Neurocomputing 494 (2022) 455–467

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

TranSalNet: Towards perceptually relevant visual saliency prediction

Jianxun Lou [a], Hanhe Lin [b], David Marshall [a], Dietmar Saupe [c], Hantao Liu [a,*]

[a] School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AX, United Kingdom
[b] National Subsea Centre, Robert Gordon University, Aberdeen AB107AQ, United Kingdom
[c] Department of Computer and Information Science, University of Konstanz, Konstanz 78464, Germany

ARTICLE INFO

Article history:
Received 16 January 2022
Revised 22 March 2022
Accepted 17 April 2022
Available online 21 April 2022
Communicated by Zidong Wang

Keywords:
Saliency prediction
Deep learning
Transformer
Convolutional neural network

ABSTRACT

Convolutional neural networks (CNNs) have significantly advanced computational modelling for saliency prediction. However, accurately simulating the mechanisms of visual attention in the human cortex remains an academic challenge. It is critical to integrate properties of human vision into the design of CNN architectures, leading to perceptually more relevant saliency prediction. Due to the inherent inductive biases of CNN architectures, there is a lack of sufficient long-range contextual encoding capacity. This hinders CNN-based saliency models from capturing properties that emulate viewing behaviour of humans. Transformers have shown great potential in encoding long-range information by leveraging the self-attention mechanism. In this paper, we propose a novel saliency model that integrates transformer components to CNNs to capture the long-range contextual visual information. Experimental results show that the transformers provide added value to saliency prediction, enhancing its perceptual relevance in the performance. Our proposed saliency model using transformers has achieved superior results on public benchmarks and competitions for saliency prediction models.

The source code of our proposed saliency model TranSalNet is available at: https://github.com/LJOVO/TranSalNet.

1. Introduction

Visual attention represents an important mechanism of the human visual system (HVS), which allows humans to select and interpret the most relevant information in the visual scene [1]. Simulating visual attention in the form of an algorithm is regarded as visual saliency prediction. Being able to automatically predict saliency is beneficial to many research fields including computer vision, robotics, healthcare, and multimedia [2–8].

Existing saliency prediction models can be categorised into two types, traditional and deep learning-based models. Traditional models [9–12] apply low-level visual features such as colour, luminance, texture, and contrast, to simulate the visually salient areas in the scene. These models remain rather limited as higher-level features such as objects are often omitted; but these features exhibit significant determinants of visual saliency [13,14]. Although some traditional models [15] have been extended with specific higher-level features, e.g., faces and texts, there are still obstacles in combining low-level and higher-level visual features. Rather than designing handcrafted features, deep learning-based saliency models automatically discover representations from images [16–24]. These methods typically use convolutional neural networks (CNNs) to construct feature encoders and decoders to generate visual saliency maps. Deep learning-based visual saliency models have achieved remarkable success, mainly due to the availability of well-established deep CNNs [25–28] and large-scale datasets relevant to human visual attention [29]. Fig. 1 illustrates examples of visual saliency prediction using both traditional and deep learning-based models, and the correspondences between the ground truth (i.e., where humans look in an image) and prediction (i.e., output of a computational saliency model).

Accurately predicting saliency as perceived by humans remains an academic challenge. One way to improve the reliability of saliency prediction is to incorporate the properties of the HVS in the construction of computational models [9,30]. Despite the significant progress made by the deep learning-based models, each convolution kernel in CNNs only receives information from a local subset of pixels in an image, which makes fully CNN-based models deficient in obtaining long-range contextual information. When humans view an image, foveal vision provides the highest resolution visual information, but in the meantime peripheral vision still provides the HVS with non-detailed but long-range visual information [31–33]. In other words, the HVS uses the long-range informa-

* Corresponding author.
E-mail address: LiuH35@cardiff.ac.uk (H. Liu).

https://doi.org/10.1016/j.neucom.2022.04.080
0925-2312/

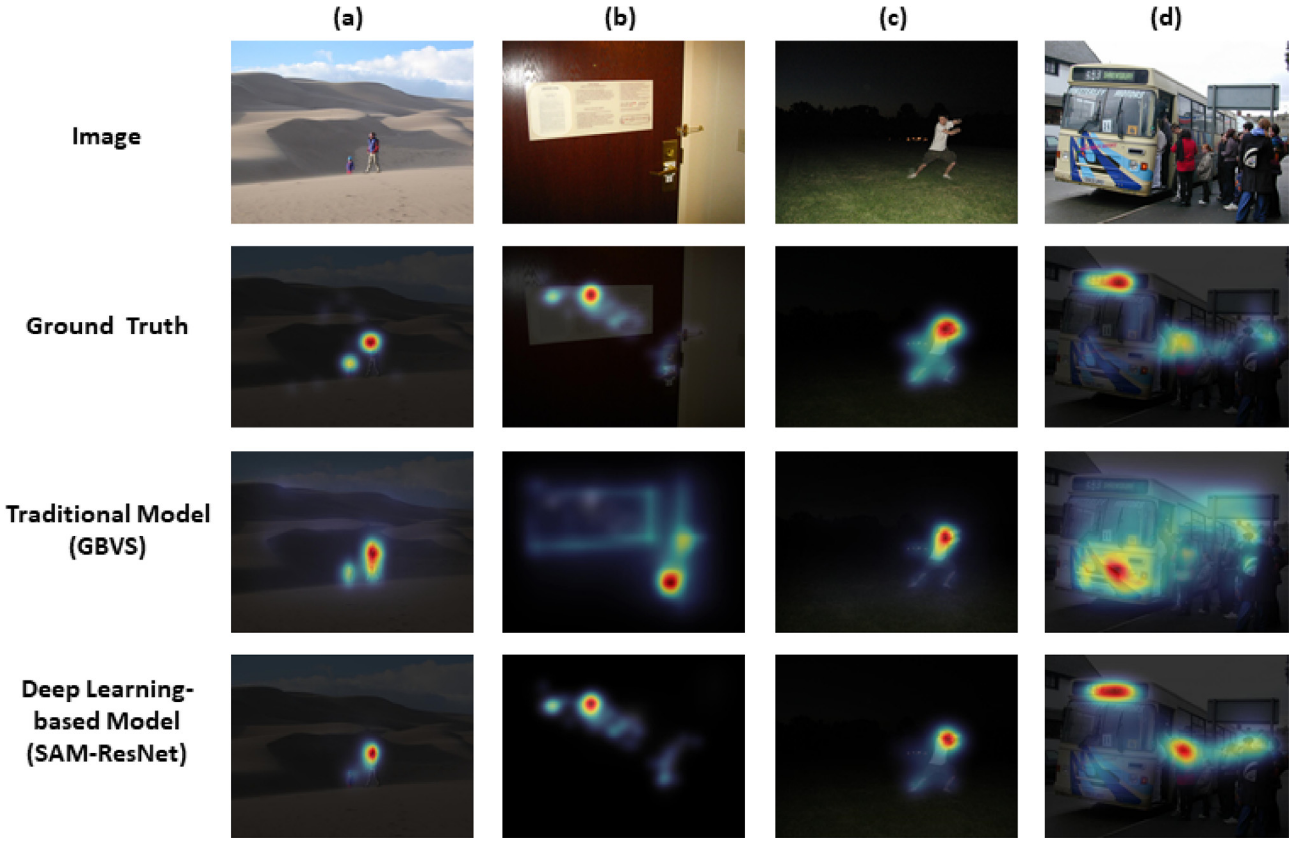


Fig. 1. Examples of visual saliency prediction. The first row shows the images that stimulate the human eye to view freely. The so-called "Ground Truth" in the second line refers to the fixation density maps, also called saliency maps, generated from the human fixation location. The third and fourth rows show the prediction results of the traditional (GBVS [10]) and deep learning-based (SAM-ResNet [18]) saliency models, respectively. Image (a) and (b) are from MIT1003 dataset [61]; (c) and (d) are from SALICON dataset [29]. It can be seen that both traditional and deep learning-based models are capable of capturing human viewing behaviour, but the deep learning-based model provides better results in demanding scenes, such as (b) and (d) to a considerable extent.

tion of an image to modulate the local maxima of saliency in the visual field [30,34]. Therefore, the ground truth saliency map represents the perceptual spatial interactions of local and non-local (i.e., long-range) information. This HVS property could be beneficial for predicting visual saliency in a perceptually more relevant manner so that the machine generated saliency map can faithfully reflect human perception. Previous studies mainly attempted to solve this problem through two approaches. One approach is capturing multi-scale information through the CNNs [35,22,19,36,17], which introduces image representations with different granularities to some extent. This approach may not provide the optimal solution as it still lacks the ability to model the way visual information is perceived by the HVS, e.g., some studies have used multi-scale images or multi-scale representations to improve saliency prediction [35,22,19,36,17], but challenges remains for models in optimally fusing multi-scale information to mimic the functionality of the HVS. Another approach is adding long-range modelling capabilities to network structures to increase spatial representations. By using a Long-Short Term Memory (LSTM)-based architecture [18,37], this approach has proven effective in handling local and long-range visual information thus refining the accuracy of saliency prediction. Although these studies have demonstrated promising outcomes, much work is needed to close the gap between saliency prediction and human perception. The transformer [38], which consists of a self-attention mechanism, provides an elegant solution to process long-range information. By effectively modelling long-range dependency, the transformer has proven efficacy in the field of natural language processing [39] and more recently achieved promising results in computer vision tasks [40,41]. However, the use of transformers in visual saliency prediction has not been fully explored until now.

To address the above-mentioned challenges and to build a human-like saliency model, we propose a novel saliency prediction model called TranSalNet, which integrates transformers into a CNN-based architecture. Transformer encoders can learn spatially long-range dependencies by using a self-attentive mechanism, resulting in a perceptually more relevant saliency representation. To the best of our knowledge, this is the first study to explore the combination of CNNs and transformers to enhance saliency prediction. Also, we demonstrate the benefits of transformer components in saliency prediction. Our model achieves superior performance not only on the MIT300 benchmark (the most widely recognised dataset for saliency benchmark) but also on the SALICON Saliency Prediction Challenge (the largest dataset available for saliency prediction).

2. Related work

We contribute towards a perceptually more relevant saliency prediction method using deep learning models with transformers. This section provides a comprehensive review on deep learning-based saliency prediction models, methods for evaluating saliency models (especially evaluating the perceptual relevance of saliency prediction), transformer applications in vision tasks, and multi-scale and long-range information in visual saliency prediction.

2.1. Deep learning-based visual saliency prediction

A number of deep learning-based visual saliency prediction models have been proposed in recent years. The *Ensembles of Deep Networks* (eDN) [42] represents one of the first models that adopts shallow CNNs to detect the visual saliency of natural images. The saliency features are extracted by CNNs and combined by a linear classifier to create saliency maps. Since then, with the application of deep neural networks and large-scale saliency datasets, deep learning-based saliency prediction has achieved further remarkable successes. *DeepGaze* and *DeepGaze II* [23], which are based on AlexNet [25] and VGGNet [26], respectively, successfully build pre-trained networks as feature extractors to train deeper networks for saliency prediction. By comparing VGGNet, AlexNet, and GoogleNet [43], Huang et al. [35] found that VGGNet detects saliency more effectively than the other two models. Many visual saliency prediction models based on VGGNet have since been proposed [16,17,19]. *EML-NET* [20] focuses on exploring the use of more sophisticated feature extractors (i.e., a parallel two-stream CNN-based encoder) to enhance the performance of saliency prediction. By comparing ResNet-50 with DenseNet [28] and NASNet [44], it is argued that in the field of saliency prediction, the widely used ResNet-50 could still be "shallow" for the large-scale saliency datasets, such as SALICON. Similarly, *DeepGaze II-E* discusses the contribution of different backbones to saliency prediction. It is found that appropriately concatenating multiple backbones pre-trained on ImageNet [45] is effective in improving the performance of saliency models.

In addition to the efforts mentioned above, there are several studies that adopt multi-scale or long-range information to improve visual saliency prediction. We discuss this issue below in Section 2.4.

2.2. Evaluation methods for saliency models

A number of metrics have been proposed to measure the agreement between the predicted saliency map and the ground truth produced by human eye movements. By investigating commonly used metrics, Bylinskii et al. [46] found that under general assumptions the *Linear Correlation Coefficient* (CC) and *Normalized Scanpath Saliency* (NSS) metrics could be used as representative metrics for benchmarking saliency models. More importantly, they also suggested different evaluation metrics should be used for different applications, for example, metrics that are more appropriate for evaluating the capability of salient object detection may not be necessarily useful for the evaluation of saliency prediction of other vision applications [46,47]. Li et al. [48] found that only a limited number of evaluation metrics, i.e., NSS, CC, and *Similarity* (SIM) are in close agreement with human judgements through a large-scale subjective experiment. Similarly, Yang et al. [47] found that CC and SIM are the most in line with human evaluation of saliency maps. Kummerer et al. [49] also demonstrated that it is difficult for a saliency model to perform equally well on all popular saliency evaluation metrics. They proposed a novel approach that allows a saliency model to generate different "saliency maps" according to the characteristics and behaviours of different metrics; and the model that adopts this evaluation method is referred to as a "probabilistic model." As a distinction, without targeting any specific evaluation metric, a saliency model that generates a single saliency map for a given image is referred to as a "classical model". Since our aim is to generate a single saliency map for each image that can faithfully reflect human perception, we evaluate models in the "classical model" framework. In evaluating models, we apply all commonly used evaluation metrics to quantify model performance, but make a clear distinction between "perception-based metrics (i.e., NSS, CC, and SIM)" and "non-perception-based metrics", as defined in [46]. By doing so, the perceptual relevance of the predicted saliency maps can be appropriately measured.

2.3. Transformer in visual tasks

The transformer was first introduced to the tasks of natural language processing (NLP) [38]. Because of its powerful long-range dependency modelling capabilities, the transformer has achieved remarkable success in the field of NLP. Consequently, a number of studies in the field of computer vision are also exploring the effectiveness of the use of transformer.

The *vision transformer* is one of the first pure transformer architectures for image processing, which uses a vanilla version of the transformer to form a network that achieves performance comparable to that of state-of-the-art CNN-based models. After this work, several models, such as *DeepViT* [50] and *Swin Transformer* [51], have achieved further success in visual tasks by using the transformer.

Currently, the transformer has also demonstrated excellent performance in the field of salient object detection [52], which is related to the current work, even though it is a substantially different task [53]. Salient object detection aims to segment salient objects from an image and generate a binary map [54]. However, in visual saliency prediction, the aim is to predict the density map of human fixations (i.e., the spatial deployment of visual attention).

In summary, the previous studies have shown the powerful representation capabilities of the transformer, particularly for capturing long-range information, which could have potential contributions to predicting gaze. However, the use of transformers in visual saliency prediction has not been fully explored until now. In this paper, we will investigate the benefits as well as application of transformer components in saliency prediction.

2.4. Multi-scale and long-range information in visual saliency prediction

By using multi-scale image representations to simulate different perceptual scales, successful results have been achieved in vision tasks such as image segmentation [55], human pose estimation [56], and salient object detection [57]. In the filed of visual saliency prediction, Huang et al. [35] and Fan et al. [22] proposed CNN-based models that extract multi-scale features from images of different resolutions separately and concatenate the results to obtain salient semantic objects with different granularities hence to optimise saliency prediction. In order to obtain multi-scale contextual information, *Deep Visual Attention* (DVA) [17] constructs three decoders of different granularities to generate multi-scale saliency estimates for saliency prediction. *EML-NET* [20] also uses multi-scale feature maps from encoder networks to obtain holistic scene features for saliency prediction. *MSI-Net* [19] adopts convolutional layers with different dilation rates to augment multi-scale information for saliency prediction. *GazeGAN* [36] is a generative adversarial network for saliency prediction, which uses a modified U-Net [55] with multi-scale information by using skip-connections to construct its generator. *UNISAL* [21] adopts skip-connections to provide the decoder network with multi-scale features. These studies have demonstrated that multi-scale information is beneficial to visual saliency prediction.

Similarly to other vision tasks [52,58], visual saliency prediction has also benefitted from neural networks with long-range modelling capabilities to simulate the spatial attentional mechanisms. *DSCLSTM* [37] extracts local feature maps by using CNNs first, and then incorporates non-local scene contexts into the local feature maps by using LSTM-based components to predict human eye fixation points in natural scenes. Cornia et al. [18] developed visual saliency models that integrate an LSTM module into the CNN-based network to simulate explicit properties of the human attention mechanism. Similarly, Fang et al. [59] used LSTM to obtain pseudo sequential information to simulate the human visual attention shift. These studies suggest that modelling the relevant dependence between spatial information can refine the saliency prediction models.

In this paper, we combine these two strategies. More specifically, we integrate transformer encoders into a CNN-based architecture to provide multi-scale image representations with enhanced long-range contextual information, resulting in perceptually more relevant visual saliency prediction.

3. The proposed model

The schematic overview of our proposed TranSalNet model is shown in Fig. 2. Firstly, a given image is fed into a CNN encoder. In order to obtain multi-scale image representations, three sets of feature maps with different spatial sizes are extracted from the CNN encoder. Due to the inherent inductive biases of CNN encoder architectures, the extracted image representations lack long-range contextual information, which potentially makes a saliency model less humanlike (note the human visual system is proficient in capturing both local and long-range visual information). Therefore, to obtain perceptually more relevant visual saliency prediction, these feature maps are passed through three transformer encoders, yielding long-range context-enhanced feature maps. Then the CNN decoder fuses these feature maps for saliency prediction.

3.1. The CNN encoder

Previous research has shown that the use of CNN-based networks to extract features for saliency prediction is effective. Likewise, we used a CNN encoder as the feature extractor in this study.

The CNN models used in this study were initially constructed for image classification. In order to provide image feature maps to the downstream networks, the fully connected layer at the end of these CNNs is removed to form a viable CNN encoder. We extract feature maps with three sets of different spatial sizes from the CNN encoder. Given an input image with size $w \times h \times 3$, the spatial dimensions of the extracted feature maps are $\frac{w}{8} \times \frac{h}{8}$, $\frac{w}{16} \times \frac{h}{16}$, and $\frac{w}{32} \times \frac{h}{32}$, respectively.

In this study, two feature extraction networks are adopted to construct two versions of TranSalNet models. One version uses ResNet-50 [27] as an encoder, which is a feature extraction network widely used in saliency prediction. This version of the model is referred to as *TranSalNet_Res*. The CNN body of ResNet-50 is composed of five convolutional blocks that are denoted as conv1 and conv2_x to conv5_x. We extract feature maps from the deeper conv3_x, conv4_x, and conv5_x blocks. However, [20] suggests that ResNet-50 itself as an encoder is probably relatively "shallow." Therefore, we use DenseNet-161 [28], which has higher performance on the ImageNet benchmark, as the CNN encoder to build another version referred to as *TranSalNet_Dense*. For DenseNet-161, it mainly consists of four "Dense Blocks" denoted as DenseBlock 1 to 4. We extract feature maps from the deeper DenseBlock 2, DenseBlock 3, and DenseBlock 4.

Although previous work [35,16,17,19,36] showed that adopting multi-scale feature maps is beneficial to saliency prediction, our experiments found that using feature maps from shallower network blocks, i.e. the conv1 and conv2_x, may cause undesired artefacts to appear in the saliency maps. Therefore, we exclude feature maps from the shallower network blocks.

3.2. The transformer encoder

The three sets of multi-scale feature maps are respectively fed into three transformer encoders to enhance the long-range and contextual information. The details of transformer are depicted at the bottom of Fig. 2. Let $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ be the feature maps that have spatial dimensions of $\frac{w}{32} \times \frac{h}{32}$, $\frac{w}{16} \times \frac{h}{16}$, and $\frac{w}{8} \times \frac{h}{8}$, respectively, first, a 1×1 convolution layer ($\mathrm{Conv}_{1\times1}$) is used to reduce the computational cost and align with the acceptable input size of the transformer encoder. More specifically, both $\mathbf{x}_1$ and $\mathbf{x}_2$ are reduced to 768 dimensions, and $\mathbf{x}_3$ changed to 512 dimensions. Following this, as there is no relative or absolute position information in the feature maps, it is necessary to utilise position embedding (POS) to enable position-awareness before feeding the input into the transformer encoders. Therefore, the absolute POS [40] is implemented before feeding input into the transformer encoders, which per-

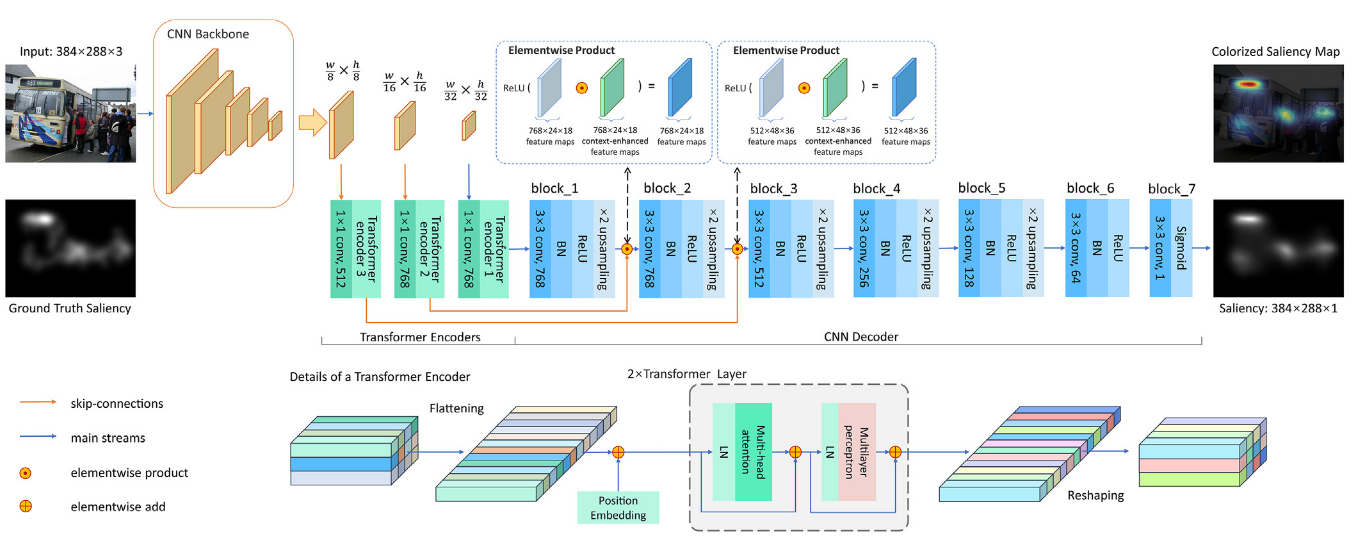


Fig. 2. Schematic overview of TranSalNet. Assume that the spatial size of inputs is $w \times h$. After the input image is processed by the CNN encoder, which provides three sets of multi-scale feature maps have spatial size of $\frac{w}{8} \times \frac{h}{8}$, $\frac{w}{16} \times \frac{h}{16}$, and $\frac{w}{32} \times \frac{h}{32}$, respectively. Then the contextual information of these feature maps is enhanced by transformer encoders. The predicted saliency map is generated by the CNN decoder, which uses skip-connection (orange arrows) and element-wise production to fuse multi-scale context-enhance feature maps. The illustration of the transformer encoder is shown below the architecture diagram, which consists of standard Multi-head Self-Attention (MSA) and Multi-layer Perceptron (MLP) blocks.

forms an element-wise addition to the input and a learnable matrix with the same shape as the input. Each transformer encoder contains two same layers of standard Multi-head Self-Attention (MSA) and Multi-layer Perceptron (MLP) blocks [40]. In our model, we apply 12-heads attention in transformer encoder 1 and 2, and 8-heads in encoder 3. The MLP block contains two layers with a GELU activation function. Besides, Layer Normalization (LN) and residual connection are applied before and after each block respectively. The processing in each transformer encoder can be represented as:

$$\mathbf{z}_0 = \mathrm{Conv}_{1\times 1}(\mathbf{x}_i) \oplus \mathrm{POS}_i, \quad i = 1,2,3 \tag{1}$$

$$\mathbf{z}'_l = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{l-1})) \oplus \mathbf{z}_{l-1}, \quad l = 1,2 \tag{2}$$

$$\mathbf{z}_l = \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_l)) \oplus \mathbf{z}'_l, \quad l = 1,2 \tag{3}$$

where $\mathbf{z}_l$ is the output feature maps of the l-th layer in transformer encoder, and $\mathbf{x}_i$ is the input feature maps from the CNN encoder. The feature maps that are passed through transformer encoder 1, 2, and 3 are context-enhanced and denoted as $\mathbf{x}_1^c, \mathbf{x}_2^c$, and $\mathbf{x}_3^c$ respectively.

3.3. The CNN decoder

A CNN decoder is used to fuse the long-range context-enhanced feature maps from the transformer encoders and restore the original image resolution. The CNN decoder is a fully CNN network containing block_1 to block_7, which is used to implement pixel-level classification to predict saliency maps. Batch normalization (BN) and the activation function (ReLU for block_1 to block_6; Sigmoid for block_7) are applied after each 3×3 convolution operation ($\mathrm{Conv}_{3\times 3}$), where the former is used to promote the convergence and the latter is used to increase the nonlinear factor of the model. Since the input image is 32-scale downsampled by the encoder network, a 2-scale upsampling that adopts nearest-neighbor interpolation is performed to the feature map in block_1 to block_5 to obtain a saliency map of the same size as the input. In order to enhance the long-range and multi-scale context of the feature map during the decoding process, the upsampled feature map and the transformer's output from the corresponding skip-connection are fused by an element-wise product operation. The processes from block_1 to block_6 can be expressed as:

$$\mathbf{x}_i^f = \begin{cases} \mathbf{x}_i^c, & i = 1 \quad (4) \\ \mathrm{ReLU}(\mathrm{Upsample}(\hat{\mathbf{x}}_{i-1}^f) \odot \mathbf{x}_i^c), & i = 2,3 \quad (5) \\ \mathrm{Upsample}(\hat{\mathbf{x}}_{i-1}^f), & i = 4,5,6 \quad (6) \end{cases}$$

$$\hat{\mathbf{x}}_i^f = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{3\times 3}(\mathbf{x}_i^f))), \quad i = 1,2,\ldots,6 \tag{7}$$

where $\mathbf{x}_i^f$ and $\hat{\mathbf{x}}_i^f$ are the input and output features of the i-th block. The output block, i.e., block_7, is used to reduce the dimensionality of the feature maps to a 2D map for pixel-level classification. Therefore, the sigmoid activation function is applied to the feature map:

$$\hat{\mathbf{y}} = \mathrm{sigmoid}(\mathrm{Conv}_{3\times 3}(\hat{\mathbf{x}}_6^f)), \tag{8}$$

where $\hat{\mathbf{y}}$ is the predicted saliency map.

3.4. Loss function

Recent saliency prediction studies [18,20,36] have shown that taking advantage of the saliency evaluation metrics to define the loss function can significantly improve the performance of saliency prediction models.

Following a similar idea, we adopt a linear combination of four metrics as the loss function to train our model, including the Normalized Scanpath Saliency (NSS), Kullback–Leibler divergence (KLD), Linear Correlation Coefficient (CC), and Similarity (SIM). Let $\mathbf{y}^s, \mathbf{y}^f$, and $\hat{\mathbf{y}}$ be the ground truth saliency map, fixation map, and predicted saliency map, and i indicates the ith pixel of $\mathbf{y}^s$ and $\hat{\mathbf{y}}$, our loss function is defined as:

$$L(\mathbf{y}^s, \mathbf{y}^f, \hat{\mathbf{y}}) = \lambda_1 L_{\mathrm{NSS}}(\mathbf{y}^f, \hat{\mathbf{y}}) + \lambda_2 L_{\mathrm{KLD}}(\mathbf{y}^s, \hat{\mathbf{y}}) + \lambda_3 L_{\mathrm{CC}}(\mathbf{y}^s, \hat{\mathbf{y}}) + \lambda_4 L_{\mathrm{SIM}}(\mathbf{y}^s, \hat{\mathbf{y}}), \tag{9}$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are the weights of each metric, and

$$L_{\mathrm{NSS}}(\mathbf{y}^f, \hat{\mathbf{y}}) = \frac{1}{\sum_i \mathbf{y}_i^f} \sum_i \frac{\hat{\mathbf{y}}_i - \mu(\hat{\mathbf{y}})}{\sigma(\hat{\mathbf{y}})} \mathbf{y}_i^f, \tag{10}$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ stand for standard deviation and mean respectively;

$$L_{\mathrm{KLD}}(\mathbf{y}^s, \hat{\mathbf{y}}) = \sum_i \mathbf{y}_i^s \log(\epsilon + \frac{\mathbf{y}_i^s}{\epsilon + \hat{\mathbf{y}}_i}), \tag{11}$$

where ϵ is a regularization constant and set to 2.2204×10^{-16};

$$L_{\mathrm{CC}}(\mathbf{y}^s, \hat{\mathbf{y}}) = \frac{\mathrm{cov}(\mathbf{y}^s, \hat{\mathbf{y}})}{\sigma(\mathbf{y}^s)\sigma(\hat{\mathbf{y}})}, \tag{12}$$

where $\mathrm{cov}(\cdot)$ is the covariance and $\sigma(\cdot)$ is standard deviation;

$$L_{\mathrm{SIM}}(\mathbf{y}^s, \hat{\mathbf{y}}) = \sum_i \min(\mathbf{y}_i^s, \hat{\mathbf{y}}_i). \tag{13}$$

In $L_{\mathrm{KLD}}, L_{\mathrm{CC}}$ and $L_{\mathrm{SIM}}, \mathbf{y}^s$, and $\hat{\mathbf{y}}$ are normalized so that $\sum_i \mathbf{y}_i^s = \sum_i \hat{\mathbf{y}}_i = 1$.

Since the higher NSS, SIM, and CC values and the lower KLD value represent the better agreement between predicted saliency maps and ground truth, we set λ_1, λ_3, and λ_4 to negative and λ_2 to positive. In order to balance the impact of different sub-loss functions on the module result, we determine the weights of individual sub-loss functions based on TranSalNet's performance on the SALICON validation set. In our experiments, the weights are adjusted to ensure these sub-loss functions (note the ranges of output values are different for these functions) contribute relatively equally to the model outcome. This is achieved by training and validating TranSalNet on the SALICON training and validation sets each time by a single sub-loss function. According to the recorded minimal loss values on the validation set, weights are initially assigned to the sub-loss functions so that their contributions to the combined loss are relatively equal. In a second step, these weights in a combined loss are further adjusted to achieve balanced results on all evaluation metrics. As per our empirical studies, the default weights $\lambda_1, \lambda_2, \lambda_3$, and λ_4 of the combined loss function are set to -1, 10, -2, and -1, respectively.

4. Experimental results and discussion

4.1. Datasets

Four commonly used benchmark saliency datasets are used to train and evaluate our proposed saliency model and variants.

- *SALICON* [29] contains 10,000 training, 5,000 validation, and 5000 testing images. The ground truth annotations of its test set are unpublished and used for a challenge named *LSUN 2017*[1] to test the performance of the saliency models. Different from other benchmark datasets, it employs mouse clicks instead of an eye tracker to record human visual attention.

[1] https://competitions.codalab.org/competitions/17136.

- *CAT2000* [60] contains 2,000 publicly available images of 20 categories such as action, art, cartoon etc, where each category includes 100 images. Each image is associated with its eye-tracking data of 24 observers.
- *MIT1003* [61] consists of 1,003 natural indoor and outdoor images with eye-tracking data of 15 observers.
- *MIT300* [62] includes 300 natural indoor and outdoor images. The eye-tracking data is unpublished, it is used as the test set of the MIT/Tübingen benchmark [63].

4.2. Evaluation metrics

Various metrics have been proposed to evaluate the agreement between the predicted saliency map and the ground truth. In general, these metrics can be described as location-based and distribution-based metrics depending on how the ground truth is represented [46]; the former adopts the fixation map (i.e., in the form of a binary image) and the latter uses the saliency map (i.e., in the form of a gray-scale image) as the ground truth for visual saliency evaluation. Six popular metrics are widely used to quantify the general performance of saliency models, including CC, SIM, KLD, NSS, AUC (Area under ROC Curve), and sAUC (Shuffled AUC). Details of these metrics can be found in [46]. The first three are distribution-based metrics, and the remaining three are location-based metrics. For KLD, the closer the value is to zero, the better the agreement between prediction and ground truth. For the other five metrics, higher values represent higher consistency.

Now, in this paper, we aims to evaluate the general performance of our proposed model, but in the meantime the perceptual relevance of the saliency model is the focus of our study. To this end, on the basis of the study of [46], we classify the six metrics into two categories based on their capability of being in close agreement with human judgements of saliency maps: "*perception-based metrics*", which include NSS, CC, and SIM; and "*non-perception-based metrics*", which include sAUC, AUC, and KLD [46]. Note, "*non-perception-based metrics*" do not necessarily mean they are not measuring the gaze behaviour, they may focus on specific properties of viewing behaviour, such as detecting salient objects in the visual field. It is stated in [46] that "*AUC, KL are appropriate for detection applications, as they penalize target detection failures. However, where it is important to evaluate the relative importance of different image regions, such as for image-retargeting, compression, and progressive transmission, metrics like NSS or SIM are a better fit.*" This provides sufficient grounds for building perceptually more relevant saliency prediction models, which is the primary goal of our work.

4.3. Setup

By following a similar procedure in the state-of-the-art [23,35,17,19,18], a model should be first initialised by the weights pre-trained on ImageNet [45], then trained on the 10,000 images of the SALICON training set to reduce the risk of overfitting. Consequently, the best model on its validation set should be selected for further testing on the SALICON test set and training on MIT1003 and CAT2000.

To obtain fair results in each dataset, k-fold cross-validation ($k = 10$) is applied for each model. More specifically, each dataset is divided into 10 non-overlapping subsets. For MIT1003, each subset contains around 100 images; For CAT2000, each subset contains 200 images (10 from each category). Each time, one subset is kept as a test set, one as a validation set, and the remaining eight subsets altogether are used as the training set. To eliminate randomness, each test set corresponds to a fixed validation set and training set. We report the overall performance of 10 times test results.

To reduce the computational cost while aligning with the aspect ratio (4:3) of the images in SALICON, all input images are resized and padded to a same size of 384×288 pixels. A consistent standard is followed in all training phases. The Adam optimizer [64] is used to minimize the loss function. The learning rate is set to 1×10^{-5}, which is then multiplied by 0.1 for every 3 epochs. Models are trained with a batch size of 4 for 30 epochs with a stop patience of 5 epochs.

4.4. Ablation study

Ablation experiments are conducted to investigate the contribution of three key components in our modelling: (1) Transformer encoders (E_1, E_2, and E_3 denote Transformer encoder 1, 2, and 3 in Fig. 2, respectively), (2) Skip-connections (SC), (3) Combined loss function (L_{CB}). To this end, nine model variants are constructed to demonstrate the added value of one or more of the above components, as shown in Table 1. Among them, BaseNet is constructed as a baseline that adopts the widely used ResNet-50 as the CNN encoder, removes all transformer encoders and skip-connections except for the $Conv_{1\times1}$ layer before transformer encoder 1, and is trained by the BCE loss function. BaseNet + adds the transformer encoder 1 based on the BaseNet. SkipNet is equipped with skip-connections based on the BaseNet. TranSalNet_Res_BCE adds the transformer encoder 1, 2, and 3 based on the SkipNet, which utilises ResNet-50 as the CNN encoder and is identical in architecture to the proposed TranSalNet (demonstrated in Fig. 2) but is trained by the BCE loss. The model variants trained by the combined loss that are consistent with the architecture of the above four model variants are denoted as BaseNet(L_{CB}), BaseNet+(L_{CB}), SkipNet(L_{CB}), and TranSalNet_Res, respectively. TranSalNet_Dense replaces ResNet-50 with DenseNet-161 as the CNN encoder. The overall performance of these model variants on the MIT1003 and CAT2000 datasets is shown in Table 2. The illustration of saliency maps of four images from these two datasets can also be found in Fig. 3.

By comparing BaseNet/BaseNet(L_{CB}) and BaseNet+/BaseNet+(L_{CB}), it can be found that adding a transformer encoder improves the overall performance, i.e., BaseNet+/BaseNet+(L_{CB}) outperforms BaseNet/BaseNet(L_{CB}) in the majority of instances. Especially, on the perception-based metrics, i.e., CC, SIM, and NSS, BaseNet+/BaseNet+(L_{CB}) give consistently better performance than BaseNet/BaseNet(L_{CB}), suggesting that the transformer encoder contributes to the perceptual relevance of saliency prediction. Besides, the benefits of enhancing saliency prediction by providing multi-scale image representations through skip-connections have been demonstrated in previous studies. Similarly, by adding skip-connections to BaseNet/BaseNet(L_{CB}), the performance of model variants SkipNet/SkipNet(L_{CB}) improves on most instances in the ablation study.

By uniting transformer encoders and skip-connections, the decoder network can obtain multi-scale feature maps with long-range context enhanced by transformer encoders. As a result, the performance of TranSalNet_Res_BCE/TranSalNet_Res is further boosted on all instances of perception-based metrics as well as most instances of non-perception-based metrics. This provides additional evidence that the transformer is of added value for visual saliency prediction. Also, this demonstrates the effectiveness of the TranSalNet architecture, which integrates transformer encoders into CNN-based models via skip-connections to obtain multi-scale representations with enhanced long-range visual information.

Table 1
Model variants purposely constructed for ablation study to explore the contribution of skip-connections (SC), transformer encoders (E_1 to E_3), and combined loss function (L_{CB}).

Model variants	E_1	E_2	E_3	SC	Loss function	Backbone
BaseNet	–	–	–	–	L_{BCE}	ResNet-50
BaseNet+	✓	–	–	–	L_{BCE}	ResNet-50
SkipNet	–	–	–	✓	L_{BCE}	ResNet-50
TranSalNet_Res_BCE	✓	✓	✓	✓	L_{BCE}	ResNet-50
BaseNet(L_{CB})	–	–	–	–	L_{CB}	ResNet-50
BaseNet+(L_{CB})	✓	–	–	–	L_{CB}	ResNet-50
SkipNet(L_{CB})	–	–	–	✓	L_{CB}	ResNet-50
TranSalNet_Res	✓	✓	✓	✓	L_{CB}	ResNet-50
TranSalNet_Dense	✓	✓	✓	✓	L_{CB}	DenseNet-161

Table 2
Ablation study: performance of nine model variants purposely constructed for ablation study to explore the contribution of skip-connections, transformer encoders, and the combined loss function based on **MIT1003** and **CAT2000** datasets. Red and orange font indicate the best and 2nd best performance, respectively.

	MIT1003						CAT2000					
	perception-based metrics			non-perception-based metrics			**perception-based metrics**			non-perception-based metrics		
Variation	CC ↑	SIM ↑	NSS ↑	sAUC ↑	AUC ↑	KLD ↓	CC ↑	SIM ↑	NSS ↑	sAUC ↑	AUC ↑	KLD ↓
BaseNet	0.7361	0.5858	2.7129	0.7460	0.9018	0.7886	0.8536	0.7277	2.3043	0.6062	0.8775	0.5190
BaseNet+	0.7446	0.6003	2.7437	0.7472	0.9046	0.8369	0.8622	0.7343	2.3286	0.6054	0.8784	0.5056
SkipNet	0.7393	0.5381	2.7571	0.7505	0.9011	0.7327	0.8582	0.7187	2.3298	0.6111	0.8782	0.3207
TranSalNet_Res_BCE	0.7593	0.6122	2.8326	0.7507	0.9086	0.7785	0.8765	0.7458	2.3887	0.6035	0.8803	0.5040
BaseNet(L_{CB})	0.7386	0.5966	2.7156	0.7508	0.9046	0.7839	0.8589	0.7332	2.3349	0.6063	0.8787	0.5146
BaseNet+(L_{CB})	0.7449	0.6046	2.7588	0.7505	0.9066	0.8321	0.8666	0.7381	2.3578	0.6076	0.8794	0.5053
SkipNet(L_{CB})	0.7471	0.5998	2.7692	0.7534	0.9067	0.7558	0.8612	0.7347	2.3409	0.6081	0.8797	0.4730
TranSalNet_Res	0.7595	0.6145	2.8501	0.7546	0.9093	0.7779	0.8786	0.7492	2.4154	0.6054	0.8811	0.5036
TranSalNet_Dense	0.7743	0.6279	2.9214	0.7547	0.9116	0.7862	0.8823	0.7512	2.4290	0.6099	0.8820	0.4715

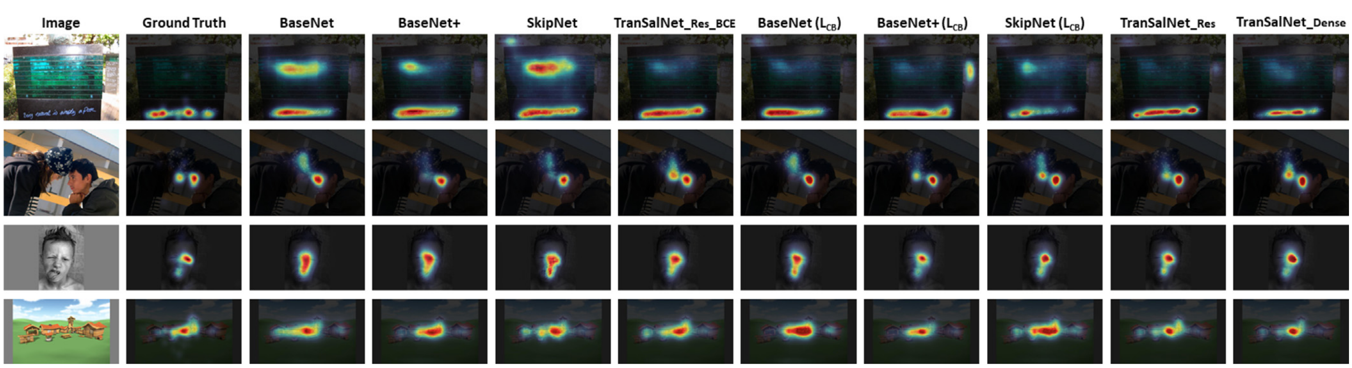


Fig. 3. Comparison of the saliency prediction performance of nine model variants in our ablation study. The images of top two rows are from the MIT1003 dataset and the bottom two rows are from the CAT2000 dataset. It can be seen that by adopting transformer encoder, skip-connection to provide multi-scale information, and combined loss function, the generated saliency maps are significantly refined relative to the ground truth.

Table 2 also demonstrates the practical plausibility of training the proposed model with the linear combination of sub-loss functions. Compired with the model variations trained by L_{BCE} (i.e., BaseNet, BaseNet+, SkipNet, and TranSalNet_Res_BCE), the model variations trained by L_{CB} (i.e., BaseNet(L_{CB}), BaseNet+(L_{CB}), SkipNet(L_{CB}), and TranSalNet_Res) achieve higher performance on the majority of saliency metrics. In particular, the TranSalNet_Res outperforms the TranSalNet_Res_BCE on all instances in the ablation study. In summary, the effectiveness of the transformer encoder, the TranSalNet architecture, and the combined loss function has now been demonstrated in this ablation study.

In addition, previous research [20,24] has shown that using backbones with greater representational capability could improve saliency prediction. Similarly, by simply replacing the backbone from the widely used but comparatively "shallow" ResNet-50 (used by TranSalNet_Res) with DenseNet-161 [28], TranSalNet_Dense has been further improved as shown in Table 2.

4.5. Comparison with state-of-the-art methods

4.5.1. On MIT1003 and CAT2000 datasets

Seven state-of-the-art deep learning-based saliency models that adopt multi-scale representations or attention mechanisms, including FastSal [65], UNISAL [21], MSI-Net [19], SAM-VGG [18], SAM-ResNet [18], ML-Net [16], and Deep Visual Attention (DVA) [17] are selected for the general performance comparison on the MIT1003 and CAT2000 datasets. In order to ensure a fair comparison, the same k-fold Cross-Validation ($k = 10$ for MIT1003 and CAT2000) strategy and the dataset splitting method used in TranSalNet are employed for fine-tuning and testing of these models. The corresponding pre-trained weights on the SALICON dataset is loaded for each fine-tuning instance. For MIT1003 and CAT2000 datasets, the overall performance of 10 times test results is reported in Table 3.

It can be seen that our models (both TranSalNet_Res and TranSalNet_Dense) achieve the best performance on all perception-based metrics in both MIT1003 and CAT2000, while producing competitive results on non-perception-based metrics (i.e., being best or 2nd best in most instances in the comparative study). It should be noted that our TranSalNet_Res and the five state-of-the-art models all use ResNet-50 or VGGNet (representing similar network capacity) as the feature extraction network. TranSalNet_Res achieves the best performance on most instances (except for sAUC and KLD in CAT2000), implying the contribution of enhanced long-range information to saliency prediction using

Table 3
Performance comparison of state-of-the-art saliency models on **MIT1003** and **CAT2000**. Red and orange font indicate the best and 2nd best performance, respectively.

	MIT1003						CAT2000					
	perception-based metrics			non-perception-based metrics			**perception-based metrics**			non-perception-based metrics		
Model name	CC ↑	SIM ↑	NSS ↑	sAUC ↑	AUC ↑	KLD ↓	CC ↑	SIM ↑	NSS ↑	sAUC ↑	AUC ↑	KLD ↓
FastSal [65]	0.5901	0.4783	2.0078	0.7060	0.8745	1.0359	0.7213	0.6031	1.8590	0.6180	0.8599	0.5523
UNISAL [21]	0.7340	0.5973	2.7593	0.7326	0.9026	1.0138	0.8417	0.7207	2.2575	0.5982	0.8758	0.5302
MSI-Net [19]	0.7473	0.6081	2.8007	0.7454	0.9068	0.8155	0.8655	0.7398	2.3547	0.6071	0.8809	0.4280
SAM-VGG [18]	0.7260	0.5976	2.7520	0.7256	0.9003	1.2195	0.8680	0.7391	2.4138	0.5966	0.8784	0.6383
SAM-ResNet [18]	0.7466	0.6068	2.8001	0.7365	0.9024	1.2470	0.8706	0.7395	2.4108	0.5932	0.8778	0.6702
ML-Net [16]	0.5979	0.4960	2.3329	0.7218	0.8623	1.3496	0.5221	0.5407	1.4485	0.6212	0.8104	1.1101
DVA [17]	0.6990	0.5663	2.5740	0.7258	0.8970	0.7528	0.8616	0.7335	2.3447	0.6014	0.8783	0.4492
TranSalNet_Res	0.7595	0.6145	2.8501	0.7546	0.9093	0.7779	0.8786	0.7492	2.4154	0.6054	0.8811	0.5036
TranSalNet_Dense	0.7743	0.6279	2.9214	0.7547	0.9116	0.7862	0.8823	0.7512	2.4290	0.6099	0.8820	0.4715

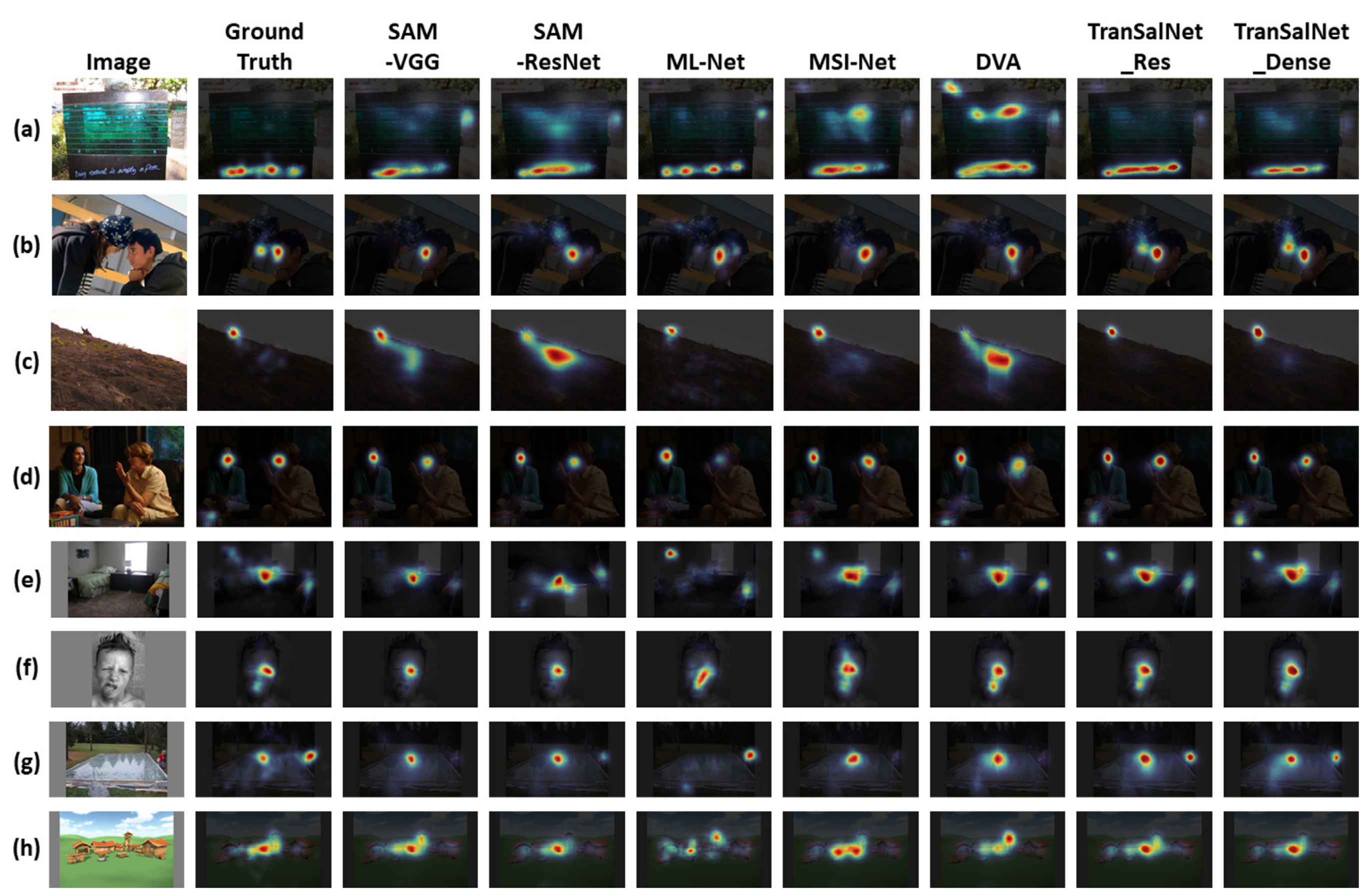


Fig. 4. Comparison of saliency maps generated by our models (TranSalNet_Res and TranSalNet_Dense) and other state-of-the-art saliency models. The images from (a) to (d) are from the MIT1003 dataset, and the images from (e) to (h) are from the CAT2000 dataset.

transformers. Moreover, the performance our TranSalNet_Res could be further enhanced by replacing ResNet-50 by a network with higher capacity, namely DenseNet-161. Fig. 4 shows saliency maps generated by our models and other models for images including common contexts such as objects, portraits, natural, indoor, social, and cartoon scenes. By visually assessing these saliency maps, our models are in closer agreement with the ground truth than other models.

4.5.2. On MIT300 competition

For the MIT300 competition, we use the MIT1003 to train an optimal model, in which 703 images are randomly selected as a training set and the rest as a validation set. The optimal model is submitted to and tested by the MIT/Tuebingen Saliency Benchmark [63]. It should be noted that the benchmark evaluates models by different standards, i.e., models must be explicitly claimed as either probabilistic or non-probabilistic models, so they can be fairly evaluated within the category they belong to [46]. In this paper, same as the original MIT Saliency Benchmark [46], we "do not assume that our model is probabilistic". Note that for evaluating probabilistic models, metric-specific adaptations are applied using regularization and scaling of saliency values, hence, a probabilistic model generates optimal saliency maps for individual metrics [49]. But a non-probabilistic model only outputs a single saliency map for all metrics. So it is nontrivial to compare a non-probabilistic (i.e., classical) model to a probabilistic model [46]. To avoid unfair model comparison under different assumptions, Table 4 shows only non-probabilistic classical saliency models on the leader-board of [63]. It can be seen that our models (both TranSalNet_Res and TranSalNet_Dense) consistently rank in the top 1st or 2nd positions on the perception-based metrics (note the only exception is for TranSalNet_Res on NSS, but its performance score is fairly comparable to the 1st or 2nd scores as shown in Table 4). On the non-perception-based metrics, our models exhibit competitive performance on sAUC and AUC, with the performance scores comparable to the results in the 1st and 2nd positions. In addition, even though we include top probabilistic models such as DeepGaze II-E [24], MSI-Net [19], UNISAL [21], SalFBNet [67], and DeepGaze II [23] for performance comparison, our model can still remain competitive in perception-based metrics (results available on website of [63]).

4.5.3. On LSUN'17 competition

Although our aim is to predict the spatial distribution of human fixations, the human attention measured by mouse tracking can still reflect eye movement behaviour to a certain extent [29]. SALICON provides so far the largest-scale saliency dataset (via mouse tracking), which allows the opportunity to examine the saliency models from the perspective of being "data rich". Moreover, for LSUN'17 competition (on SALICON test set), a unified evaluation process is adopted, i.e., the saliency models are not treated differently because of their type of being probabilistic or classical. In the competition each model submitted should generate one single saliency map for each image. Therefore, in order to provide a complementary comparison of state-of-the-art saliency models, Table 5 reports the results of models submitted to the competition based on the 2017 version (i.e., the latest version). It can be seen that our TranSalNet_Res and TranSalNet_Dense achieve superior performance on the perception-based metrics and promising results on other non-perception-based metrics. This shows that our model are competitive on the LSUN 2017 leaderboard, in particular for prediction saliency in a perceptually relevant manner.

4.5.4. Discussion

It is crucial to note that metric selection for saliency model evaluation should be based on specific modelling assumptions and specific target applications [46]. The study in [46] concludes that "*under the assumptions of non-probabilistic modelling, NSS and CC provide the fairest comparison*"; "*if evaluating probabilistic models, KLD is recommended*"; and "*specific tasks and applications also call for a difference choice of metrics*". In [48], researchers have verified that "NSS, CC and SIM best correspond to human perception". In a recent study [47], it is found that CC and SIM are the most appropriate saliency evaluation metrics for image quality assessment applications. Therefore, as the results demonstrated in Table 3, Table 4, and Table 5, the proposed saliency models (TranSalNet_Res and TranSalNet_Dense) could be the best "human-like" models (i.e., based on perception-based metrics CC and SIM) to evaluate the relative importance of different image regions for the applications such as image re-targeting, image compression and transmission, and visual quality assessment.

Using skip-connections to provide multi-scale features from encoder to decoder has been shown in previous studies to be an effective method for computer vision tasks. For example, the

Table 4
MIT300 competition for saliency models. The results are administered and reported by the Benchmark [66]. Red and orange font indicate the best and 2nd best performance, respectively.

Model name	perception-based metrics			non-perception-based metrics		
	CC ↑	SIM ↑	NSS ↑	sAUC ↑	AUC ↑	KLD ↓
EML-NET [20]	0.7893	0.6756	2.4876	0.7469	0.8762	0.8439
CASNet II [22]	0.7054	0.5806	1.9859	0.7398	0.8552	0.5857
GazeGAN [36]	0.7579	0.6491	2.2118	0.7316	0.8607	1.3390
SAM-VGG [18]	0.6630	0.5986	1.9552	0.7305	0.8473	1.2746
SAM-ResNet [18]	0.6897	0.6122	2.0628	0.7396	0.8526	1.1710
DVA [17]	0.6631	0.5848	1.9305	0.7257	0.8430	0.6293
ML-Net [16]	0.6633	0.5819	1.9748	0.7399	0.8386	0.8006
eDN [42]	0.4518	0.4112	1.1399	0.6180	0.8171	1.1369
SalGAN [66]	0.6740	0.5932	1.8620	0.7354	0.8498	0.7574
HATES	0.7897	0.5313	2.3762	0.7549	0.8744	0.7146
TranSalNet_Res	0.7991	0.6852	2.3758	0.7471	0.8730	0.9019
TranSalNet_Dense	0.8070	0.6895	2.4134	0.7467	0.8734	1.0141

Table 5
LSUN'17 competition (on SALICON test set) for saliency models. Results are provided by the authors. Red and orange font indicate the best and 2nd best performance, respectively.

Model name	perception-based metrics CC ↑	SIM ↑	NSS ↑	non-perception-based metrics sAUC ↑	AUC ↑	KLD ↓
DeepGaze II-E [24]	0.872	0.733	1.996	0.767	0.869	0.285
SalFBNet [67]	0.892	0.772	1.952	0.740	0.868	0.236
MSI-Net [19]	0.889	0.784	1.931	0.736	0.865	0.307
UNISAL [21]	0.879	0.775	1.952	0.739	0.864	–
EML-NET [20]	0.886	0.780	2.050	0.746	0.866	0.520
GazeGAN [36]	0.879	0.773	1.899	0.736	0.864	0.376
SAM-ResNet [18]	0.899	0.793	1.990	0.741	0.865	0.610
TranSalNet_Res	0.901	0.796	1.998	0.742	0.866	0.414
TranSalNet_Dense	0.907	0.803	2.014	0.747	0.868	0.373

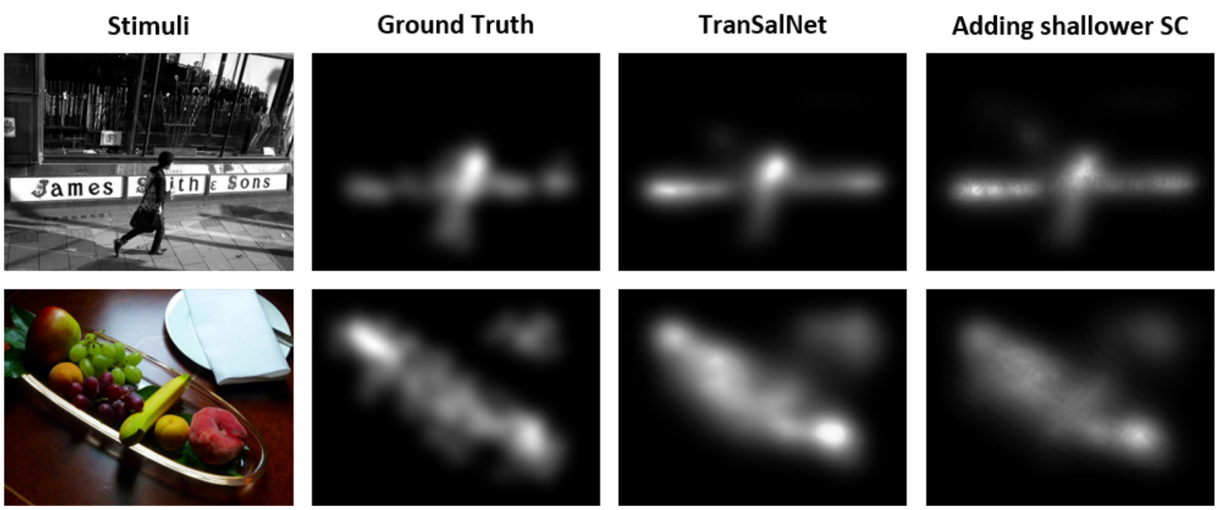


Fig. 5. The column on the righthand side illustrates the salinency maps with undesired artefacts caused by adding skip-connections to TranSalNet_Res to connect shallow encoder blocks with decoder block_4 and block_5. From left to right, the remaining three columns are: stimuli, ground truth saliency maps, and saliency maps generated from TranSalNet_Res, respectively.

widely used U-Net [55] style networks usually connect feature maps of each spatial size to the decoders from shallow to deep encoder blocks. However, as can be seen in Fig. 5, using skip-connections to connect shallow encoder blocks (i.e., the blocks provide feature maps with spatial sizes of $\frac{w}{4} \times \frac{h}{4}$ and $\frac{w}{2} \times \frac{h}{2}$) with decoder blocks (i.e., block_4 and block_5 in the decoder) may lead to some shapes of objects and texts appearing in the predicted saliency maps, which are not consistent with the ground truth. This implies that adding low-level features from the encoder directly to the decoder may interfere with the saliency prediction of TranSalNet.

Multi-head Self Attention (MSA) is part of the transformer encoder. Previous research has shown that the number of heads of MSA could affect the model's performance [68]. According to the suggestions from [46], we use CC and SIM as the performance metrics to illustrate the impact of the head number of MSA on our proposed TranSalNet in Fig. 6. For each head number combination, the model is trained on the SALICON training set, validated with 2000 images of its validation set, and tested on the rest of the validation set three times. The demonstrated results are the mean results. As can be seen in Fig. 6, the scores of CC and SIM tend to increase with the increase in the head number of MSA. However, when the transformer encoders 1 and 2 (E_1 and E_2) adopt 12 heads

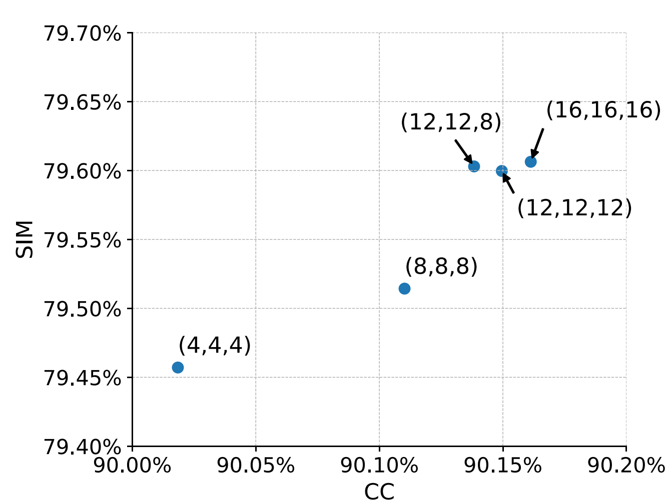


Fig. 6. Illustrations of the impact of the head number of MSA on the performance of TranSalNet. Values in brackets (i, j, k) indicate the head numbers of transformer encoder 1, 2, and 3 are set to i, j, and k respectively.

each, and transformer encoder 3 (E_3) adopts 8 heads of MSA, the performance of the model tends to be saturated in the CC-SIM performance space. Therefore, considering the trade-off between computational resource consumption and model performance, we chose 12 heads for the transformer encoder 1 and 2, and 8 heads for the transformer encoder 3 in this study.

5. Conclusion

In this paper, we have proposed a novel saliency model for predicting saliency maps that are perceptually in close agreement with the ground truth. By integrating transformers into CNNs, saliency models can significantly benefit from capturing long-range spatial information at multiple perceptual levels. An ablation study has demonstrated the contributions of the transformer encoders to a CNN model, especially the added value of transformers in enhancing the perceptual relevance of saliency prediction. Experimental results show that the proposed models have achieved superior performance on the public benchmarks and competitions for saliency models, particularly having yielded notable results on perception-based saliency evaluation metrics. The perceptually more relevant saliency models have the potential to advance many image processing applications.

CRediT authorship contribution statement

Jianxun Lou: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft. **Hanhe Lin:** Software, Validation, Writing – review & editing. **David Marshall:** Investigation, Validation, Formal analysis, Writing – review & editing. **Dietmar Saupe:** Supervision, Validation, Formal analysis, Writing – review & editing. **Hantao Liu:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 and the China Scholarship Council – ID 202008220129.

References

[1] J. Jonides, D. Irwin, S. Yantis, Integrating visual information from successive fixations, Science 215 (4529) (1982) 192–194.
[2] A. Borji, Saliency prediction in the deep learning era: Successes and limitations, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2) (2021) 679–700.
[3] K. Song, T. Yao, Q. Ling, T. Mei, Boosting image sentiment analysis with visual attention, Neurocomputing 312 (2018) 218–228.
[4] Y. Han, B. Han, X. Gao, Human scanpath estimation based on semantic segmentation guided by common eye fixation behaviors, Neurocomputing 453 (2021) 705–717.
[5] K. Gu, G. Zhai, X. Yang, W. Zhang, C.W. Chen, Automatic contrast enhancement technology with saliency preservation, IEEE Trans. Circuits Syst. Video Technol. 25 (9) (2015) 1480–1494.
[6] J. Chen, Z. Li, Y. Jin, D. Ren, H. Ling, Video saliency prediction via spatio-temporal reasoning, Neurocomputing 462 (2021) 59–68, https://doi.org/10.1016/j.neucom.2021.07.088.
[7] J. Han, E.J. Pauwels, P. de Zeeuw, Fast saliency-aware multi-modality image fusion, Neurocomputing 111 (2013) 70–80.
[8] D. Mishra, S.K. Singh, R.K. Singh, D. Kedia, Multi-scale network (MsSG-CNN) for joint image and saliency map learning-based compression, Neurocomputing 460 (2021) 95–105.
[9] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Networks 19 (9) (2006) 1395–1407.
[10] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS), NIPS'06, MIT Press, Cambridge, MA, USA, 2006, pp. 545–552.
[11] E. Erdem, A. Erdem, Visual saliency estimation by nonlinearly integrating features using region covariances, J. Vision 13 (4) (2013), 11–11.
[12] S. Fang, J. Li, Y. Tian, T. Huang, X. Chen, Learning discriminative subspaces on random contrasts for image saliency analysis, IEEE Trans. Neural Networks Learn. Syst. 28 (5) (2017) 1095–1108.
[13] J. Stoll, M. Thrun, A. Nuthmann, W. Einhäuser, Overt attention in natural scenes: Objects dominate features, Vision. Res. 107 (2015) 36–48.
[14] W. Einhäuser, M. Spain, P. Perona, Objects predict fixations better than early saliency, J. Vision 8 (14) (2008), 18–18.
[15] M. Cerf, E.P. Frady, C. Koch, Faces and text attract gaze independent of the task: Experimental data and computer model, J. Vision 9 (12) (2009), 10–10.
[16] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, A deep multi-level network for saliency prediction, in: 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 3488–3493..
[17] W. Wang, J. Shen, Deep visual attention prediction, IEEE Trans. Image Process. 27 (5) (2018) 2368–2378.
[18] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, Predicting human eye fixations via an LSTM-based saliency attentive model, IEEE Trans. Image Process. 27 (10) (2018) 5142–5154.
[19] A. Kroner, M. Senden, K. Driessens, R. Goebel, Contextual encoder–decoder network for visual saliency prediction, Neural Networks 129 (2020) 261–270.
[20] S. Jia, N.D. Bruce, EML-NET: An expandable multi-layer network for saliency prediction, Image Vis. Comput. 95 (2020) 103887.
[21] R. Droste, J. Jiao, J.A. Noble, Unified image and video saliency modeling, in: A. Vedaldi, H. Bischof, T. Brox, J. Frahm (Eds.), Computer Vision - ECCV 2020–16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V, Vol. 12350 of Lecture Notes in Computer Science, Springer, 2020, pp. 419–435..
[22] S. Fan, Z. Shen, M. Jiang, B.L. Koenig, J. Xu, M.S. Kankanhalli, Q. Zhao, Emotional attention: A study of image sentiment and visual attention, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7521–7531..
[23] M. Kummerer, T.S.A. Wallis, L.A. Gatys, M. Bethge, Understanding low- and high-level contributions to fixation prediction, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4799–4808..
[24] A. Linardos, M. Kümmerer, O. Press, M. Bethge, DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 12919–12928..
[25] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 2015 International Conference on Learning Representations (ICLR), 2015..
[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778..
[28] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269..
[29] M. Jiang, S. Huang, J. Duan, Q. Zhao, SALICON: Saliency in context, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1072–1080..
[30] A. Toet, Computational versus psychophysical bottom-up image saliency: A comparative evaluation study, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2131–2146.
[31] E.E.M. Stewart, M. Valsecchi, A.C. Schütz, A review of interactions between peripheral and foveal vision, J. Vision 20 (12) (2020), 2–2.
[32] R. Rosenholtz, Capabilities and limitations of peripheral vision, Annual Review of Vision Science 2 (1) (2016) 437–457.
[33] L. Spillmann, J.S. Werner, Long-range interactions in visual perception, Trends Neurosci. 19 (10) (1996) 428–434.
[34] L. Itti, C. Koch, A saliency-based search mechanism for overt and covert shifts of visual attention, Vision. Res. 40 (10) (2000) 1489–1506.
[35] X. Huang, C. Shen, X. Boix, Q. Zhao, SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 262–270..
[36] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, P. Le Callet, How is gaze influenced by image transformations? Dataset and model, IEEE Trans. Image Process. 29 (2020) 2287–2300.
[37] N. Liu, J. Han, A deep spatial contextual long-term recurrent convolutional network for saliency detection, IEEE Trans. Image Process. 27 (7) (2018) 3264–3274.
[38] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, et al. (Eds.), Advances in Neural Information Processing Systems, 30, Curran Associates, Inc., 2017.

[39] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186..

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR), 2021..

[41] M. Chen, A. Radford, R. Child, et al., Generative pretraining from pixels, in: H.D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proc. of Machine Learning Research, PMLR, 2020, pp. 1691–1703..

[42] E. Vig, M. Dorr, D. Cox, Large-scale optimization of hierarchical features for saliency prediction in natural images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2798–2805..

[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9..

[44] B. Zoph, V. Vasudevan, J. Shlens, Q.V. Le, Learning transferable architectures for scalable image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8697–8710..

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252.

[46] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, F. Durand, What do different evaluation metrics tell us about saliency models?, IEEE Trans Pattern Anal. Mach. Intell. 41 (3) (2019) 740–757.

[47] X. Yang, F. Li, H. Liu, A measurement for distortion induced saliency variation in natural images, IEEE Trans. Instrum. Meas. 70 (2021) 1–14.

[48] J. Li, C. Xia, Y. Song, S. Fang, X. Chen, A data-driven metric for comprehensive evaluation of saliency models, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 190–198..

[49] M. Kümmerer, T.S.A. Wallis, M. Bethge, Saliency benchmarking made easy: Separating models, maps and metrics, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Lecture Notes in Computer Science, Springer International Publishing, 2018, pp. 798–814.

[50] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, J. Feng, DeepViT: Towards deeper vision transformer (2021). arXiv:2103.11886..

[51] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022..

[52] N. Liu, N. Zhang, K. Wan, L. Shao, J. Han, Visual saliency transformer, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 4722–4732..

[53] Y. Liu, D. Zhang, Q. Zhang, J. Han, Part-object relational visual saliency, IEEE Trans. Pattern Anal. Mach. Intell. (2021), 1–1.

[54] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, J. Han, RGB-T Salient Object Detection via Fusing Multi-Level CNN Features, IEEE Trans. Image Process. 29 (2020) 3321–3335.

[55] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[56] B. Cheng, B. Xiao, J. Wang, H. Shi, T.S. Huang, L. Zhang, HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5385–5394..

[57] Y. Liu, J. Han, Q. Zhang, C. Shan, Deep salient object detection with contextual information guidance, IEEE Trans. Image Process. 29 (2020) 360–374.

[58] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr, L. Zhang, Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6877–6886..

[59] Y. Fang, H. Zhang, Y. Zuo, W. Jiang, H. Huang, J. Yan, Visual attention prediction for autism spectrum disorder with hierarchical semantic fusion, Signal Process.: Image Commun. 93 (2021) 116186.

[60] A. Borji, L. Itti, CAT2000: A large scale fixation dataset for boosting saliency research, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015 workshop on Future of DatasetsArXiv preprint arXiv:1505.03581 (2015)..

[61] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: IEEE International Conference on Computer Vision (ICCV), 2009, pp. 2106–2113..

[62] T. Judd, F. Durand, A. Torralba, A benchmark of computational models of saliency to predict human fixations, Tech. Rep. MIT-CSAIL-TR-2012-001, MIT Computer Science and Artificial Intelligence Lab (CSAIL), Cambridge, MA, USA (01 2012)..

[63] M. Kümmerer, Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, A. Torralba, MIT/Tübingen Saliency Benchmark,https://saliency.tuebingen.ai/..

[64] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015..

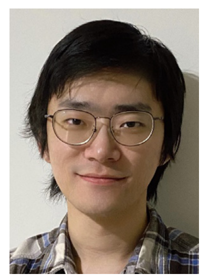
[65] F. Hu, K. McGuinness, FastSal: a Computationally Efficient Network for Visual Saliency Prediction, in: 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 9054–9061..

[67] G. Ding, N. Imamouglu, A. Caglayan, M. Murakawa, R. Nakamura, Salfbnet: Learning pseudo-saliency distribution via feedback convolutional networks (2021). arXiv:2112.03731..

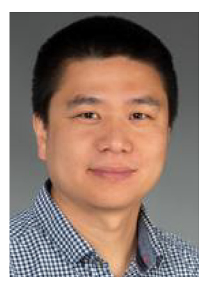
[68] P. Michel, O. Levy, G. Neubig, Are sixteen heads really better than one?, in: Advances in Neural Information Processing Systems, vol. 32, Curran Associates Inc, 2019..

Further Reading

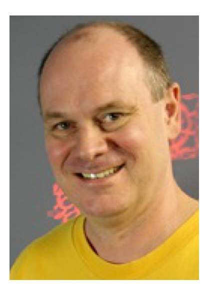
[66] J. Pan, C.C. Ferrer, K. McGuinness, N.E. O'Connor, J. Torres, E. Sayrol, X.G. i Nieto, Salgan: Visual saliency prediction with generative adversarial networks (2018). arXiv:1701.01081..



Jianxun Lou received the B.Eng. from Central South University, Changsha, China, in 2018 and the M.S. degree from Cardiff University, Cardiff, UK, in 2020. He is now pursuing his Ph.D. degree at the School of Computer Science and Informatics, Cardiff University, Cardiff, UK.



Hanhe Lin received his Ph.D. at the Department of Information Science, University of Otago, New Zealand in 2016. From 2016 to 2021, he was a postdoc at the Department of Computer and Information Science at the University of Konstanz, Germany, where he was working on project A05 (visual quality assessment) of SFB-TRR 161, funded by the German Research Foundation (DFG). Currently, he is a research fellow at the National Subsea Centre at Robert Gordon University, UK. His research interests include image processing, computer vision, machine learning, deep learning, and visual quality assessment.



Prof. David Marshall has been working in the field of computer vision since 1986. In 1989 he joined Cardiff University as lecturer and is now Professor of Computer Vision in the School of Computer Science and Informatics. David's research interests include articulated modelling of human faces, models of human motion, statistical modelling, high dimensional subspace analysis, audio/video image processing, and data/sensor fusion. He has published over 150 papers and one book in these research areas and has attracted over £4 M in research funding over his academic career. He is currently Head of the Visual Computing Research Group and Director of the Human Factors Technology Centre. http://users.cs.cf.ac.uk/Dave.Marshall/

Dietmar Saupe was born in Bremen, Germany, in 1954. He received the Dr.rer.nat. degree in mathematics from the University of Bremen, Germany, in 1982. From 1985 to 1993, he was an Assistant Professor with the Departments of Mathematics, first at the University of California, Santa Cruz, USA, and then at the University of Bremen, resulting in his habilitation. From 1993 to 1998, he was a Professor of computer science with the University of Freiburg, Germany, the University of Leipzig, Germany, until 2002, and since then, the University of Konstanz, Germany. He is the coauthor of the book Chaos and Fractals, which won the Association of American Publishers Award for Best Mathematics Book of the Year, in 1992, and well over 100 research articles. His research interests include image and video processing, computer graphics, scientific visualisation, dynamical systems, and sport informatics.

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands in 2011. He is currently an Associate Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology and IEEE Signal Processing Letters.