4], (Senior Member, IEEE), AND ALESSANDRO BRUNO[5]

[1]Laboratoire PRISME, Université d'Orleans, 45067 Orleans, France
[2]Department of Electrical and Computer Engineering, Faculty of Engineering, Elmergib University, Khoms, Libya
[3]Institute of Interactive Technologies, University of Applied Sciences and Arts Northwestern Switzerland, 4132 Windisch, Switzerland
[4]Department of Computer Science, The University of British Columbia, Kelowna, BC V6T 1Z4, Canada
[5]Department of Computing and Informatics, Faculty of Science and Technology, Bournemouth University, Poole BH12 5BB, U.K.

Corresponding author: Alessandro Bruno (abruno@bournemouth.ac.uk)

ABSTRACT Human visual Attention modelling is a persistent interdisciplinary research challenge, gaining new interest in recent years mainly due to the latest developments in deep learning. That is particularly evident in saliency benchmarks. Novel deep learning-based visual saliency models show promising results in capturing high-level (top-down) human visual attention processes. Therefore, they strongly differ from the earlier approaches, mainly characterised by low-level (bottom-up) visual features. These developments account for innate human selectivity mechanisms that are reliant on both high- and low-level factors. Moreover, the two factors interact with each other. Motivated by the importance of these interactions, in this project, we tackle visual saliency modelling holistically, examining if we could consider both high- and low-level features that govern human attention. Specifically, we propose a novel method SAtSal (Self-Attention Saliency). SAtSal leverages both high and low-level features using a multilevel merging of skip connections during the decoding stage. Consequently, we incorporate convolutional self-attention modules on skip connection from the encoder to the decoder network to properly integrate the valuable signals from multilevel spatial features. Thus, the self-attention modules learn to filter out the latent representation of the salient regions from the other irrelevant information in an embedded and joint manner with the main encoder-decoder model backbone. Finally, we evaluate SAtSal against various existing solutions to validate our approach, using the well-known standard saliency benchmark MIT300. To further examine SAtSal's robustness on other image types, we also evaluate it on the Le-Meur saliency painting benchmark.

INDEX TERMS Eye movements, low and high vision, saliency prediction, self-attention, visual attention.

I. INTRODUCTION

Visual attention consists of perceptual and cognitive mechanisms that empower humans to rapidly select and interpret the most interesting parts of a complex visual scene. For human information processing, selective mechanisms associated with attention work as a "data prepossessing bottleneck". Often, the selective mechanisms are a result of so-called bottom-up processes, in which the viewer is guided by perceptual signals and analyses the surroundings with no conscious intentions [8], [74]. However, cognitively-driven top-down mechanisms are equally important in the way humans direct their attention to selected elements, whether they are visual, auditory, olfactory or otherwise [22].

In addition to top-down vs bottom-up dichotomy, visual attention literature distinguishes overt from covert attention. Covert attention enacts when the eyes are not moving because focusing on a specific fixation point (one might intentionally pay attention to the peripheral information without moving the eyes). On the other side, overt attention relies on eye movements shifting from a location to another of a given visual scene; foveal processing enables capturing high levels of detail from objects of interest while suppressing the context information into a low-resolution, low-colour

The associate editor coordinating the review of this manuscript and approving it for publication was Mingbo Zhao.

processing mode [52], [70]. The scene is constantly analysed through rapid eye movements, i.e., saccades, by which visual attention processes scan rapidly new objects of interest. Such mechanisms help humans prioritise and filter stimuli from the early visual processing stages to later stages where higher-level cognitive processing can occur. The human ability to detect saliency of objects enables efficient scene scanning, and as such, it is one of the fundamental attention mechanisms [51].

Therefore, a visual attention-related modelling is known as saliency prediction [50]. Saliency prediction deals with detecting the most attention-grabbing regions of a given visual scene from a bottom-up perceptual perspective. For a given input, be it an image or a video sequence, saliency prediction algorithms encode each pixel of the visual scene with an intensity value [0,255] or [0,1], indicating the probability of the pixel to be salient [60]. The corresponding map returned by saliency algorithms is known as a *saliency map*. The dominant understanding in the field is that the higher the saliency value, the more likely observers' eye movements will be drawn to that area in the image or video frame, assuming that there are no top-down cognitive or task-driven bias. Saliency maps are usually visualized as blobs distributed around the regions that naturally stand out (or pop out) of the visual scene. Therefore, saliency maps are typically represented as density maps (or heat-map) of probabilities. The accuracy of a predicted saliency of a given scene is measured against the recorded eye movements on the same scene. That relies on the understanding of close relationships between eye movements and visual attention [62]. Examining statistics on different levels of visual saliency enables an in-depth understanding of the processes that govern human attention and, by extension, human behaviour. Due to its broad relevance, predicting human eye movement patterns and visual saliency has an impressive range of applications in computer vision and related fields such as image compression [37], image captioning [24], image retrieval [33], image re-targeting [63], quality assessment of multimedia content (i.e. image [3], [19], [20], stereo [64], 3D meshes [1], etc.), remote sensing [32], watermarking [36], map viewing [5], [53], indoor localization [31], perception [15], image enhancement [13], [21], healthcare [40] among many others.

Saliency prediction links to the pioneering work by Treisman and Gelade [74] on the feature integration theory. According to Treisman and Gelade [74], early visual features are registered as viewers perceive a visual scene for, then, being combined into a complete object-based perceptual identification. The latter also introduces the so-called pre-attentive and attentive stages, corresponding to bottom-up and top-down information processing. Being able to separate and organize the visual information hierarchically based on its perceptual saliency and its importance paves the way for mimicking human attention in mathematical models and makes feature-based models eligible for predicting salient regions from stimuli that are viewed freely (not task-driven). In terms of algorithmic developments, the seminal work by Koch and Ullman [50] establishes the basis of central saliency and incorporates low-level information, and building on this model, Itti *et al.* [42] proposed the first computational model of saliency. From this point forward, many biologically inspired computational architectures have been proposed, such as the graph-based visual salience, or GBVS [38]. Several computer vision and image anaylsis methods focus on the extraction of low-level and high-level features to model and detect objects in images and videos [4], [12].

The ever-increasing success of machine learning and deep learning approaches in vision-related computational processing has allowed a critical development for saliency prediction over the last few years. Convolutions Neural Networks (CNNs) reached out high accuracy rates in learning complex semantic representations from large-scale image recognition datasets [54]. Due to the advance of deep learning techniques in mimicking human behaviour, recent CNN-based saliency architectures reduced the gap between human eye movements (typical baseline for saliency studies) and the performance of prediction models remarkably. Current CNN-based models focus on high-level semantically-informative representations because low-level features contain a high ratio of noise signals which are not semantically helpful [30]. Another main drawback that affects the reliance only on these deep hierarchical CNNs representations is the problem of limited receptive field proportional to the network depth layer. The consistency of scene semantics has an effect on eye movements, i.e., the eye tends to remain fixated longer on objects that are semantically informative regarding a scene's content [39], [73]. If a visual scene contains too many objects, representational inconsistency of scene semantics increases, highly correlated with human eye movement during free viewing. Given the above, predicting object characteristics links to the accuracy of the visual attention model. Despite the availability of several "deep saliency models" (i.e., deep learning-based saliency models), much of the knowledge in psychology and neuroscience describing various aspects of human visual attention has not been adequately tackled yet [10], [51]. In some tasks, even the traditional saliency models (those proposed before deep-learning models) offer decent saliency predictions, which can be superior to psychophysical evaluations [51], possibly explained by the importance of the low-level features in an image in detecting the early fixations [29]. Both high and low-level features may serve a purpose in the way humans process visual information. Inspired by the previous studies considering human attention as a multilevel selection process [47], we integrate low and intermediate-level feature mapping to leverage the discriminant part of both low-level and deep semantic features to propose a new saliency model. To implement and test the proposed model, we incorporate convolutional Self-Attention modules on skip connections from the encoder to the decoder architecture opposite layers. As a result, the proposed model can effectively predict visual saliency patterns from multilevel contextual scene representations and overcomes the limitation of narrow receptive fields by employing the ability of Self-Attention to capture the context from an extended range of sequence dependencies.

The main contributions of this publication are summarised as follows:

- We develop a novel approach for visual saliency prediction using both high and low-level factors in learning multilevel features for producing static saliency maps. In addition, a self-attention module has been incorporated on the encoder-decoder skip connections to boost the global information in the deep layer and generate a highly representative saliency distribution.
- We evaluate the effectiveness of our model on the established MIT300 preserved benchmark and Le Meur [59] paintings dataset. All comparisons demonstrate that the proposed model is consistent, efficient, and superior to or competitive with other state-of-the-art methods.
- We further test out the robustness of the proposed approach in an ablation study on challenging scene samples that include both high and low-level features. The results reveal that multilevel skip attentive connections are effective and boost the performance of the backbone encoder-decoder model.

The rest of the manuscript is organized as follows: Section II introduces an overview of related saliency literature. Section III provides a detailed description of the proposed approach. Section IV demonstrates the benchmarking experiments and compares results to state-of-the-art methods. Finally, conclusions and outlook are given in Section V.

II. RELATED WORK

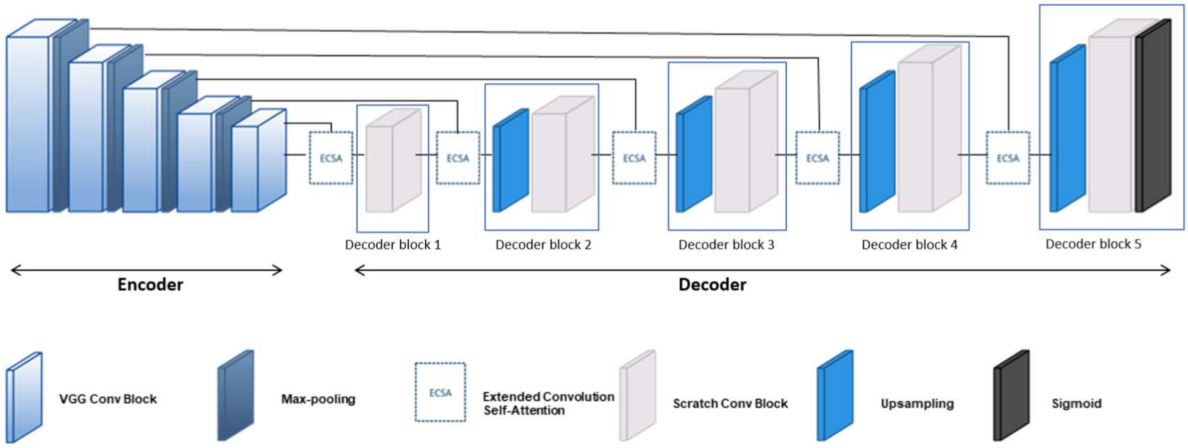
Visual attention modelling has been a topic of interest to computer vision for many decades, starting from the seminal work by Koch *et al.* [50], which was then implemented by [42] as a bottom-up model that predicts saliency using multi-scale low-level features. On top of the previous efforts, the GBVS [38] framework extracts image features to predict saliency using graph theory-based formulas that define Markov chains over different input maps. Zhang *et al.* [58] proposed a Bayesian framework tackling bottom-up saliency as self-information over linear visual features and the overall saliency as the point-wise mutual information between features and target. Around the same time, Bigdely-Shamlo *et al.* [65] proposed a method to detect visual saliency relying on the Kalman filter. Achanta *et al.* [2] extracted salient pixels in images using features of colour and luminance in the Fourier domain. Colour spaces and their role in saliency extraction were further investigated later by several researchers, e.g., [14].

The approaches mentioned above focus on visual attention using low-level spatial features. Sun and Fisher [73] introduced a hierarchical object and location-based visual attention model using a grouping-based salience. They treat complex visual tasks that depend on the current scene and the observer's goals, thus introducing a top-down cognitive aspect to saliency prediction. Integrating another feature of human cognitive processing, Jin-Gang and Gui-Song [43] presented an object-based saliency detection with a paradigm based on the Gestalt grouping cues. Kai-Yueh *et al.* [46] method introduced a model that is reliant on the relationships between saliency and "objectness", a concept in which a scene element is ranked for its meaning.

Recently saliency modelling gained remarkable performance by applying deep learning techniques which can learn top-down representations. This was achieved due to the construction of large scale eye movement datasets such as [7], [41], [44], [45] and [16]. The eye movement datasets mentioned above were collected using free-viewing eye-tracking sessions. The latter differs from task-driven scenarios where the variability of tasks could result in an unbalanced specificity of eye movements toward visual features related to the tasks. Among the deep learning approaches to saliency prediction, Ensemble of Deep Networks (eDN),Vig *et al.* [75] trained an early shallow CNN architecture that learns end-to-end saliency by merging different layer feature maps. The achieved performance did not mark an important result leap, as shallow networks cannot learn high-level features.

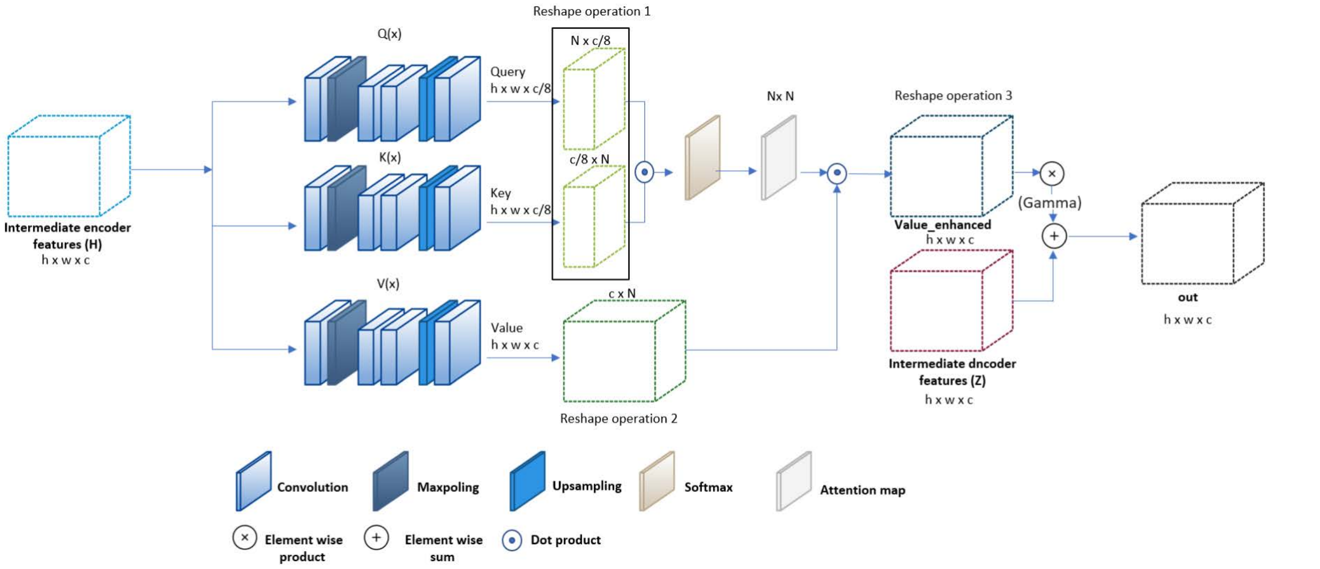
Lots of recent methods leveraged classification architectures pre-trained on ImageNet dataset [26]. These architectures have a superior ability at extracting the deep semantic representations from images [28].

As the deep learning-based models started to populate the scene, Oyama and Yamanaka [66] explored the influence of classification accuracy of the models on saliency estimation. The well-known DeepGaze1 by Kümmerer *et al.* [55] reemploys an early light object recognition network to explore the limits of deep learning in saliency prediction. Kümmerer *et al.* later introduced a new network, namely DeepGaze2 [57], based on the VGG [72] classification network. Both DeepGaze1 and DeepGaze2 models use fixed prior maps to regulate the possible biases in the data. Pan *et al.* [68] compare two approaches for predicting saliency in an end-to-end fashion. The first one is reliant on a lightweight network whose parameters are learnt from scratch. The second one is deeper and takes advantage of a pre-trained image classification network. In contrast, Cornia *et al.* [23] propose an architecture that extracts and combines features from multiple levels, then use a learned before tackling the bias of the dataset and introducing a new custom pixel-based loss function. Huang *et al.* [41] use a two-stream VGG network in their SALICON with different input scales, in which both output streams feature maps are concatenated to model the final saliency. DVA [76] learns multi-level information from different layers with different receptive field sizes, the decoders composed of a series of deconvolution layers with upsampling operations, the resulting multi-level maps are fused to produce the final saliency map. SalGAN [67] uses a deep convolution generative adversarial neural (GAN) network. The model design architecture contains a generator of saliency, and a discriminator network, where the two compete in a min-max game between the generator and the discriminator to produce a saliency map, which



- Apart from the first and the fourth ones, all decoder blocks consist of scratch conv and upsampling operation.
- We add a sigmoid activation to the last decoder block to have the output as a map of probabilities.

FIGURE 1. Network architecture of SATSal.



Reshape operation 1 : it reshapes the query tensor to (N x c/8). it reshapes the key tensor to (c/8 x N). Reshape operation 2 : it reshapes the value tensor to (c x N). Reshape operation 3 : it reshapes the resulted tensor form the dot product operation of the attention map and the vlaue tensor to (h x w xc).

* The goal from reshape operations is to calculate the similarity over all features locations using the dot product between 2D matrices.

FIGURE 2. The design of extended convolution self-attention module.

is qualitatively indistinguishable from the ground truth based on eye movement recordings. Liu and Han [61] propose a deep spatial contextual long-term recurrent convolutional network that learns local features on each image location in parallel via fine-tuning a pre-trained CNN model. Afterwards, the model simultaneously learns to incorporate global scene context to predict saliency. Cornia *et al.* [25] introduce a set of prior maps generated by a Gaussian function, the use of a neural attention mechanism and convLSTM (convolutional LSTM) layers on feature maps to refine the predicted saliency maps iteratively. Most recently, [48] proposed SALYPATH, an architecture to simultaneously predict saliency and associated scan-path, using a combined loss function that uses pixel level and distribution functions and a Noise Sensitivity Score (NSS) [69] metric.

Table 1 provides a summary overview of the most prominent methods mentioned above in terms of their major contributions to predict saliency on natural images.

TABLE 1. An overview of previous state the art methods.

Models	Major contribution to predict saliency maps
SAM-ResNet [25]	ResNet and RNN to refine the features
SalGAN [67]	Trains a deep convolution generative neural network
DVA [76]	Combination VGG high level features
SAM-VGG [25]	VGG and RNN to refine the features
ML-Net [23]	High multilevel VGG features and learnable bias
DeepGaze [55]	VGG and fixed prior maps to regulate possible range of biases
SALICON [41]	Two streams of VGG network on two different scales

III. PROPOSED MODEL

In light of the prior work, this section describes the architecture of the proposed model, as illustrated in Figure 1. It consists of a VGG-Encoder network and the Decoder-network composed of five deconvolution blocks interspersed with an extended Self-Attention module (ECSA). ECSA takes the previous block decoded hidden-vector and the opposite layer in Encoder features as input using a skip connection; the ECSA output is fed into the next Decoder block to produce a saliency map at the final stage. The motivation behind the proposed model are explained below:

- CNNs can only process features in a local neighbourhood, thus making it inefficient to model long-range, multilevel dependencies across spatial regions. Instead, incorporating self-attention modules connecting multilevel Encoder and Decoder networks yield robust predictions. Therefore, fine details at every position are properly considered with others in distant portions of low-level attention from early layers. That also helps to overcome the limited receptive field issue in learning object-based correlation from deep semantic representations.
- Most recent studies consider the high-level features as they focus on solving the complex top-down attention problem, meanwhile underestimating the importance of capturing global and local low-level attention. Multilevel skip connections help in leveraging the deep semantic representations from the last decoder layers and simple, attractive structural features from the first encoder layers, which enhance the modelling of better, more representative saliency distributions than those examining only high-level features.

A. EXTENDED CONVOLUTIONAL SELF-ATTENTION

The so-called integrated attention mechanism has recently shown important improvements in the performance of various downstream computer vision tasks [18], [27], [48]. Unlike the absolute attention mechanisms, the mechanism mentioned above learns in a fully adaptive, joint, and task-oriented manner, which allows the network to prioritise and associate weights to feature vectors. The self-attention or intra-attention calculates the response at a position in a vector by attending all positions within the same vector. In greater detail, the self-attention module draws the relationship between distant features, incorporating the module at multilevel connections on the Encoder-Decoder network layers. The latter prompts the model to generalise better static visual attentive cues at low and high levels, boosting the representation capability of the full network. The prediction performance of this design generalises well across various static saliency datasets fig 1.

The main goal of self-attention is to determine a new set of vector values representing global vector features dependency. Thus, Self-attention reveals the set of values to pay more attention to the interaction of input vector features. In simpler words, for a given vector, we need to extract query, key and value vectors from it, simulating the selection process applied in system retrieval. The latter measures attention by calculating a similarity between a query and best related key features using a score function. The output scores go through the normalisation step to have the sum of probability values to one. The final value vector is a weighted combination of the previous value vectors based on the normalised score result. The overall architecture of the proposed extended self-attention is described in figure 2.

In equation 1, the hidden, encoded features of the ith VGG encoder block are given as a function of the input image $X \in \mathbb{R}^{256\times128\times3}$,

$$H_i = f_{0-i}(X) \in \mathbb{R}^{h_i \times w_i \times c}, \tag{1}$$

where f_{0-i} is the ith VGG encoder block, h_i and w_i are the down-sampled input height and width after the ith max-pooling operation, except for $i = 5$ which denotes the last encoder output $H_5 = Z_0$, here the max-pooling is not applied. We denote the decoded variable after the ith block by Z_i. Each of the ECSA modules is placed just before each decoder block and takes as input both of Z_i and H_{5-i}, and transform the intermediate features H_{5-i} into three variables Q, K and V, unlike [77] that incorporate just one layer of 1*1 convolution without activation function, we extend our implementation by a shallow series of activated CNNs interspersed with down-sampling and upsampling operation, the ECAS module architecture slightly differ corresponding to the ith positional block, because we are extracting an attention vector from the Encoder layer position and inject it into the Decoder which mean that the two vector spaces are not similar, so that a deep transformation need to be applied, of course taking into consideration the computation efficiency of the whole architecture. The resulted couple feature spaces (Query, Key) $\in \mathbb{R}^{\bar{C}\times N}$ from Q(H_i) and K(H_i) respectively, simplifying the dimension of $H_i \in \mathbb{R}^{C\times N}$, where N $= h_i \times w_i$ representing the number of feature location,and $\bar{C}$ the number of output channel from both of Q and K stream which is equal to the C/8. The attention map resulted after normalizing the output of dot product between Query vectors and key vectors using a Softmax function, where S represent the similarity between Query and key feature spaces:

$$S_{lj} = Query[l]^T.Key[j] \tag{2}$$

$$A_{j,l} = \frac{\exp(S_{lj})}{\sum_{j=1}^{N}\exp(S_{lj})}, \tag{3}$$

The attention map A$\in \mathbb{R}^{N\times N}$ shows the likelihood that a particular positional feature in l_{th} location appears in the jth location in N feature locations, $(_{j,l})\in \mathbb{R}^{N}$, the Value feature space is further enhanced by multiplying it to the attention map:

$$Value_{enhanced} = Value.A \tag{4}$$

The dimension of the context vector, which is the enhanced value feature space, is equal to Z_i dimensions, moreover we scaled the context by a learnable parameter γ in order to learn how much the decoder network should relay on the context from the the encoder features at each stage. Finally, we add it to the decoded variable Z_i

$$out = \gamma \times Value_{enhanced} + Z_i \tag{5}$$

Furthermore, our approach offers flexibility in that it has no restriction regarding receptive field dimensions, using the capability of self attention in capturing distant features, i.e., the system theoretically can work with any width and height input image.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

1) LOSS FUNCTION AND TRAINING STAGE

The whole model, including the encoder-decoder and the ECSA modules, was trained using the loss function noted as L. The L loss function is defined as a combination between the Kullback-Leibler Divergence(KLD), the Normalized Scanpath Saliency (NSS), and the Binary Cross-Entropy (BCE). Each term (KLD, NSS, BCE) covers a particular aspect for learning the best set of weights [11]. Specifically, KLD evaluates the mutual distribution between the predicted output and the ground truth, BCE is used for binary classification of each CNN output vector independently, and NSS [69] provides a saliency metric that measures the mean saliency value at ground-truth fixation locations. We detail each approach and their contribution to L loss function in our model below. Our assumption behind using a combination of metrics as an objective function, back to optimize the loss toward the weights that lead to the best results in capturing more accurate representative saliency distributions.

Let the predicted map $Y \in \left[0, 1\right]^{256\times 192}$, the fixation map $F \in \{0, 1\}^{256\times 256}$, and the ground truth saliency $\hat{Y} \in \left[0, 1\right]^{256\times 192}$.

$$L = 0.7 \times \mathcal{L}_{\mathrm{KLD}}(Y, \hat{Y}) + \mathcal{L}_{\mathrm{BCE}}(Y, \hat{Y}) - 0.3 \times \mathcal{L}_{\mathrm{NSS}}(Y, \hat{P}) \tag{6}$$

BCE is mainly designed for calculating the distance between two normalized distributions in the interval [0, 1]. In probabilistic terms, BCE measures the accuracy of the modeled probability distribution of saliency for a given input image pixel.

$$BCE(Y, \hat{Y}) = -1/m \sum_{i=1}^{\mathrm{m}} Y_i \log(\hat{Y}_i) + (1 - Y_i)\log(1 - \hat{Y}_i) \tag{7}$$

$\mathcal{L}_{\mathrm{KLD}}$ has been widely used for training saliency models as it often used as one of the metrics in different benchmarks. It is chosen as a weighted main loss in our work.

$$\mathcal{L}_{\mathrm{KLD}}(Y, \hat{Y}) = \sum_i \hat{Y}_i \log\left(\epsilon + \frac{\hat{Y}_i}{\epsilon + Y_i}\right). \tag{8}$$

$\mathcal{L}_{\mathrm{NSS}}$ is adopted from the standard NSS metric, which is a similarity metric. Their negatives are used for minimization in order to optimise the model weight in the right direction, the goal from adding the $\mathcal{L}_{\mathrm{NSS}}$ loss is to maximise the similarity metric results:

$$\mathcal{L}_{\mathrm{NSS}}(Y, \hat{P}) = -\frac{1}{N}\sum_i \bar{Y}_i \times \hat{P}_i, \tag{9}$$

where $\bar{Y}_i = (Y_i - \mu(Y_i))/\sigma(Y_i)$. and $N =$ refer to the sum of fixations.

2) IMPLEMENTATION DETAILS

We implemented our model in PyTorch and trained the model on the MIT1003 dataset, using 900 images for training and 103 images for validation. We initialised the encoder with the pre-trained VGG [72], and both the decoder and the attention modules were randomly initialised using the Xavier method [35]. We used the Adam optimiser [49] to train the model. We opted in for a learning rate of 10^{-4} and a scheduler step with a dividing factor of 2 every 20 epochs. During the first ten epochs, the ECSA parameter γ was set to zero to focus on learning the main task. At the same time, the decoder layers gradually froze, starting from the bottom to the top and progressively increasing the complexity. After the first ten epochs, the whole model was trained end-to-end, including all parameters.

3) COMPUTATIONAL LOAD

The embedded self-attention modules are trained in an end-to-end manner with the encoder-decoder backbone model. The entire training procedure takes about 5 hours on Google Colab environment with a single NVIDIA Tesla T4 GPU and a 2.0GHz Intel(R) Xeon(R) CPU. Since our model does not need any pre or post-processing steps, it takes only about 0.0106 s to process an image of size 256 × 192.

B. EXPERIMENTAL RESULTS

In this section, we evaluate our model on the MIT300 benchmark dataset [56], which is one of the most well-known benchmarks for saliency models. The dataset consists of 300 natural images; the corresponding saliency maps are preserved privately for a fair comparison. We also used the newly published Le Meur [59] Paintings dataset, which offers a different, more specialised stimuli space as the paintings differ in many ways from natural scenes. Testing our approach on multiple types of stimuli helps us study our model's performances on different datasets. We also want to demonstrate the effectiveness of the extended self-attention module in capturing the global representations on another type of space, in which inherently different cues would attract the viewers' gaze compared to natural scenes. Le Meur's dataset consists of 150 painting images related to five different art periods and their respective saliency maps. We used the entire dataset for testing.

1) COMPETITORS

We compare our model with a representative set of stat-of-the-art models, namely, SALICON [41], DeepGaze1 [55], SAMCornia [25], and ML Net [23]. We selected these models due to their ability to address visual attention on different stimuli domains, e.g., indoor, outdoor, painting. For the sake of generality on low-level attention, we further compare our model with some previous static attention models and frameworks, i.e., Itti & Koch model [42], and the GBVS [38].

2) METRICS

We conducted comparisons of our model's results against the selected competitors using six saliency metrics, which are divided into two categories:

- **Distribution-based metrics:** These metrics allow comparing the predicted saliency map to the ground-truth distribution from eye movement recordings. We used three of them, namely, (KLD) Kullback-Leibler Divergence, Similarity Metric (SIM), Linear Correlation Coefficient (CC).
- **Location-based metrics:** These metrics compute some statistics of fixation locations, such as Normalized Scanpath Saliency (NSS), Area under Curve (AUC) and its derivative AUC-Judd (AUC-J), and shuffled AUC (s-AUC).

Reference articles [6], [17] provide more detailed descriptions of all the metrics used in our experiments.

3) PERFORMANCE

We calculated the results on the MIT300 dataset by sending the output prediction to the active benchmark service. At the same time, we tested our method over Le Meur's dataset using the same protocol described in their work workLeMeur. Table 2 shows our results on the MIT300 benchmark. As Table 2 demonstrates, our model scored the highest among the comparative models on both CC and SIM metrics for this dataset while achieving a very close second place for the AUC and scoring competitive results for the NSS and KLD. Next, we made the same comparisons for the second dataset (Le Meur paintings dataset). The outcomes from this comparison are shown in Table 3. As Table 3 demonstrates, our model achieves the highest score with the KLD metric and close second place with the SIM and CC for the painting data set, and it remains competitive for the remaining metrics.

TABLE 2. Comparative performance of different saliency models on MIT300 benchmark.

Models	AUC ↑	sAUC ↑	NSS ↑	CC↑	KLD ↓	SIM↑
SATSal (our model)	**0.851**	0.703	1.947	**0.703**	0.854	**0.614**
SAM-ResNet [25]	**0.852**	**0.739**	**2.062**	**0.689**	1.171	**0.612**
SalGAN [67]	0.849	0.735	1.862	0.674	0.757	0.593
DVA [76]	0.843	0.725	1.930	0.663	**0.629**	0.584
SAM-VGG [25]	0.847	0.730	1.955	0.663	1.274	0.598
ML-Net [23]	0.838	0.739	1.974	0.663	0.800	0.581
DeepGaze I [55]	0.842	0.723	1.723	0.614	0.667	0.571
SALICON [41]	0.814	0.739	1.702	0.562	0.782	0.516
GVBS [38]	0.806	0.629	1.245	0.479	0.887	0.483
IttiKoch [42]	0.543	0.535	0.408	0.130	1.496	0.337

TABLE 3. Comparative performance of different saliency models on Le Meur paintings dataset.

Models	AUC ↑	AUC B ↑	NSS ↑	CC↑	KLD ↓	SIM↑
SATSal (our model)	0.827	**0.790**	1.530	**0.647**	**0.737**	**0.579**
GVBS [38]	0.817	**0.809**	1.256	0.506	0.962	0.446
RARE2012 [71]	0.786	0.777	1.103	0.443	1.020	0.438
AIM [9]	0.735	0.723	0.772	0.315	1.245	0.371
AWS [34]	0.769	0.762	1.083	0.427	1.045	0.430
ML-Net [23]	0.818	0.770	1.524	0.576	**0.832**	0.513
DeepGaze II [57]	0.804	0.679	1.394	0.485	0.869	0.488
SALICON [41]	0.827	0.708	1.445	0.538	0.880	0.517
SAM-ResNet [25]	**0.862**	0.782	**1.834**	**0.700**	0.984	**0.613**
SAM-VGG [25]	0.846	0.752	1.603	0.617	0.970	0.561

Figures 3 and 4 illustrate the qualitative results (i.e., the visual outputs) of our model against the "ground truth" (i.e., the eye movement data) and other state-of-the-art models. In these two figures (Fig 3 and 4), we can see the stimuli overlain with predictions and ground truth saliency maps for the MIT300 and Le Meur dataset. It is immediately clear that our model can capture both the global and the local attention patterns, demonstrating an important generalisation capability for different image (i.e., scene content) distributions. The continuity between the most intensive salient regions to the effect of the self-attention in extracting the global context of the scene captures the most salient objects. It also provides intuition on how our attention could be swapped from one object to another. Thus, it would be beneficial for the case of scan path prediction. Figures 5 shows the effectiveness of our model to predict salient regions on synthetic images characterised with low-level features standing out of the visual stimulus e.g., shape, contrast, colour, orientation.

Based on these findings, SAtSal (our proposed model) performs superior to and competitive with previous state-of-the-art models. For the sake of fair comparisons, we used some standard metrics accounting for the effectiveness of saliency detection in static images. Thus, overall results demonstrate that the method is robust over multiple datasets.

C. ABLATION STUDY

This section provides a detailed evaluation of the proposed approach from several aspects through an ablation study to verify the effectiveness of the proposed multilevel Self-Attention modules and examine the influence of different training protocols. We conducted the ablation study on two subsets, one that contains natural images from (MIT1003 by [45]), and another that has explicitly low-level features from (CAT2000 by [7]). We consider this protocol to study the effect of multilevel self-attention modules on natural images using different settings. We first tackled images that contain both bottom-up and top-down attention stimuli. The same patterns exist in the distribution of images from Le Meur

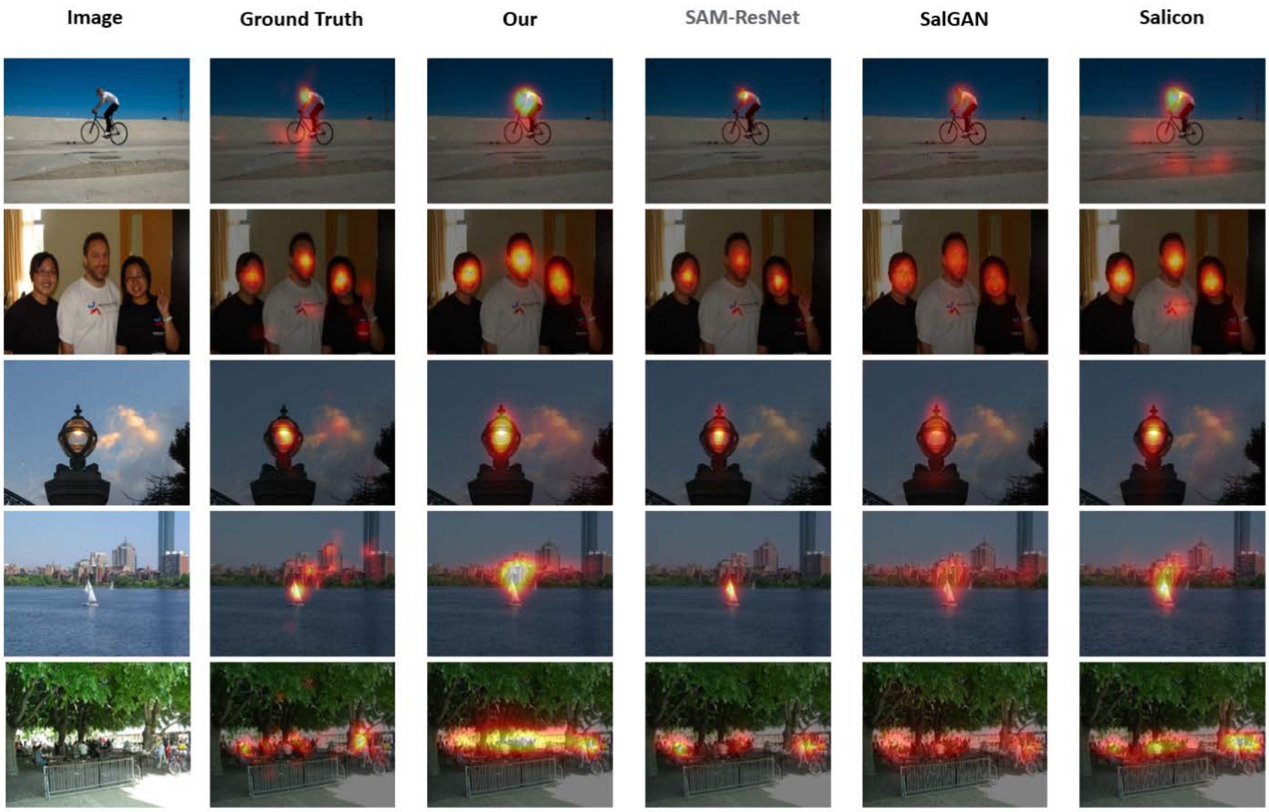


FIGURE 3. Visualisation of the results: Saliency maps for the samples from MIT1003 cross validation-set.

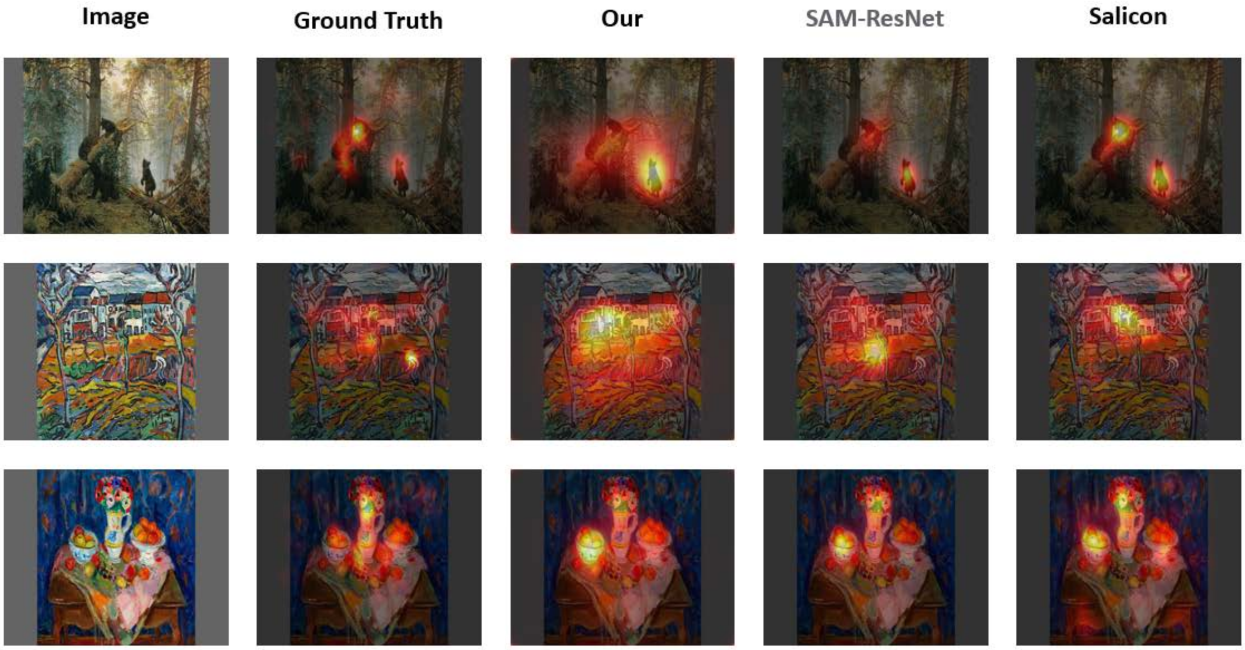


FIGURE 4. Visualisation of the results: Saliency maps for the samples from Le Meur dataset.

and MIT300 test sets. Therefore, to further examine the robustness regarding only the bottom-up cues, we conduct the same test protocol on a specific category of images containing only low-level features with no semantic meaning. Also, we are restricted in this ablation test protocol to exclude models trained on the same data distribution to avoid

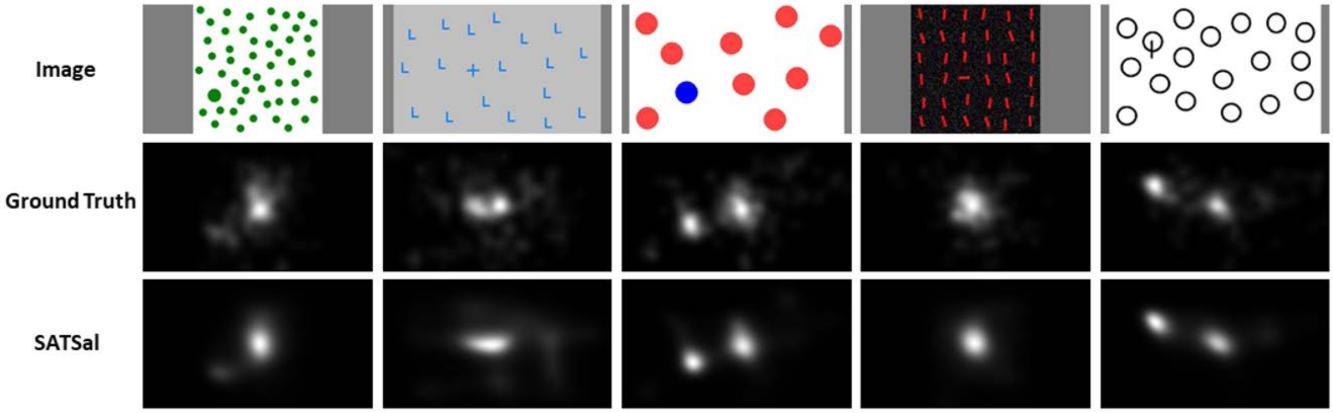


FIGURE 5. Visualisation of the results: Saliency maps for the images with low-level features.

training-test overlap. Below we detail the outcomes from the two ablation studies.

1) EFFECTIVENESS OF MULTILEVEL SELF-ATTENTION MODULES ON NATURAL IMAGES

First, we study the effect of multilevel self-attention modules (ECSA modules) by disabling the main components in the following three settings. Note that for each setting, we have used the same training protocol as the original model and tested the model on 100 images from the MIT1003 dataset [45]. The results are summarized in Table 3.

- **Setting 1**: We remove the last skip connection and the associated self-attention module (ECSA module) to evaluate the importance of the low-level features carried through this connection. We could remark a slight drop in performance regarding overall metrics (e.g., metric: value before ablation → value after ablation: AUC-J: 0.9207→ 0.9162, NSS: 3.2400→ 3.0933, CC: 0.8651→ 0.8435).
- **Setting 2**: We remove the mid-level skip connections and the associated self-attention modules (ECSA modules) on these connections. We observe a significant drop in performance even about Setting 1, demonstrating the importance of mid-level skip connections in modelling a better saliency distribution.
- **Setting 3**: We remove the high level skip connection and the associated self-attention module (ECSA module), resulting in a remarkable shot in performance (e.g., AUC-J: 0.9207→ 0.9144, NSS: 3.2400→ 2.9854, KLD: 0.3.975→ 0.0.4765), clearly demonstrating the importance of self-attention in capturing long-range of spatial dependencies and enhancing the high-level representation with an enlarged receptive field.

TABLE 4. Results of the ablation study on 100 images from MIT1003.

Model	Auc Judd ↑	NSS↑	CC↑	SIM↑	KLD ↓
Salgan [67]	0.8662	1.9460	0.5836	0.4908	1.0470
MLNet [23]	0.8509	2.1678	0.5787	0.4815	1.3083
SAM-VGG [25]	0.9050	2.9409	0.8144	0.6650	0.8500
SAM-ResNet [25]	0.9124	3.0934	0.8570	**0.7045**	0.8515
SATSal (our model)	**0.9207**	**3.2400**	**0.8651**	0.6961	**0.3975**
Setting 1	0.9162	3.0933	0.8435	0.6714	0.4503
Setting 2	0.9072	2.8250	0.7746	0.6004	0.5861
Setting 3	0.9144	2.9854	0.8340	0.6537	0.4765

TABLE 5. Results of ablation study on low-level patterns from CAT2000.

Model	CC↑	SIM↑	KLD ↓
SAMResnet [25]	0.9142	0.7861	0.5393
SATSal (our model)	**0.9448**	**0.8218**	**0.1550**
Setting 1	0.9334	0.8008	0.1623
Setting 2	0.9362	0.7941	0.1663
Setting 3	**0.9470**	**0.8227**	**0.1410**

The significant drop in Settings 2 and 3 compared to Setting 1 is caused by the nature of the testing dataset, which portrays images representing objects with high semantic meaning. Even though it is relatively subtle, the drop in performance with Setting 1 indicates the importance of the low-level features, which would be even more pronounced in other stimuli.

2) EFFECTIVENESS OF MULTILEVEL SELF-ATTENTION MODULES ON LOW-LEVEL IMAGE FEATURES

Since one of the strong points of our model is to integrate the low-level feature detection into deep learning-based saliency prediction in combination with mid and high-level features, we conducted an additional ablation study with a focus on low-level features. We repeated the ablation study on 100 images from the CAT2000 [7] using the same settings as in the previous section. Images from CAT2000 contain patterns prepared for perceptual psychology studies, with low-level features, including geometrical elements, pop-out, conjunction, search asymmetry, textures, etc. We present the results from this study in Table 4.

We select just the distribution-based metrics on this part of the study because we are interested in testing the model's accuracy in revealing one region of interest from the other non-attractive low-level features. Other fixations on this kind of scene located far from the areas of interest can be considered outliers that do not represent the bottom-up saliency of the scene.

With the CAT2000 dataset, we see a minor improvement (e.g., CC: 0.9448→ 0.9470) in Setting 3 compared to the proposed approach. We believe this may be due to the nature of the scenes in CAT2000, as they do not contain much semantic meaning. Thus there is no need to calculate the attention for high-level deep representations. However, the drop in performance is quite evident in Settings 1 and 2 compared to the results obtained from SATSAL. The learned information from low-level features on the multilevel skip self-attention modules are essential for modelling better saliency and could boost the performance on a given general scene.

V. DISCUSSION AND CONCLUSION

In this paper, we were set out to build, implement and test a new architecture for visual attention modelling, specifically, for saliency prediction. Unlike most previous methods, we designed our approach to predict saliency from a more "holistic" perspective, accounting for both bottom-up (low level) and town-down (high level) features in a scene. Our model has shown great flexibility (thus, early signs of generalisability) in predicting visual saliency over datasets containing images with inherently different visual characteristics, precisely, natural scenes, paintings, as well as highly simplified perceptual psychology stimuli. SATSal's saliency scores are either superior or competitive against the state-of-the-art models based on multiple metrics.

We introduced an extended CNN self-attention module, using skip connection on multiple levels to model the representation of low and high-level features equally to capture local and global factors that attract human attention. Our approach enables local features to model human visual attention after filtering them out of the noise and merging them with deeper global representations. The steps mentioned above finally allow global and local visual information to generate more accurate predictions than models focusing only on low-level or high-level features. Specifically, the main contribution of our work is a new architecture that can capture relations between separated spatial dependencies from multiple hierarchical levels. Furthermore, the steps, as mentioned earlier, improved the accuracy of the extracted saliency maps because it takes all stimulus features into account.

We evaluated our model on a well-known benchmark and a newly proposed dataset, attaining competitive results with a representative set of state-of-the-art models. Although the model is trained on a small set of data, both quantitative and qualitative outcomes demonstrate the effectiveness and robustness of our model and its capability to generalise against different data distributions. Furthermore, SATSal's performance provides evidence on the importance of taking multiple level features into account in improving saliency prediction. As a future extension, we intend to address the temporal dimension to predict fixations and their duration. To do that, we aim to employ the capability of self-attention in capturing the temporal dimension while exploiting the contextual and semantic characteristics of the stimuli. This work also opens questions about the interpretability of deep saliency models, the features responsible for improving saliency prediction, the reason behind the accuracy rates from one distribution to another covering different cues. Finally, we consider bottom-up and deep semantic cues contributions in qualitative and quantitative results.

REFERENCES

[1] I. Abouelaziz, A. Chetouani, M. E. Hassouni, L. Latecki, and H. Cherifi, "3D visual saliency and convolutional neural network for blind mesh quality assessment," *Neural Comput. Appl.*, vol. 32, pp. 16589–16603, Oct. 2019.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[3] E. Ardizzone and A. Bruno, "Image quality assessment by saliency maps," in *Proc. VISAPP*, 2012, pp. 479–483.

[4] Q. Bai, S. Li, J. Yang, Q. Song, Z. Li, and X. Zhang, "Object detection recognition and robot grasping based on machine learning: A survey," *IEEE Access*, vol. 8, pp. 181855–181879, 2020.

[5] K. Bektaş, A. Çöltekin, J. Krüger, and A. T. Duchowski, "A testbed combining visual perception models for geographic gaze contingent displays," in *Proc. Eurograph. Conf. Vis., EuroVis-Short Papers*. The Eurographics Association, 2015, pp. 67–71, doi: 10.2312/eurovisshort.20151127.

[6] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2012.

[7] A. Borji and L. Itti, "CAT2000: A large scale fixation dataset for boosting saliency research," 2015, *arXiv:1505.03581*.

[8] A. Borji, D. N. Sihite, and L. Itti, "What stands out in a scene? A study of human explicit saliency judgment," *Vis. Res.*, vol. 91, pp. 62–77, Oct. 2013.

[9] N. Bruce and J. Tsotsos, "Attention based on information maximization," *J. Vis.*, vol. 7, p. 950, Jun. 2010.

[10] N. D. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos, "On computational modeling of visual saliency: Examining what's right, and what's left," *Vis. Res.*, vol. 116, pp. 95–112, Nov. 2015.

[11] A. Bruckert, H. Tavakoli, Z. Liu, M. Christie, and O. Meur, "Deep saliency models: The quest for the loss function," *Neurocomputing*, vol. 453, pp. 693–704, Oct. 2021.

[12] A. Bruno, L. Greco, and M. L. Cascia, "Video object recognition and modeling by sift matching optimization," in *Proc. ICPRAM*, 2014, pp. 662–670.

[13] A. Bruno, F. Gugliuzza, E. Ardizzone, C. C. Giunta, and R. Pirrone, "Image content enhancement through salient regions segmentation for people with color vision deficiencies," *I-Perception*, vol. 10, no. 3, 2019, Art. no. 2041669519841073.

[14] A. Bruno, F. Gugliuzza, R. Pirrone, and E. Ardizzone, "A multi-scale colour and keypoint density-based approach for visual saliency detection," *IEEE Access*, vol. 8, pp. 121330–121343, 2020.

[15] A. Brychtova and A. Coltekin, "Discriminating classes of sequential and qualitative colour schemes," *Int. J. Cartography*, vol. 1, no. 1, pp. 62–78, Jan. 2015.

[16] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vis. Res.*, vol. 116, pp. 165–178, Nov. 2015.

[17] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Oct. 2019.

[18] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[19] A. Chetouani, "Convolutional neural network and saliency selection for blind image quality assessment," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2835–2839.

[20] A. Chetouani and L. Li, "On the use of a scanpath predictor and convolutional neural network for blind image quality assessment," *Signal Process., Image Commun.*, vol. 89, p. 115963, 2020.

[21] A. Chetouani, M. A. Qureshi, M. Deriche, and A. Beghdadi, "A novel ranking algorithm of enhanced images using a convolutional neural network and a saliency-based patch selection scheme," in *Proc. 11th Int. Conf. Quality Multimedia Exp. (QoMEX)*, Jun. 2019, pp. 1–6.

[22] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biol.*, vol. 14, no. 19, pp. R850–R852, 2004.

[23] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3488–3493.

[24] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying more attention to saliency: Image captioning with saliency and context attention," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 2, pp. 1–21, Apr. 2018.

[25] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2016.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Feb. 2009, pp. 248–255.

[27] Y. Dahou, M. Tliba, K. McGuinness, and N. O'Connor, "ATSal: An attention based architecture for saliency prediction in 360 videos," 2020, *arXiv:2011.10600*.

[28] J. Donahue, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 32. Jun. 2014, pp. 647–655.

[29] M. Dorr, R. Karl Gegenfurtner, and E. Barth, "The contribution of low-level features at the centre of gaze to saccade target selection," *Vis. Res.*, vol. 49, no. 24, pp. 2918–2926, 2009.

[30] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vis.*, vol. 8, no. 14, p. 18, 2008.

[31] W. Elloumi, K. Guissous, A. Chetouani, and S. Treuillet, "Improving a vision indoor localization system by a saliency-guided detection," in *Proc. IEEE Vis. Commun. Image Process. Conf.*, Dec. 2014, pp. 149–152.

[32] W. Feng, H. Sui, J. Tu, W. Huang, and K. Sun, "A novel change detection approach based on visual saliency and random forest from multi-temporal high-resolution remote-sensing images," *Int. J. Remote Sens.*, vol. 39, no. 22, pp. 7998–8021, Nov. 2018.

[33] Y. Gao, M. Shi, D. Tao, and C. Xu, "Database saliency for fast image retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 359–369, Mar. 2015.

[34] A. Garcia-Diaz, V. Leborán, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *J. Vis.*, vol. 12, no. 6, p. 17, 2012.

[35] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.

[36] M. Hamidi, A. Chetouani, M. E. Haziti, M. E. Hassouni, and H. Cherifi, "Blind robust 3-D mesh watermarking based on mesh saliency and QIM quantization for copyright protection," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.*, A. Morales, J. Fierrez, J. Salvador Sánchez, and B. Ribeiro, Eds. Cham, Switzerland: Springer, 2019, pp. 170–181.

[37] S. Han and N. Vasconcelos, "Image compression using object-based regions of interest," in *Proc. Int. Conf. Image Process.*, 2006, pp. 3097–3100.

[38] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NIPS*, 2006, pp. 545–552.

[39] J. M. Henderson and A. Hollingworth, "The effects of semantic consistency on eye movements during complex scene viewing," *J. Exp. Psychol., Hum. Perception Perform.*, vol. 25, no. 1, pp. 210–228, 1999.

[40] F. Hongwen, C. Nian, Z. Jingwen, B. Youfang, L. Jian, and W. Han, "Automatic detection of ultrasound breast lesions: A novel saliency detection model based on multiple priors," *Signal, Image Video Process.*, vol. 4, pp. 1–12, Oct. 2021.

[41] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 262–270.

[42] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[43] Y. Jin-Gang, X. Gui-Song, G. Changxin, and S. Ashok, "A computational model for object-based visual saliency: Spreading attention along gestalt cues," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 273–286, Oct. 2015.

[44] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," MIT, Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2012-001, 2012.

[45] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.

[46] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 914–921.

[47] M. Pinsk and S. Kastner, "Visual attention as a multilevel selection process," *Cognit., Affect., Behav. Neurosci.*, vol. 4, no. 4, pp. 483–500, 2004.

[48] M. A. Kerkouri, M. Tliba, A. Chetouani, and R. Harba, "SALYPATH: A deep-based architecture for visual attention prediction," *CoRR*, abs/2107.00559, pp. 1–45, Oct. 2021.

[49] P. Diederik Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, abs/1412.6980, pp. 1–15, Dec. 2015.

[50] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Hum. Neurobiol.*, vol. 4, pp. 219–227, Dec. 1985.

[51] I. Kotseruba, C. Wloka, A. Rasouli, and K. John Tsotsos, "Do saliency models detect odd-one-out targets? New datasets and evaluations," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2019, pp. 1–14.

[52] G. KR, "The interaction between vision and eye movements," *Perception*, vol. 12, no. 45, pp. 1333–1357, 2016.

[53] K. Krejtz, A. Coltekin, A. Duchowski, and A. Niedzielska, "Using coefficient K to distinguish ambient/focal visual attention during map viewing," *J. Eye Movement Res.*, vol. 10, no. 2, Apr. 2017.

[54] A. Krizhevsky, I. Sutskever, and E. Geoffrey Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, Red Hook, NY, USA, 2012, pp. 1097–1105.

[55] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet," *CoRR*, abs/1411.1045, pp. 1–12, Nov. 2015.

[56] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," in *Proc. Conf. 15th Eur. Conf. Comput. Vis. (ECCV)*, in (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11220, Y. Weiss, V. Ferrari, C. Sminchisescu, and M. Hebert, Eds. Munich, Germany: Springer-Verlag, Sep. 2018, pp. 798–814. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85055114685&doi=10.1007%2f978-3-030-01270-0_47&partnerID=40&md5=b38a0b0cbfd3123c6f54d66a6a2d5fd2, doi: 10.1007/978-3-030-01270-0_47.

[57] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "DeepGaze II: Reading fixations from deep features trained on object recognition," 2016, *arXiv:1610.01563*.

[58] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, p. 32, Dec. 2008.

[59] R. L. M. O. Cozot and P. T. Le, "Can we accurately predict where we look at paintings? *PLoS ONE*, vol. 15, no. 10, 2020, Art. no. e0239980.

[60] J. Li and W. Gao, *Visual Saliency Computation: A Machine Learning Perspective* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8408. Springer-Verlag, 2014, pp. 1–249. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84904736815&doi=10.1007%2f978-3-319-05642-5&partnerID=40&md5=fe7b18c66cc2e05191487722fc73edb3, doi: 10.1007/978-3-319-05642-5.

[61] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3264–3274, Jul. 2018.

[62] A. K. Mackenzie and J. M. Harris, "A link between attentional function, effective eye movements, and driving ability," *J. Exp. Psychol., Hum. Perception Perform.*, vol. 43, no. 2, pp. 381–394, Feb. 2017.

[63] A. Mahdi, K. Nader, and S. Shadrokh, "Context-aware saliency detection for image retargeting using convolutional neural networks," *Multimedia Tools Appl.*, vol. 80, no. 8, pp. 11917–11941, 2021.

[64] O. Messai, A. Chetouani, F. Hachouf, and Z. A. Seghir, "Deep quality evaluator guided by 3d saliency for stereoscopic images," *Electron. Imag., Hum. Vis. Electron. Imag.*, vol. 11, pp. 110–117, Jan. 2021.

[65] N. Bigdely-Shamlo, A. Vankov, R. R. Ramirez, and S. Makeig, "Brain activity-based image classification from rapid serial visual presentation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 5, pp. 432–441, Oct. 2008.

[66] T. Oyama and T. Yamanaka, "Influence of image classification accuracy on saliency map estimation," *CAAI Trans. Intell. Technol.*, vol. 3, no. 3, pp. 140–152, Sep. 2018.

[67] J. Pan, C. Canton Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-I-Nieto, "SalGAN: Visual saliency prediction with generative adversarial networks," 2017, *arXiv:1701.01081*.

[68] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. OConnor, "Shallow and deep convolutional networks for saliency prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 598–606.

[69] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 8, pp. 2397–2416, 2005.

[70] M. I. Posner, "Orienting of attention," *Quart. J. Exp. Psychol.*, vol. 32, no. 1, pp. 3–25, Feb. 1980.

[71] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit, "RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Process., Image Commun.*, vol. 28, no. 6, pp. 642–658, Jul. 2013.

[72] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.

[73] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artif. Intell.*, vol. 146, no. 1, pp. 77–123, 2003.

[74] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[75] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.

[76] J. S. W. Wang, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[77] H. Zhang, I. Goodfellow, N. D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. ICML*, 2019, pp. 7345–7363.

MAROUANE TLIBA received the Engineering degree in computer networking and the M.S. degree in advanced telecommunication systems from the Institut National Des Télécommunications et des Technologies de l'Information et de la Communication (INTTIC), Oran, Algeria, in 2020 and 2021, respectively. He is currently pursuing the Ph.D. degree in computer vision with Orleans University, France. His thesis project is about assessing the perceptual quality of 3-D scenes. His research interests include deep learning and computer vision.

MOHAMED A. KERKOURI received the Engineering degree in computer networking from the Institut National Des Télécommunications et des Technologies de l'Information et de la Communication (INTTIC), Oran, Algeria. He is currently pursuing the Ph.D. degree in computer vision with Orleans University, France, more particularly his thesis works revolved around eye-movement prediction on painting images. He worked as a Teaching Assistant with Orleans University. His research interests include machine learning and computer vision.

BASHIR GHARIBA received the B.Sc. degree in electrical and computer engineering from Elmergib University, Khoms, Libya, in 1998, and the M.Sc. degree from Libyan Academy, in April 2010. He became a Lecturer with the Faculty of Engineering, Elmergib University, in December 2011. Recently, he defended his Ph.D. thesis in computer engineering with the Memorial University of Newfoundland, St. John's, Canada. His research interests include image processing, computer vision, and deep learning.

ALADINE CHETOUANI (Member, IEEE) received the master's degree in computer science from University Pierre and Marie Curie, France, in 2005, the Ph.D. degree in image processing from the University of Paris 13, France, in 2010, and the Habilitation degree from the Université d'Orléans, titled "On the use of visual attention and deep learning for blind quality assessment of multimedia contents," in 2020. From 2010 to 2011, he was a Postdoctoral Researcher with the L2TI Laboratory, Paris 13 University. In 2020, he benefited from a CNRS Delegation Year with the L2S Laboratory, Centrale Supélec, Université Paris Saclay, France. He is currently an Associate Professor with the Laboratory PRISME, Orleans, France. He led and participated in several research projects. He supervised more than 20 students (Ph.D. and master's). He is the coauthor of more than 100 research publications in international refereed journals and conference proceedings. He served in several program committees. His current research interests include image quality, perceptual analysis, and visual attention and for cultural heritage using deep learning models for different multimedia content (image, stereo, and 3-D). He serves as a reviewer for major conferences and journals in the field of image analysis and pattern recognition. He co-edited different special issues in international journals and organized different special sessions in international conferences (IEEE ICIP and IEEE ICME).

ARZU ÇÖLTEKIN works with the Institute for Interactive Technologies, University of Applied Sciences and Arts Northwestern Switzerland, as a Professor in human–computer interaction and extended reality. She is also a Research Affiliate with the Harvard-Smithsonian Center for Astrophysics, Seamless Astronomy Group, Harvard University, Cambridge, USA, chairs the International Geovisualization, Augmented and Virtual Reality Working Group within the ISPRS, the Co-Chair of Commission on Visual Analytics within the ICA, and a Council Member with the International Society of Digital Earth. Her interdisciplinary work covers topics related to information science, visual analytics, visualization and cartography, virtual/augmented reality, gaze-contingent displays, eye-tracking, vision (perception and cognition), and human–computer interaction.

MOHAMED SHEHATA (Senior Member, IEEE) received the B.Sc. degree (Hons.) and the M.Sc. degree in computer engineering from Zagazig University, Egypt, in 1996 and 2001, respectively, and the Ph.D. degree from the University of Calgary, Canada, in 2005. Following his Ph.D. degree, he worked as a Postdoctoral Fellow with the University of Calgary, and after that, he joined Intelliview Technologies, Inc., as the Vice-President of Engineering and Research. He joined Memorial University as an Assistant Professor, in January 2013, and then as an Associate Professor, in January 2019. In August 2019, he joined the Department of Computer Science, Math, Physics, and Statistics, The University of British Columbia, and became an Adjunct Professor with the Department of Computer Engineering, Memorial University. He served as the IEEE Newfoundland Section Chair, from 2017 to 2019. He is also the Editor-in-Chief of the IEEE Canadian Journal of Electrical and Computer Engineering.

ALESSANDRO BRUNO received the master's degree in computer engineering, in 2008, and the Ph.D. degree in computer engineering from DINFO, Palermo University. He is currently a Lecturer in computing with the Department of Computing and Informatics, Bournemouth University. In April 2012, was a Postdoctoral Research Fellow with Palermo University, focusing on image forensics, object recognition and visual perception. He worked at INAF IASF Palermo (Italian National Institute for Astrophysics), focusing his efforts on remote sensing applications and cosmic rays analysis with deep learning techniques. Furthermore, he was a Research Visitor at Mullard Space Science Laboratory (MSSL), University College London (UCL). He worked also as a Postdoctoral Research Fellow in computer vision with Bournemouth University. His research interests include computer vision, artificial intelligence, and image analysis. He has mostly dealt with visual attention and visual saliency, biomedical imaging, crowd behavior analysis, image and video forensics, remote sensing, and human–computer interaction.