electronics

MDPI

Article

A Saliency Prediction Model Based on Re-Parameterization and Channel Attention Mechanism

Fei Yan [1], Zhiliang Wang [1], Siyu Qi [1] and Ruoxiu Xiao [1,2,*]

[1] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China; b20140367@xs.ustb.edu.cn (F.Y.); wzl@ustb.edu.cn (Z.W.); m202120772@xs.ustb.edu.cn (S.Q.)

[2] Beijing Engineering and Technology Center for Convergence Networks and Ubiquitous Services, School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

* Correspondence: xiaoruoxiu@ustb.edu.cn

Abstract: Deep saliency models can effectively imitate the attention mechanism of human vision, and they perform considerably better than classical models that rely on handcrafted features. However, deep models also require higher-level information, such as context or emotional content, to further approach human performance. Therefore, this study proposes a multilevel saliency prediction network that aims to use a combination of spatial and channel information to find possible high-level features, further improving the performance of a saliency model. Firstly, we use a VGG style network with an identity block as the primary network architecture. With the help of re-parameterization, we can obtain rich features similar to multiscale networks and effectively reduce computational cost. Secondly, a subnetwork with a channel attention mechanism is designed to find potential saliency regions and possible high-level semantic information in an image. Finally, image spatial features and a channel enhancement vector are combined after quantization to improve the overall performance of the model. Compared with classical models and other deep models, our model exhibits superior overall performance.

Keywords: visual attention; visual saliency; saliency prediction; deep learning; re-parameterization

Citation: Yan, F.; Wang, Z.; Qi, S.; Xiao, R. A Saliency Prediction Model Based on Re-Parameterization and Channel Attention Mechanism. *Electronics* **2022**, *11*, 1180. https://doi.org/10.3390/electronics11081180

Academic Editor: Eva Cernadas

Received: 19 March 2022
Accepted: 5 April 2022
Published: 8 April 2022

1. Introduction

The human visual system (HVS) receives hundreds of megabytes of visual data per second, but processes only 40 bits per second [1]. The visual attention mechanism plays an important role in this process [2]. When facing a complex scene, HVS will immediately select a few regions of interest related to the current behavior or task for priority processing, considerably decreasing the amount of input visual data and selectively processing each scene in different orders and strengths to avoid waste of calculation and reduce the difficulty of analysis.

The saliency detection task imitates the HVS mechanism to detect areas that can attract people's attention from the environment. This concept exhibits strong subjectivity, including related knowledge in many fields, such as neurobiology, psychology, and computer vision. Early saliency prediction models used this related knowledge, adopting the method of handcrafted features or artificial design tasks. However, the performance of saliency models gradually encountered a bottleneck. With the widespread application of deep models, the field of visual saliency detection has achieved considerable progress and played an important role in various studies. Multilayer deep models can automatically capture more features and train in an end-to-end manner. They combine feature extraction and saliency prediction, resulting in remarkable improvement in performance compared with the classical model. As shown in Figure 1, a deep saliency model can efficiently extract common features, such as human and contexture. However, the most

interesting or significant parts of an image are not necessarily these objects. The human visual model frequently has a reasoning process based on sensory stimulation. Although deep models have made significant achievements in saliency prediction, saliency models still require a higher-level concept to approach human-level performance. The important problem is how to imitate a human analysis scene and understand the mechanism of human gaze.

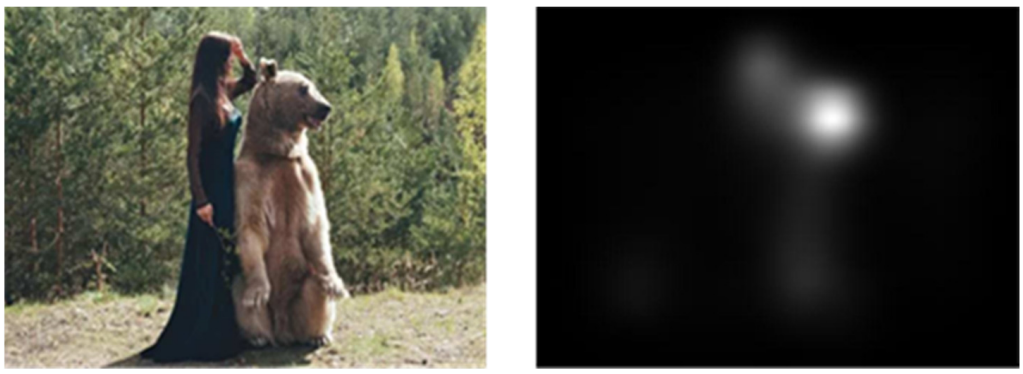

Figure 1. In the picture, the animal attracts considerably more attention than the human.

The saliency detection tasks consist of two parts: saliency prediction and salient object detection (SOD). In recent years, researchers have gradually changed to SOD tasks, omnidirectional images, and dynamic models. As a pure computer vision application, SOD can be easily applied to many different fields and has shown outstanding achievements [3]. Wang et al. [4] proposed a parameter- and weight-sharing model to obtain the sharing information, and they proposed a PAGE-Net [5] to obtain the edge information. Zhang et al. [6] proposed a dual refinement network (DRFNet) to process high-resolution images. However, a saliency model is more related to neuroscience and psychology basics, which still play a crucial role in promoting a variety of interdisciplinary tasks, such as human social interactions [7], end-to-end driving [8], medical diagnosis [9,10], and health monitoring [11,12].

To improve the performance of saliency models further and explore the role and importance of advanced features such as emotion or contexture in saliency prediction, a multiscale, deep network model is proposed in this study. Our major contributions are as follows:

- We propose a new, multilevel, deep neural network (DNN) model that adds an identity block to the network through re-parameterization. By integrating the identity block and improving the receptive field, we obtain more robust and accurate features. Simultaneously, the proposed model effectively reduces computational cost compared with the commonly used multiscale networks.
- We design a semantic perception subnetwork by adjusting channel features and exploring the correlation between high-level semantic information. The priority and importance of high-level information in visual saliency prediction are verified by testing and comparing datasets with rich semantic targets.

The organization of this paper is as follows: Section 2 summarises the related work on classical and deep saliency prediction models. Section 3 presents the network architecture and optimization method proposed in this study. Section 4 describes the experimental steps, including the evaluation measures and the results in two datasets on the basis of the analysis and comparison of public standards. Section 5 presents the model visualization and ablation analysis. Section 6 is the conclusion.

2. Related Work

2.1. Visual Saliency and Attention Mechanism

The attention mechanism has always been an important topic in neuroscience and psychology. Cognitive psychology emphasizes the initiative of human psychological activities and the importance of consciousness. It considers attention an important mechanism of human brain information processing, promoting the research and development of the attention mechanism. With the rapid development of cognitive psychology, many attention theories have emerged and exerted an important effect on the field of computer vision. These theories include the feature integration theory (FIT) proposed by Treisman [13] and the return–inhibition mechanism, based on the FIT, proposed by Koch and Ullman [14]. Early human visual attention system simulation also uses important achievements in physiology and psychology, such as center surround antagonism, maximization of information, and global rarity. Psychologists have determined that among many advanced concepts, the content that comprises human and facial expressions and human-related objects and words can exert considerable effects on people. These studies have guided and standardized subsequent saliency prediction models.

2.2. Visual Saliency Models

Early visual saliency prediction models can be divided into two categories: task agnostic (bottom-up) and task specific (top-down). Bottom-up visual saliency models are modeling by extracting low-level features, such as contrast, color, and texture. This attention–prediction mechanism is an autonomous and fast information process. For example, the earliest Itti et al. [15] model can simulate the process of human visual attention transfer without giving any prior information. Since then, some scholars have made improvements, such as local contrast analysis [16], global contrast analysis [17], conditional random field [18], sparse coding analysis [19], and superpixel [20,21]. Considering the diversity and complexity of top-down factors, top-down saliency modeling is a difficult task. Top-down visual saliency models are mostly Bayesian models [22,23]. In addition, Bayesian models can be regarded as the special case of decision theory models [24,25]. Both models simulate the biological calculation process of human visual saliency. Although the modeling methods of these classical models are diverse and creative, handcrafted features or tasks still induce a bottleneck in model performance.

With the development of machine learning and big data computing, Vig et al. [26] proposed the ensemble of deep networks (eDN) model in 2014. This model used a self-driving method to search for optimal features on a large scale for the first time. Since then, more researchers have adopted the deep learning method to study saliency prediction. Combined with target recognition networks commonly used in deep learning, such as AlexNet [27], VGG-16 [28], and GoogleNet [29], the deep learning method has achieved good performance. Deepgaze I [30] first used AlexNet and softmax layers to generate a saliency probability distribution map by using a classification method and applied transfer learning in the field of saliency prediction. Then, the Deepfix [31] model was changed to the VGG-16 network, and Deepgaze II [32] used the VGG-19 [28] network as its primary feature extraction network. In addition, many models optimize the network by adjusting different resolutions. Saliency in context (SALICON) [33] used convolutional neural networks (CNNs) trained with double tributary multiscale features. Pan et al. [34] proposed a shallow CNN (juntingnet) and a deep CNN (salnet) for saliency prediction. The probability distribution prediction proposed by Jetley et al. [35] defined saliency as a generalized Bernoulli distribution. The deep spatial contextual long-term recurrent convolutional network (DSCLRCN) proposed by Liu and Han [36] used the deep spatial long-term short-term (LSTM) model to capture global features. Subsequently, the ML-Net model proposed by Cornia et al. [37] combined the advantages of the aforementioned models. This model was composed of a feature extraction CNN, a feature coding network, and a priori learning network. Cornia et al. [38] subsequently proposed the SAM-ResNet and SAM-VGG models that combined

the full convolutional network and the LSTM to obtain spatial information. The loss function of this network was weighted by normalized scanning path saliency (NSS), a correlation coefficient (CC), and a similarity metric (SIM). Thereafter, SalGAN [39] conducted network training by using a countermeasure network. The saliency model EML-Net proposed by Jia et al. [40] used extreme learning machines (ELMs) to learn saliency prediction from each image in a set. The deep visual attention (DVA) model proposed by Wang et al. [41] used a skip-layer network to train multiple scales by the cooperation of the global and local predictions. The model proposed by Gorji and Clark [42] used shared attention to generate saliency maps, and the performance of these models was further improved. A fully convolutional network based on the deep learning framework automatically received global information and trains in an end-to-end manner to better identify the most significant region in an image. Such a network has gradually become the mainstream direction of saliency prediction. In recent years, saliency prediction has gradually developed in the field of dynamic and omnidirectional images (ODIs). Wang et al. [43] proposed the Attentive CNN-LSTM Network (ACLNet) that used the CNN-LSTM to encode static saliency information. Wang et al. [44] proposed the spatiotemporal residual attentive network (STRA-Net) that used global attention priors to capture information. Xu et al. [45] used adversarial networks to capture the head trajectory and train deep models.

2.3. Understanding Advanced Semantic Information

Deep models have made remarkable achievements in the field of saliency prediction; however, existing saliency models still cannot clearly understand the high-level semantics of a scene. How the significance of objects in an advanced semantic model is predicted has yet to be understood. A "semantic gap" still exists. To approach human-level prediction, many scholars have conducted useful exploration by reasoning the relative importance of image regions, and then learning higher-level features, such as emotion and body posture. With the deep learning framework gradually becoming mainstream, a variety of models with automatic learning features have also been produced in emotion-prediction tasks, such as multimodal learning models [46] and multitasking frameworks [47]. With the help of the attention mechanism, the classified emotional state (CES) or dimensional emotional space (DES) models can automatically learn the importance of different channels and improve robustness and accuracy. The emotion analysis model and saliency prediction model based on attention mechanisms can promote and integrate with each other.

3. Proposed Approach

To further explore the effect of high-level semantic information on saliency, we propose an improved multilayer network as the primary feature prediction network and use a subnetwork to determine the importance of different spaces and channels of an image, strengthening the possible saliency channels that contain high-level semantic information. The model uses a bottom-up method; that is, it adopts spatial and channel features and then refines them from top to bottom. The network obtains multilevel information in an end-to-end manner, effectively reducing computational cost while retaining important spatial and channel information. The overall network can be divided into two parts: a multilevel feature re-parameterization network and a semantic feature-aware network. The whole network is illustrated in Figure 2.

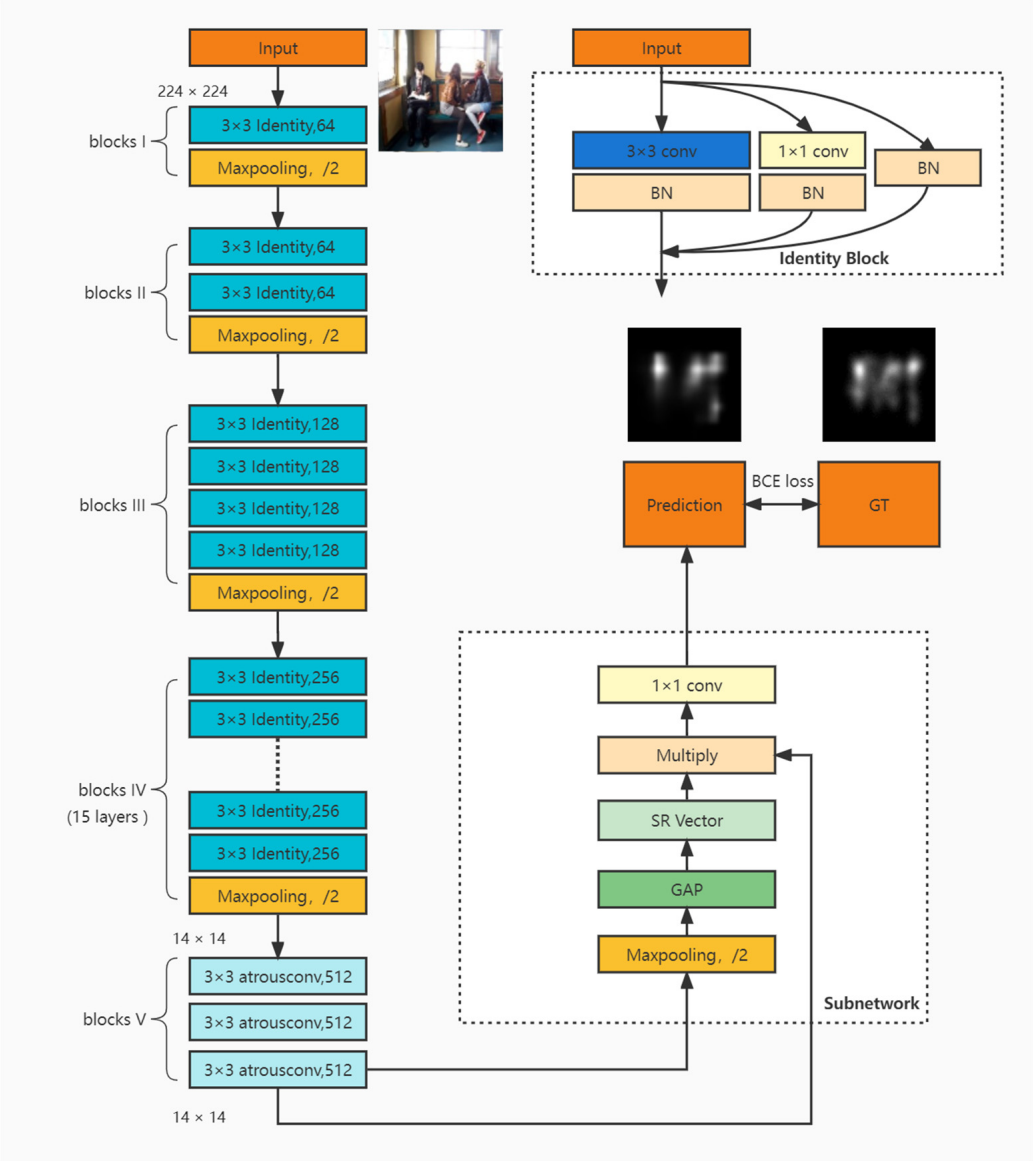

Figure 2. Overall network architecture.

3.1. Multilevel Feature Re-Parameterization Network

Inspired by the RepVGG [48] network, we added non-cross layers to the classical VGG basic block as our backbone network. Simultaneously, we optimized and adjusted it to adapt better to the saliency prediction task. On the basis of neurobiological research on memory and forgetting dependence, RepVGG uses a new lossless channel pruning network to simplify a CNN by reducing the number of output channels of the convolution layer. This procedure can equivalently convert the re-parameterized model into the original architecture with narrower layers, realizing structural sparsity and parameter reorganization. RepVGG is a lossless pruning model under an extremely high compression ratio.

Although ResNet, Inception, and other networks use a multi-branch structure to improve the performance of a network, such complex structures will affect inference speed because the branch results should be saved and resource consumption is large. DepthwiseConv, ShuffleNet, and other networks have increased memory consumption, but floating-point operations (FLOPs) are not directly proportional to speed. In addition, a multi-branch structure affects the flexibility of a network. For example, the shape of the input and output of the residual part of a residual structure must be the same to ensure the feasibility of the residual. The use of VGG-like networks has many advantages. Firstly,

a VGG network structure only includes 3 × 3 CNNs, a batch normalization (BN) layer, and a rectified linear unit (ReLU) activation function. Existing computing libraries and hardware are deeply optimized for 3 × 3 CNNs, while VGG is composed of a series of 3 × 3 CNNs, which exhibits evident speed advantages. However, VGG networks typically suffer from model degradation with the deepening of network depth. The deepening of a network makes training it difficult, and training errors will be initially reduced and then increased. The addition of an identity block to VGG-like networks has been proven to compensate for the shortcomings of a network, achieving better network performance [35]. Simultaneously, we optimized the whole network architecture to better play the role of the backbone network in the saliency prediction task.

We adopted a network similar to RepVGG-A0 to build the backbone network, which used a non-cross layer as a multi-branch identity block to replace the original CNN. The identity block consists of the CNN, the 1 × 1 branch, and the identity branch. We also replaced the convolutions of the head to better adapt to our saliency prediction task. As shown in Figure 2, the network is largely a five-block structure. The first layer uses a VGG style structure, including 1 identity block with 64 channels in Block I and 64 channels in Block II. The number of channels in Block III is changed to 128, including 4 layers of the identity block. Block IV uses 14 layers of the identity block, and the number of channels is changed to 256. The last layer uses a 3-layer atrous convolution with 512 channels. Our model adds multiple gradient flow paths to the network, which is equivalent to integrating multiple networks into one network and will be simpler and more efficient than other multiscale methods. The identity block is adopted during training, and this procedure is equivalent to the calculation made in a block as follows:

$$\mathrm{Out} = \mathrm{F}_1(\mathrm{X}) + \mathrm{F}_3(\mathrm{X}) + \mathrm{X} \tag{1}$$

where $\mathrm{F}_1(\mathrm{X})$ represents 1 × 1 convolution layers and $\mathrm{F}_3(\mathrm{X})$ indicates 3 × 3 convolution layers. The values before and after the identity block remain unchanged. The branch is equivalent to the special weight convolution layer, which is equivalent to using a convolution kernel with a weight of one to separate channels for convolution.

The convolution and BN layer structures are as follows:

$$\mathrm{Conv}(\mathrm{x}) = \mathrm{w}(\mathrm{x}) + \mathrm{b} \tag{2}$$

$$\mathrm{BN}(\mathrm{x}) = \alpha * \frac{(\mathrm{x} - \mathrm{E}(\mathrm{x}))}{\sqrt{\mathrm{S}}} + \beta \tag{3}$$

The whole fusion result can be expressed as

$$\mathrm{BN}(\mathrm{Conv}(\mathrm{x})) = \mathrm{w_f}(\mathrm{x}) + \mathrm{b_f} \tag{4}$$

where $\mathrm{w_f}(\mathrm{x}) = \frac{\alpha * \mathrm{w}(\mathrm{x})}{\sqrt{\mathrm{S}}}$, $\mathrm{b_f} = \frac{\alpha * (\mathrm{b} - \mathrm{E}(\mathrm{x}))}{\sqrt{\mathrm{S}}} + \beta$, and α and β are learnable parameters introduced by the BN layer. S is the variance and $\mathrm{E}(\mathrm{x})$ is the mean value. The identity block can be integrated well into the main network. The integrated structure is same as the original CNN layer. It can efficiently capture more robust features and deal with the gradient disappearance problem in the deep layer of the network. In model inference, the three-branch convolution layer and the subsequent BN layer can be equivalently transformed into a convolution layer with bias. After the obtained 1 × 1 convolution kernel padded into 3 × 3, the convolution kernel and bias obtained by the three branches are added, respectively. In this way, the trained model can be equivalently transformed into a one-way model with only 3 × 3 convolution layers and finally realize "re-parameterization", which can take advantage of the high performance of the multi-branch model in training and the advantages of fast speed and memory saving of the one-way model in inference.

Given the particularity of the saliency prediction task, the input image is considerably scaled during the downsampling of layers. If the pre-trained RepVGG-A0 network is used, then the input is 224 × 224. After five layers of maximum pooling, it will be reduced

to 7×7. For the saliency prediction task, excessively small feature maps may reduce prediction accuracy. Therefore, we adjusted the last block, removed the maximum pool layer, and increased the output resolution. A three-layer stacked atrous convolution with two holes and 512 channels was used. Atrous convolutions can efficiently offset the loss of spatial information caused by the pooling layer. Combined with 512 channel convolution layers, the network can obtain a better receptive field without losing the size of the characteristic image to better capture image spatial information. The module uses a dense structure to approximate a sparse CNN, allowing the network to use a larger number of channels without increasing the amount of computation. Through the preceding method, our saliency mapping is adjusted to 14×14 instead of 7×7. Simultaneously, our model obtains a better receptive field and avoids accuracy loss. The features extracted from the primary network are sent to the semantic feature-aware and merging networks.

3.2. Semantic Feature-Aware Network

In emotion recognition models, information weights of different levels are obtained through squeeze, excitation, or attention modes [49] to form the emotion feature vector for emotion classification or intensity discrimination. Referring to a variety of emotion classification models, we also use a subnetwork to evaluate and extract high-level semantic information. In contrast with the emotion classification model, our model's task is to generate a saliency map rather than emotion discrimination.

The features obtained from the last layer of the primary network are sent to the subnetwork. After the features are sent to the max-pooling layer to reduce feature dimension and spatial variance, we use the global average pooling layer to compress the extracted spatial information into a vector, generating a semantic representation vector (SRV) for extracting high-level semantic information on the basis of channel enhancement. The global average pooling layer can regularize the structure of the whole network, prevent overfitting, eliminate the characteristics of black boxes in the full connection layer, and provide practical saliency to each channel. Simultaneously, our parameters are reduced by 80% compared with some full connection layer models [50,51].

An SRV can learn its relative weights in accordance with the spatial position or semantic features of different objects or regions in a scene, change the feature intensity of different channels in a saliency map, and find the region of interest. Assuming that the spatial information saliency map extracted by the primary network is F as a whole, the N-dimensional information is compressed into a vector V by using the global average pooling layer, which can be expressed as:

$$V_C = \frac{1}{N*N} \sum_{i=1}^{N} \sum_{j=1}^{N} F_C(i,j) \tag{5}$$

where V represents the generated $1 \times 1 \times C$ vector. We use the channel enhancement vector as a weighting module to multiply the saliency map of the primary network to obtain the final saliency map. We use the sigmoid activation function to weigh each channel and perform weighted merging through the 1×1 convolution layer of the last layer. The diagram of the combined features is illustrated as follows:

$$M = \frac{1}{C} \sum_{i=1}^{C} \text{sigmoid}(F_C \times V_C \times \text{Relu}(W_C)) \tag{6}$$

We use the binary cross-entropy (BCE) loss function for network training. The pixel-level prediction of a saliency map can be understood as a classification problem with a gray value of 255. The sigmoid layer quantifies 255 to be between 0 and 1. In this chapter, we use BCE as the loss function. In contrast with the mean squared error, which focuses on the difference between prediction probability and real probability in all categories, BCE focuses on the prediction probability of the correct category, and thus, it exhibits the advantage of fast convergence. The BCE formula is as follows:

$$C = -\frac{1}{N} \sum_{i}^{N} y \log a + (1-y) \log(1-a) \tag{7}$$

where a is the predicted value given by the CNN and y is the label that corresponds to the true image pixel value in saliency prediction.

Finally, we obtain the saliency map with a size of 14×14 and restore it to the size of the input image through bilinear upsampling. The whole network uses the bottom-up method to segment features and then refines them from top to bottom to combine information from shallow to deep. Compared with the original structure, our model retains important space and channel information.

4. Experiments and Analysis

We used the SALICON [33] training dataset to train our saliency model and two image datasets (Emod [51] and Cat2000 [52]) with rich semantic information to evaluate our saliency model. The SALICON dataset is a large database that contains context saliency in images selected from Microsoft common objects in context, including 10,000 images for training and 5000 images for testing. The Cat2000 dataset contains 2000 images under 20 different categories, ranging from natural images to indoor and outdoor scenes, cartoons, and emotions. Different categories of images are suitable for a variety of attention behavior research. The Emod dataset contains 1019 positive and negative emotional images, including 4302 targets with fine contour, emotional tags, and semantic tags.

Our model uses the first four blocks of pre-trained parameters initialized as RepVGG-A0, which trained 40 epochs on the SALICON training set, with a momentum of 0.9, a weight decay of 0.0002, and an initial learning rate of 10^{-4}. Binary cross entropy is used as the loss function, the random gradient descent training image is used in end-to-end parameter learning, and Pytorch is adopted as the primary framework to train on NVIDIA Titan X 3090Ti GPU.

4.1. Evaluation Measures

Measures for visual saliency prediction are mostly used to evaluate the similarity and difference between saliency maps and ground truth (GT), and then output an evaluation score to evaluate the degree of similarity or difference between them. Given a set of true values to define the scoring function, the saliency prediction chart can be used as input and then returned to evaluate prediction accuracy. Considering different GTs, many metrics are used to evaluate the saliency prediction model. Firstly, the most widely used location-based measure is the area under the curve (AUC), which can be used as a binary classifier. We used its variant, called AUC-Judd, which uses uniform random sampling of non-concerns to calculate the false positive rate, reducing the effect of center deviation. Although AUC is widely used as an important criterion, it cannot distinguish the relative importance of different regions. Therefore, we also adopt three of the most commonly used similarity evaluation measures based on distribution, namely, NSS, CC, and earth mover distance (EMD). Their descriptions are as follows:

1. NSS can represent consistency between mappings, taking the average value of $\overline{P}$ at point Q of human eye attention, where n represents the total number of human eye fixation, $\overline{P}$ represents the unit normalized saliency map P, i represents the ith pixel, and N is the total number of pixels at the fixation point. NSS value is negatively correlated with model performance.

$$NSS = \frac{1}{N} \sum_{i=1}^{N} \overline{P}(i) \times Q(i) \tag{8}$$

2. Linear CC is a statistical metric for measuring the linear correlation between two random variables. For the prediction and evaluation of saliency, a prediction saliency map (P) and a ground truth density map (G) are regarded as two random variables. Then, the calculation formula of CC is:

$$CC = \frac{\mathrm{cov}(P, G)}{\sigma(P) \times \sigma(G)} \tag{9}$$

where cov is the covariance and σ is the standard deviation. CC can equally punish false positive and false negative, with a value range of (−1, 1). When the value is close to both ends, the model performs better.

3. EMD represents the distance between the two 2D maps, G and S. It is the minimum cost of transforming the probability distribution of the estimated saliency map S into the probability distribution of the GT. Therefore, a lower EMD corresponds to a high-quality saliency map. In the field of saliency prediction, EMD represents the minimum cost of converting the probability distribution of a saliency map into one of a human eye attention map.

4.2. Experimental Results and Analysis

To evaluate the performance of our model comprehensively, we use several classical and deep models for comparison. Three of the models are typical bottom-up methods, including two classical models: graph-based visual saliency (GBVS) [53], IttiKoch 2 [15], and the boolean graph-based saliency model (BMS) [54]. Three DNN models with superior performance are SALICON [33], SAM-ResNet [38], and EML-Net [40]. All networks have no additional center bias mechanism. The experimental results are shown in Figure 3 and discussed below.

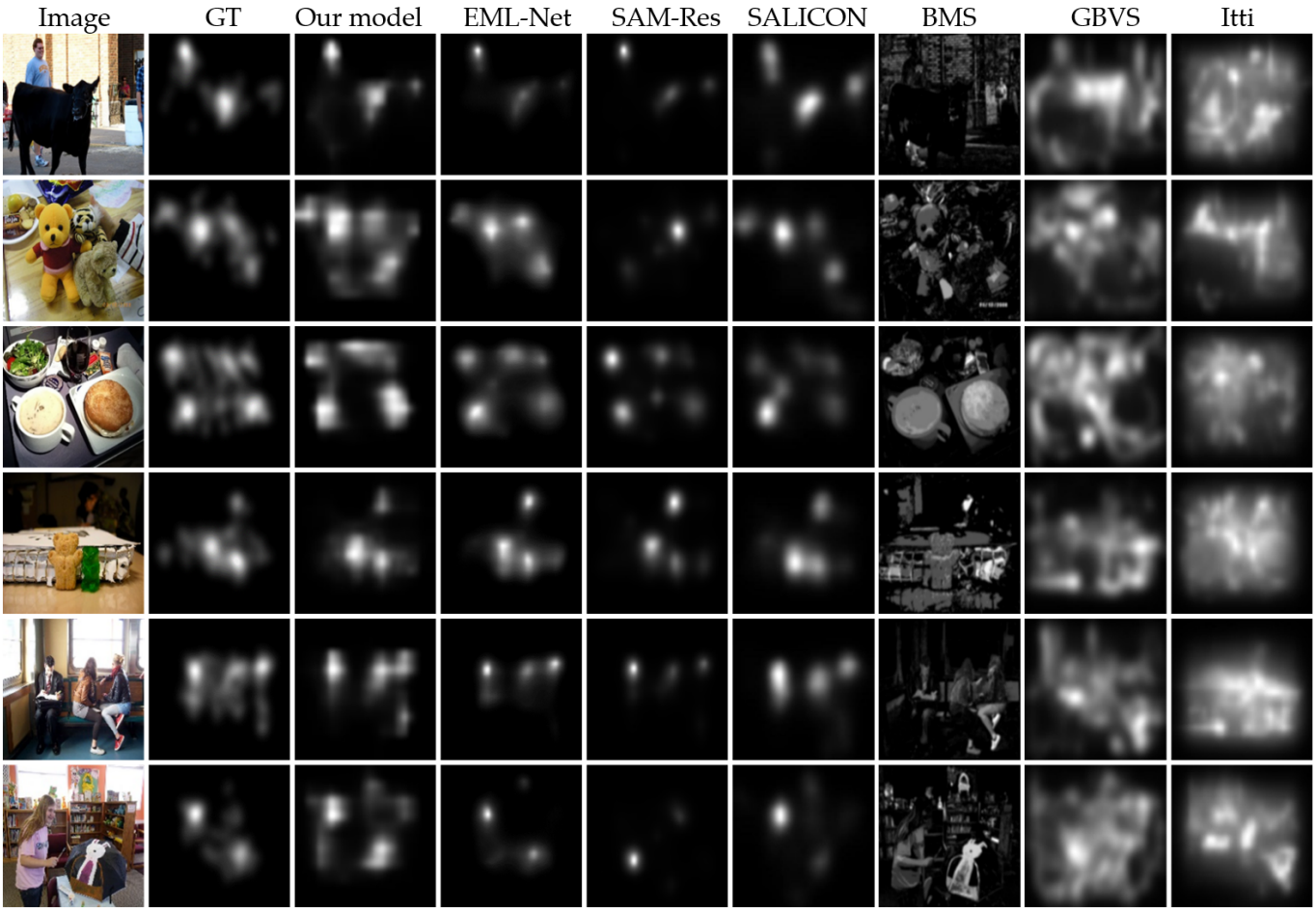

Figure 3. Saliency maps of various models on the SALICION validation dataset.

Tables 1 and 2 list the quantitative evaluation results of the model on the Cat2000 and Emod datasets, respectively. The best scores are marked in bold. The model we used achieved the best overall performance across datasets without additional center bias mechanism, that probably because these datasets have more semantic and emotional content than the other datasets. Our model can capture more relatively important features and exhibits an advantage in datasets that are rich in these features. The performance based on all the metrics is better than those of the other deep learning models, considerably exceeding the performance of classical models. Among them, the score of metrics is similar to SAM-ResNet and EML-Net on the basis of the Cat2000 dataset, while those of NSS, CC, and EMD are higher than SAM-ResNet by about 1.14%, 2.23%, and 1.75% and higher than EML-Net by about 0.56%, 2.23%, and 0.89% on the basis of the Emod dataset. This may be due to the richer contexture and emotional information in Emod. These scores are considerably higher than those of the classical models. Although the improvement of our model is limited compared with EML-Net, our model is simpler and the number of parameters is greatly reduced (from 23.5 M to 14.8 M).

Table 1. Quantitative results of evaluation measures on the Cat2000 validation dataset.

Metric	AUC-Judd	NSS	CC	EMD
SALICON	0.86	2.18	0.79	1.13
SAM-ResNet	**0.88**	2.38	**0.89**	**1.04**
EML-Net	0.87	2.38	0.88	1.05
IttiKoch2	0.77	1.06	0.42	3.44
GBVS	0.80	1.23	0.50	2.99
BMS	0.78	1.16	0.39	1.95
Our Model	**0.88**	**2.39**	**0.89**	1.05

Table 2. Quantitative results of evaluation measures on the Emod validation dataset.

Metric	AUC-Judd	NSS	CC	EMD
SALICON	**0.87**	1.59	0.84	1.32
SAM-ResNet	**0.87**	1.74	0.86	1.14
EML-Net	**0.87**	1.75	0.86	1.13
IttiKoch2	0.73	0.98	0.39	3.2
GBVS	0.79	1.18	0.47	2.92
BMS	0.77	1.12	0.49	2.06
Our Model	**0.87**	**1.76**	**0.88**	**1.12**

NSS, CC, and EMD consider the relative importance of saliency regions. They are important metrics for evaluating the roles of context and semantic information in saliency prediction. These metrics for our model are better than the other methods in the two datasets, demonstrating the advantage of the subnetwork in distinguishing the relative importance of saliency regions. As discussed in Section 1, people tend to focus on human and human-related actions or objects and regard them as saliency goals. Simultaneously, these goals are frequently high-level factors rich in semantic information and often have high saliency values. During the training process, the advanced feature detector as a subnetwork can correct the feature detector one by one and activate the feature channels of these advanced areas.

5. Model Visualization and Ablation Analysis

To better verify the role of each part of our network, we analyzed our model on the Cat2000 validation dataset. We separately removed the identity block and subnetwork to compare with the overall network and conducted joint training with the same loss and input to verify the effect of the identity block and subnetwork on the performance of the model. We divided the model structure into three parts: basic VGG-like model without

branch (B), identity block (I), and subnetwork (S). The visualization results of the model are presented in Figure 4.

1. Influence of identity block: Our backbone network uses a non-cross layer to capture spatial features that are equal to multiscale networks. Multiscale models have been widely used in recent years. Multiscale features and residual blocks are considered the key elements of the saliency prediction task that can further improve the performance of a saliency model. To verify the similar effect of the backbone network in our model, we compared the whole network (B + I + S) with the basic network (B) and basic network with subnetwork (B + S). For each image, we calculated the score of the metrics. As shown in Table 3, the non-cross layer can significantly improve the performance of the model. Also, it can greatly reduce the over-fitting phenomenon that may occur, and the network is too sensitive to detect error saliency areas (e.g., fourth line). Our model also exhibits some advantages in parameters and reasoning time because we used the re-parameterization network instead of directly using the multiscale network. Compared with other models (EML-Net: 23.5 M and SAM-ResNet50: 70.1 M), the parameters of our model (14.8 M) decreased significantly.
2. Influence of subnetwork: Similar architectures that use emotion or semantic features can act well on emotion or semantic priority and predict the relative importance of an image area by enhancing the ability of channel weighted subnetworks. To illustrate this phenomenon, we calculated the same difference score between our model (B + I + S) and the basic network without subnetwork (B + I) predictions. By correlating the difference of each model with the authenticity of background in the image, the degree of relative saliency of the human-related object predicted by the model was evaluated (e.g., the computer in the first line). As indicated in Table 3, a large correlation shows that the model performed better in predicting relative saliency. The best scores are marked in bold.

Table 3. Quantitative results of model ablation on the Cat2000 validation dataset.

Metric	AUC-Judd	NSS	CC	EMD
B + I + S	**0.88**	**2.39**	**0.89**	**1.05**
B + S	0.80	2.11	0.81	1.33
B + I	0.79	2.03	0.85	1.43
B	0.75	1.85	0.71	2.03

These observations demonstrate that the subnetwork can obtain important information in different regions, and this condition is more evident in the overall model. A multi-branch network is more successful in improving the gradient disappearance problem of the primary network and realizing a certain multiscale function to achieve better network characteristics. Although the channel-increasing vector obtained from the subnetwork is used for emotion classification in the emotion model, the ability of the channel weighted subnetwork is not limited to emotion priority, but it can predict the relative importance of the object to avoid missing some important areas. The main network and sub-network can cooperate to find the saliency area while avoiding overfitting. Although we have achieved some success, our prediction results still have some problems. When the image is more complex and has many objects, model metrics decrease and some parts are not detected, or a certain deviation exists in detection. These phenomena may be due to judging only from the relative importance of channels but cannot really start from the semantic perspective of high level.

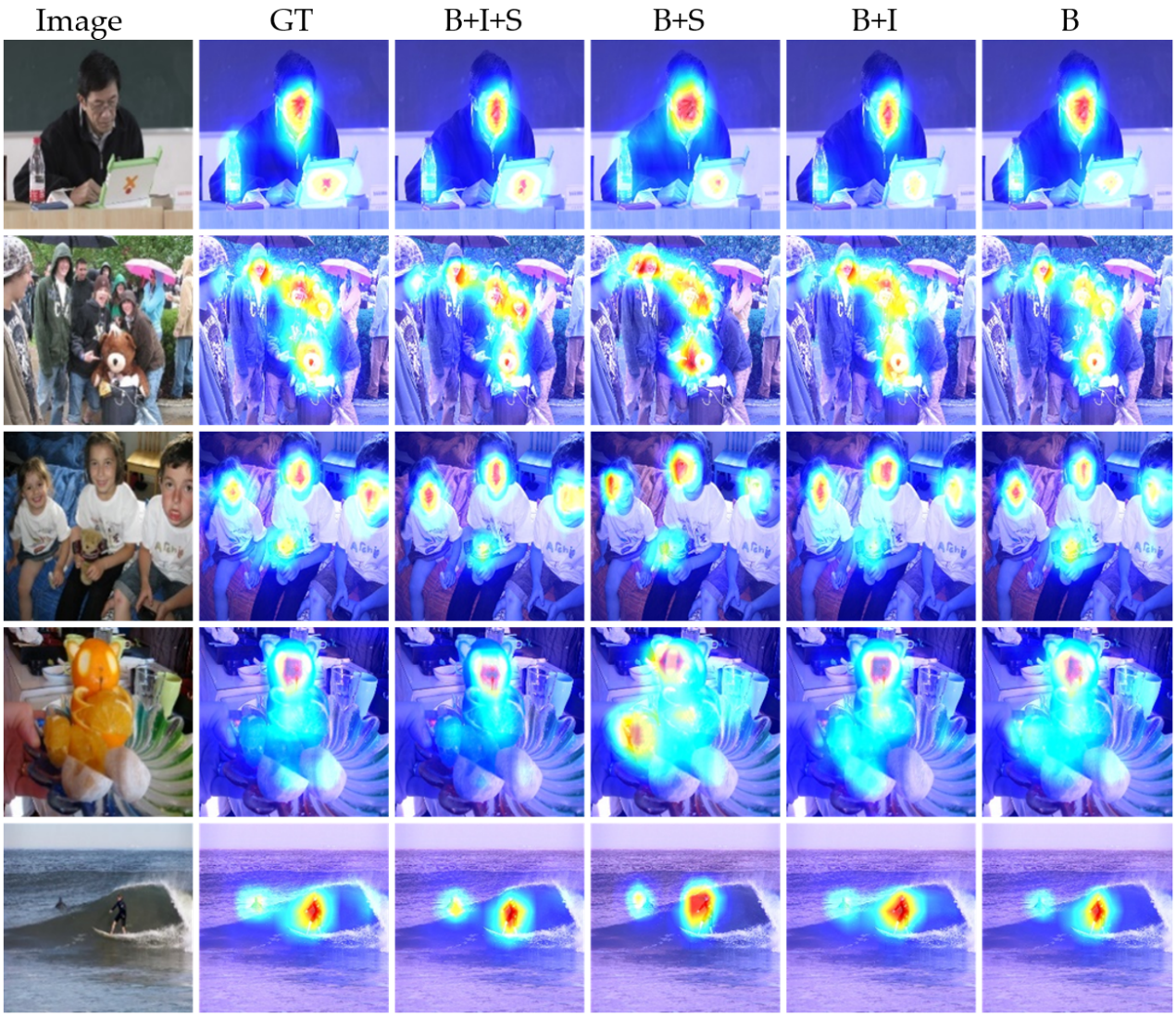

Figure 4. Visualization results of model ablation test on the Cat2000 dataset. We quantized the image saliency from 0 to 1. Approaching 0 is blue and approaching 1 is red.

6. Conclusions

With the continuously improving performance of saliency prediction models and the gradual saturation of evaluation measures, researchers have begun to look for higher-level concepts to make saliency prediction closer to human-level performance. In this work, we discuss the role of these features in saliency prediction, design a new DNN model to simulate human attention effectively in complex scenes, and quantify the relationship between high-level semantic information and visual attention. To detect the relative importance of prominent areas, we use two image datasets with rich semantic features to quantitatively investigate. Through experiments, we prove that high-level semantic features exhibit a strong correlation with saliency prediction and given priority in saliency maps. Our model combines the lightweight main network and semantic feature-aware network, which reduces the consumption of computing resources and achieves good results.

Author Contributions: Conceptualization, F.Y. and R.X.; investigation, F.Y.; resources, S.Q.; writing—original draft preparation, F.Y.; writing—review and editing, R.X.; supervision, Z.W.; project administration, R.X.; funding acquisition, R.X. and Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62176268), the Non-profit Central Research Institute Fund of the Chinese Academy of Medical Sciences (2020-JKCS-008), the Major Science and Technology Project of Zhejiang Province Health Commission (WKJ-ZJ-2112), and the Fundamental Research Funds for the Central Universities (FRF-BD-20-11A).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sziklai, G.C. Some studies in the speed of visual perception. *Inf. Theory IRE Trans.* **1956**, *76*, 125–128. [CrossRef]
2. Koch, K.; Mclean, J.; Segev, R.; Freed, M.A.; Michael, I.I.; Balasubramanian, V.; Sterling, P. How Much the Eye Tells the Brain. *Curr. Biol.* **2006**, *16*, 1428–1434. [CrossRef]
3. Yan, F.; Chen, C.; Xiao, P.; Qi, S.; Wang, Z.; Xiao, R. Review of Visual Saliency Prediction: Development Process from Neurobiological Basis to Deep Models. *Appl. Sci.* **2021**, *12*, 309. [CrossRef]
4. Wang, W.; Shen, J.; Cheng, M.M.; Shao, L. An Iterative and Cooperative Top-Down and Bottom-Up Inference Network for Salient Object Detection. In Proceedings of the CVPR19, Long Beach, CA, USA, 16–20 June 2019.
5. Wang, W.; Lai, Q.; Fu, H.; Shen, J.; Yang, R. Salient Object Detection in the Deep Learning Era: An In-depth Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, 1448–1457. [CrossRef]
6. Zhang, P.; Liu, W.; Zeng, Y.; Lei, Y.; Lu, H. Looking for the Detail and Context Devils: High-Resolution Salient Object Detection. *IEEE Trans. Image Processing* **2021**, *30*, 3204–3216. [CrossRef]
7. Fan, L.; Wang, W.; Huang, S.; Tang, X.; Zhu, S.C. Understanding Human Gaze Communication by Spatio-Temporal Graph Reasoning. *arXiv* **2019**, arXiv:1909.02144.
8. Aksoy, E.; Yazc, A.; Kasap, M. See, Attend and Brake: An Attention-based Saliency Map Prediction Model for End-to-End Driving. *arXiv* **2020**, arXiv:2002.11020.
9. Chen, C.; Zhou, K.; Zha, M.; Qu, X.; Guo, X.; Chen, H.; Wang, Z.; Xiao, R. An effective deep neural network for lung lesions segmentation from COVID-19 CT images. *IEEE Trans. Ind. Inform.* **2021**, *17*, 6528–6538. [CrossRef]
10. Chen, C.; Xiao, R.; Zhang, T.; Lu, Y.; Guo, X.; Wang, J.; Chen, H.; Wang, Z. Pathological lung segmentation in chest CT images based on improved random walker. *Comput. Methods Programs Biomed.* **2021**, *200*, 105864. [CrossRef]
11. Jia, Z.; Lin, Y.; Wang, J.; Wang, X.; Xie, P.; Zhang, Y. SalientSleepNet: Multimodal Salient Wave Detection Network for Sleep Staging. *arXiv* **2021**, arXiv:2105.13864.
12. O'Shea, A.; Lightbody, G.; Boylan, G.; Temko, A. Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture. *arXiv* **2021**, arXiv:2105.13854. [CrossRef]
13. Treisman, A.M.; Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **1980**, *12*, 97–136. [CrossRef]
14. Koch, C.; Ullman, S. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Hum Neurobiol.* **1987**, *4*, 219–227.
15. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
16. Ma, Y.-F.; Zhang, H.-J. Contrast-based image attention analysis by using fuzzy growing. In Proceedings of the Eleventh ACM International Conference on Multimedia, Berkeley, CA, USA, 2–8 November 2003; pp. 374–381.
17. Feng, J.; Wei, Y.; Tao, L.; Zhang, C.; Sun, J. Salient object detection by composition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1028–1035.
18. Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; Shum, H.-Y. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 353–367.
19. Borji, A.; Itti, L. Exploiting local and global patch rarities for saliency detection. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 July 2012; pp. 478–485.
20. Zhi, L.; Zhang, X.; Luo, S.; Meur, O.L. Superpixel-Based Spatiotemporal Saliency Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *24*, 1522–1540.
21. Huang, G.; Pun, C.M.; Lin, C. Unsupervised video co-segmentation based on superpixel co-saliency and region merging. *Multimed. Tools Appl.* **2016**, *76*, 12941–12964. [CrossRef]
22. Oliva, A.; Torralba, A.; Castelhano, M.S.; Henderson, J.M. Top-down control of visual attention in object detection. In Proceedings of the International Conference on Image Processing, Barcelona, Spain, 14–17 September 2003; pp. I-253–I-256.
23. Xie, Y.; Lu, H.; Yang, M.H. Bayesian Saliency via Low and Mid Level Cues. *IEEE Trans. Image Process. A Publ. IEEE Signal Process. Soc.* **2013**, *22*, 1689–1698.
24. Gao, D.; Vasconcelos, N. Discriminant saliency for visual recognition from cluttered scenes. *Adv. Neural Inf. Process. Syst.* **2004**, *17*, 481–488.
25. Gao, D.; Han, S.; Vasconcelos, N. Discriminant Saliency, the Detection of Suspicious Coincidences, and Applications to Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 989–1005.
26. Vig, E.; Dorr, M.; Cox, D. Large-scale optimization of hierarchical features for saliency prediction in natural images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2798–2805.

27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
28. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
29. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
30. Kümmerer, M.; Theis, L.; Bethge, M. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv* **2014**, arXiv:1411.1045.
31. Kruthiventi, S.; Ayush, K.; Babu, R.V. DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations. *IEEE Trans. Image Process.* **2017**, *26*, 4446–4456. [CrossRef]
32. Kümmerer, M.; Wallis, T.S.; Bethge, M. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv* **2016**, arXiv:1610.01563.
33. Jiang, M.; Huang, S.; Duan, J.; Zhao, Q. Salicon: Saliency in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1072–1080.
34. Pan, J.; Sayrol, E.; Giro-i-Nieto, X.; McGuinness, K.; O'Connor, N.E. Shallow and deep convolutional networks for saliency prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 598–606.
35. Jetley, S.; Murray, N.; Vig, E. End-to-end saliency mapping via probability distribution prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5753–5761.
36. Liu, N.; Han, J. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE Trans. Image Process.* **2018**, *27*, 3264–3274. [CrossRef]
37. Cornia, M.; Baraldi, L.; Serra, G.; Cucchiara, R. A deep multi-level network for saliency prediction. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 3488–3493.
38. Marcella, C.; Lorenzo, B.; Giuseppe, S.; Rita, C. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Trans. Image Process.* **2016**, *27*, 5142–5154.
39. Pan, J.; Ferrer, C.C.; McGuinness, K.; O'Connor, N.E.; Torres, J.; Sayrol, E.; Giro-i-Nieto, X. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv* **2017**, arXiv:1701.01081.
40. Jia, S.; Bruce, N.D. Eml-net: An expandable multi-layer network for saliency prediction. *Image Vis. Comput.* **2020**, *95*, 103887. [CrossRef]
41. Wang, W.; Shen, J. Deep visual attention prediction. *IEEE Trans. Image Process.* **2017**, *27*, 2368–2378. [CrossRef]
42. Gorji, S.; Clark, J.J. Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 2510–2519.
43. Wang, W.; Shen, J.; Xie, J.; Cheng, M.-M.; Ling, H.; Borji, A. Revisiting Video Saliency Prediction in the Deep Learning Era. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 220–237. [CrossRef]
44. Lai, Q.; Wang, W.; Sun, H.; Shen, J. Video Saliency Prediction using Spatiotemporal Residual Attentive Networks. *IEEE Trans. Image Process.* **2019**, *29*, 1113–1126. [CrossRef]
45. Xu, M.; Yang, L.; Tao, X.; Duan, Y.; Wang, Z. Saliency Prediction on Omnidirectional Image With Generative Adversarial Imitation Learning. *IEEE Trans. Image Process.* **2021**, *30*, 2087–2102. [CrossRef]
46. Pang, L.; Zhu, S.; Ngo, C.-W. Deep multimodal learning for affective analysis and retrieval. *IEEE Trans. Multimed.* **2015**, *17*, 2008–2020. [CrossRef]
47. Yang, J.; She, D.; Sun, M. Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network. In Proceedings of the IJCAI, Melbourne, Australia, 19–25 August 2017; pp. 3266–3272.
48. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13733–13742.
49. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
50. Cordel, M.O.; Fan, S.; Shen, Z.; Kankanhalli, M.S. Emotion-aware human attention prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4026–4035.
51. Fan, S.; Shen, Z.; Jiang, M.; Koenig, B.L.; Xu, J.; Kankanhalli, M.S.; Zhao, Q. Emotional attention: A study of image sentiment and visual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7521–7531.
52. Borji, A.; Itti, L. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv* **2015**, arXiv:1505.03581.
53. Harel, J.; Koch, C.; Perona, P. Graph-based visual saliency. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 4–9 December 2006; Volume 19.
54. Zhang, J.; Sclaroff, S. Exploiting surroundedness for saliency detection: A boolean map approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 889–902. [CrossRef]