IEICE TRANS. INF. & SYST., VOL.E106–D, NO.9 SEPTEMBER 2023

PAPER

Siamese Transformer for Saliency Prediction Based on Multi-Prior Enhancement and Cross-Modal Attention Collaboration

Fazhan YANG[†], **Xingge GUO**[†a)], ***Nonmembers***, **Song LIANG**[†], ***Member***, **Peipei ZHAO**[†], ***and*** **Shanhua LI**[†], ***Nonmembers***

SUMMARY Visual saliency prediction has improved dramatically since the advent of convolutional neural networks (CNN). Although CNN achieves excellent performance, it still cannot learn global and long-range contextual information well and lacks interpretability due to the locality of convolution operations. We proposed a saliency prediction model based on multi-prior enhancement and cross-modal attention collaboration (ME-CAS). Concretely, we designed a transformer-based Siamese network architecture as the backbone for feature extraction. One of the transformer branches captures the context information of the image under the self-attention mechanism to obtain a global saliency map. At the same time, we build a prior learning module to learn the human visual center bias prior, contrast prior, and frequency prior. The multi-prior input to another Siamese branch to learn the detailed features of the underlying visual features and obtain the saliency map of local information. Finally, we use an attention calibration module to guide the cross-modal collaborative learning of global and local information and generate the final saliency map. Extensive experimental results demonstrate that our proposed ME-CAS achieves superior results on public benchmarks and competitors of saliency prediction models. Moreover, the multi-prior learning modules enhance images express salient details, and model interpretability.

key words: *saliency prediction, Siamese transformer, multi-prior, cross-modal*

1. Introduction

With the continuous advancement of Internet technology, massive amounts of video and image data are generated every day. How to quickly obtain useful information from these images and videos has become an increasingly pressing issue. The human visual system has the characteristics of visual attention. In complex external scenes, it can ignore the interference information and quickly perceive and process critical areas [1]. This mechanism has significant meaning for us to sort out the information that are needed or interested in from a large amount of external information. Inspired by visual attention, people have introduced this visual attention into the field of computer vision, and a large number of visual saliency methods for predicting human eye attention have emerged. Image processing based on this visual attention mechanism can better allocate limited computer resources to interesting targets, and has made good progress in various applications, such as image segmentation [2], [3], target recognition and detection [4], [5], image compression [6], [7] and visual quality assessment [8], [9], etc.

In the past two decades, many visual saliency prediction methods have been proposed, and the prediction effect has been gradually improved, but there are still many problems. Traditional saliency prediction methods [10]–[12] mostly use bottom-up methods driven by data or stimuli to extract low-level information, such as texture, spectrum, and color, to find salient regions of images. Traditional methods are characterized by simplicity, intuition, ease of design, and a lack of complexity, so early research mostly used bottom-up methods [13]–[15]. However, the lower-level information cannot understand the structure, position, and semantic information contained in the image, which limits the feature representation ability of saliency prediction, and it is very difficult to predict saliency in complex scenes.

Convolutional neural networks have achieved significant success in computer vision [16]–[18]. Its powerful feature extraction and expression capabilities have further improved the performance of saliency prediction algorithms, and then saliency prediction algorithms based on deep learning have gradually become mainstream. However, neural networks also have many shortcomings, perhaps the most well-known of which is their "black box" nature, which means that without knowing how and why a neural network produces a certain output, what is the cause to this prediction. Lack of interpretability. At the same time, convolution operation lacks the global understanding of image, cannot model the dependency between features, and cannot make full use of context information. The convolution operation based solely on the sliding window makes the significant prediction of the image's global contrast model ineffective, which is very important for the significant detection.

Compared with CNN, the self-attention mechanism of transformer [19] is not limited by local interactions, and can not only mine long-distance dependencies but also perform parallel calculations. It has also achieved remarkable success in computer vision tasks. However, the pixels in the image have a very high resolution. The computational complexity of the transformer used in the visual field is the square of the image scale, which will lead to an enormous amount of calculation. Many models will reduce the size of the image to reduce the amount of calculation, but this useful information will be lost. The Swin Transformer [20] model is an improved version proposed by Microsoft based

Manuscript received December 15, 2022.
Manuscript revised April 30, 2023.
Manuscript publicized June 20, 2023.
[†]The authors are with School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China.
a) E-mail: guoxingge@163.com (Corresponding author)
DOI: 10.1587/transinf.2022EDP7220

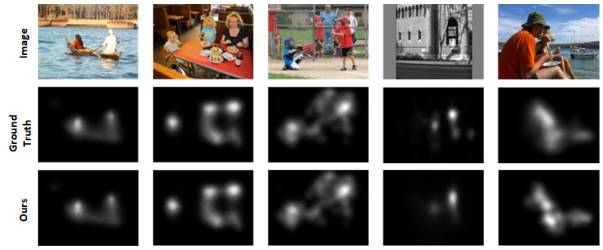

Fig. 1 Example of visual saliency prediction. The first row shows natural images that are recognizable by the human eye. The second row is the truth map. The third row is the saliency map predicted by our method.

on the transformer model. Transformer not only focuses on global modeling of information, but also proposes a hierarchical network structure to solve the multi-scale problem of visual images using sliding window operations, thereby reducing the computational complexity of transformer.

Modeling based on prior knowledge is helpful for learning the details of the underlying features. Biological vision studies have shown that the visual system is more sensitive to the contrast of the received information. Therefore, people will pay more attention to the areas in image that have a higher contrast than the surrounding objects [21]. The filter response can well reflect the salient regions of the image. In addition, people generally place objects in the center of the image or slightly off-center. That is, the central part of the image is often the most salient area.

Based on this, we propose a Siamese transformer saliency prediction model with multi-prior augmentation and cross-modal attention synergy. Transformer extracts the context information of the image under the self-attention mechanism, simulates the focus of human perception, concentrates on the salient area, and obtains the global saliency map. At the same time, it combines prior knowledge and neural networks to obtain high-level local saliency information, and guides cross-modal collaborative learning of saliency features through attention calibration, thereby optimizing the saliency features of the entire scene.

In summary, the main contributions of this paper are as follows:

1) We propose a multi-prior knowledge module. In complex contexts, a single prior does not perform well. Based on the contrast prior and frequency prior, we consider that people generally pay more attention to the center area, so we add the center prior knowledge module. It enhances the salient details of the image and also enhances the interpretability of the model.

2) We propose a cross-modal attention collaborative Siamese network, which combines prior knowledge with deep learning. By using the self-attention mechanism, the first transformer branch extracts the global features of the image, and the second Siamese branch enhances the image significantly based on prior knowledge. Local detail feature learning, cross-modal collaborative learning of global and local information under the guidance of the attention calibration module.

3) Our ME-CAS model is verified on three mainstream saliency prediction datasets. Experiments show that the ME-CAS algorithm is superior to other mainstream algorithms on multiple evaluation metrics, and its effectiveness has been verified.

2. Related Work

In this section, we will review the relevant research results of saliency detection and the development of transformer, and give a brief overview of related methods. Visual saliency detection is mainly divided into two types, one is to predict the focus of human eyes, and the other is to detect salient objects. Both have their research value. The work of this paper is mainly to predict the focus of human eyes.

2.1 Visual Saliency Prediction

The traditional saliency model detection mainly uses the characteristics of the image such as brightness, edge, and contrast, and considers the difference between the pixel and the surrounding neighborhood in terms of features, to calculate the saliency map of the pixel. Due to the widespread application of deep learning techniques, saliency prediction has progressed greatly compared to traditional methods. Vig et al. proposed an image saliency detection model eDN based on a convolutional neural architecture in [22]. The model uses CNN to output feature vector maps, and these feature vectors are combined and then input into the linear SVM classifier. These feature vectors are trained by supervised learning to obtain the prediction results of the saliency map.

Kummerer proposed DeepGaze I [23], the first model that applies transfer learning to the saliency field. It uses the features of the trained ImageNet object detection AlexNet [18] network to train the model of human eye focus. The prediction effect of the eDN model was obtained, and then the model was improved, and VGG19 [24] was used for feature extraction. Since then, there have been a lot of models about human eye focus detection like mushrooms after rain.

The DeepFix [25] network utilizes a large convolutional layer of different scales of the receptive field to capture the semantic information in the picture and introduces a convolutional layer with a location bias (LBC) to simulate the central bias of people seeing images. Pan et al. proposed SalGAN [26], a deep network for saliency prediction trained with adversarial examples. Like all other generative adversarial networks, it consists of two modules, a generator and a discriminator, which work together to generate saliency maps. The SAM-VGG [27] model breaks through the standard method of saliency prediction using a feed-forward network to calculate fixation maps and proposes an accurate saliency prediction model that combines neural attention mechanisms. The EML-NET [28] model introduces a scalable approach to combine multiple deep convolutional networks of arbitrary complexity as encoders for

visual saliency-related features.

The DeepGaze IIE [29] model discusses the contribution of various backbones to saliency detection, and it is found that cascading and fusing multiple backbone networks pre-trained on ImageNet can effectively improve the performance of the saliency detection model. TranSalNet [30] is the first saliency model combining CNN and transformer. Based on the CNN architecture, it uses transformer's self-attention mechanism to learn feature information for feature maps of different scales, thereby generating feature maps. However, the computational complexity of the classic transformer module it utilizes is the square of the input scale, which will cause huge computational overhead. Therefore, the model will reduce the size of the image to reduce the computational load or apply it to smaller features, but this will lose useful information.

Recently, the heterogeneity of saliency maps across different subjects has attracted the attention of researchers in computer vision community. In [63], Jiang et al. proposed the use of visual attention to identify individuals with autism spectrum disorders (ASDs). In [51], Fan et al. integrated emotional factors into the traditional visual attention prediction and used emotional factors to correct the prediction results of visual attention. In [64], Xu et al. proposed a saliency prediction method for individual differences, which can generate different saliency maps for different individuals, and can be applied in personalized recommendation, personalized advertising and other fields.

2.2 Transformer Development

Transformer was proposed by Google for the machine translation task in 2017. It uses the multi-head self-attention mechanism to effectively describe the long-distance dependencies between words in the sequence. This model has brought a profound shock to the field of natural language processing. It is a landmark model [19]. With the deepening of the research, it has been paid more and more attention in the computer vision task.

Image transformer [31] was the first to migrate the transformer architecture to the field of computer vision. Since 2019, the visual model based on the transformer architecture has developed rapidly, and a large number of noteworthy results have emerged. Dosovitskiy et al. proposed the ViT (vision transformer) model [32], an image classification scheme based entirely on the self-attention mechanism. This is also the first work of the transformer to replace standard convolution, which extracts images as non-overlapping images. Blocks are used as the input of the encoder to achieve a word-like sequence, and then the local and global information between image blocks in the sequence is simultaneously extracted through the self-attention mechanism and positional encoding, respectively. Carion et al. built a new object detection framework DETR (detection transformer) [33] and applied transformer to the field of target detection for the first time. Liu et al. [34] applied transformer to the salient target detection task. The mode uses transformer's self-attention mechanism to describe the global dependence on salient targets. At the same time, T2T and reverse T2T methods are used to enhance the multi-scale characteristics of the representation. The boundary information is further introduced to refine the boundary of salient target prediction. After that, the transformer-based basic network Swin Transformer was applied to various vision tasks. The SwinNet model [35] applies Swin Transformer to salient target detection and proposes a salient target detection model for RGB-D and RGB-T, which achieves better performance. Inspired by this, this paper applies Swin Transformer to visual human eye focus detection as a backbone network to extract multi-scale features.

3. Model Architecture

The overall structure of the model is shown in Fig. 2. First of all, the image is combined with the center prior, contrast prior, and frequency prior that simulate the human center bias, aiming to further suppress invalid information, strengthen the expression of details and highlight local information. Then images and prior features are input into the Siamese Swin Transformer [20] network to learn remote context information and local information, and extract four sets of salient feature maps of different scales from the backbone network. Then the attention calibration module (ACM) is used for detail redirection to guide cross-modal collaborative learning of saliency features. Finally, the CNN decoder fuses feature maps for saliency prediction.

3.1 Multi-Prior Enhancement

3.1.1 Contrast Prior

To reduce the complexity of the contrast prior, the SLIC (simple linear iterative clustering) algorithm [36] is used to form adjacent pixels with similar characteristics in the image into irregular pixel blocks with certain visual significance. The position P_i and color C_i of the pixel p_i are obtained from the mean value of the spatial positions of all pixels in the pixel block and the mean value of the LAB color. The color distance and space distance between pixel blocks p_i and p_j are defined as:

$$d_c\left(p_i, p_j\right) = \left\|C_i - C_j\right\| \tag{1}$$

$$d_p\left(p_i, p_j\right) = \left\|P_i - P_j\right\| \tag{2}$$

where $\|\cdot\|$ is the L_2 norm.

Contrast is the degree of difference between pixel blocks relative to the overall image. Contrast highlights salient regions more, and high contrast in adjacent regions draws more attention than high contrast in distant regions. For the pixel block p_j, if the spatial distance between the pixel block p_j is smaller, the influence of p_i on the calculation of p_j contrast is greater, and its saliency value is calculated by calculating its color contrast with other pixel blocks, so the global contrast of a pixel block p_i is defined

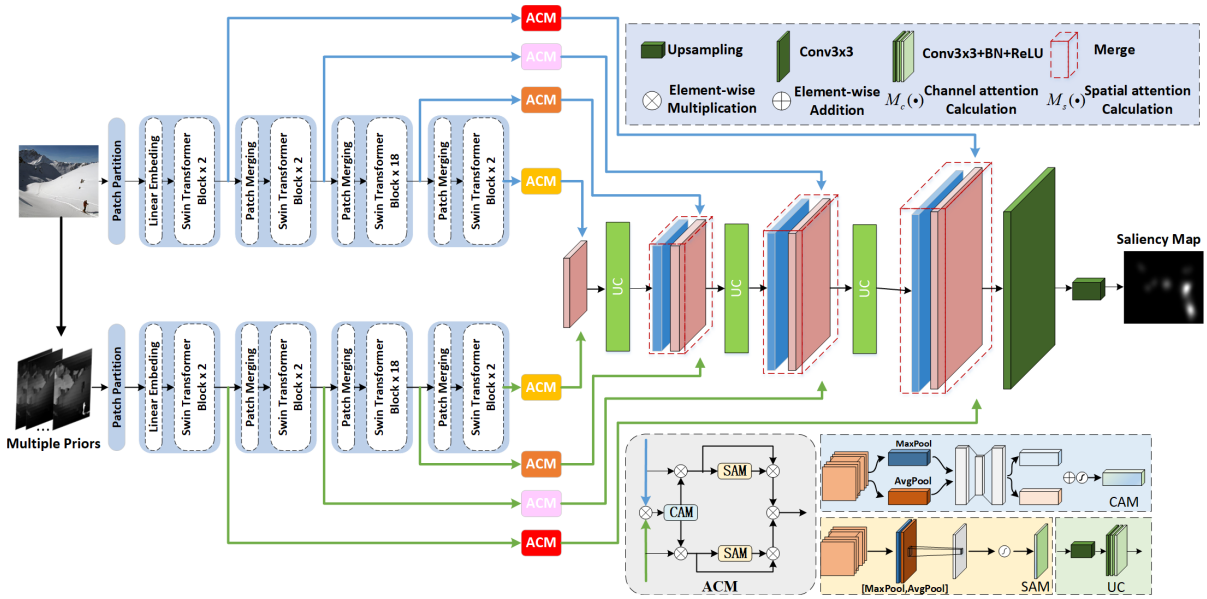

Fig. 2 Architecture of our ME-CAS network.

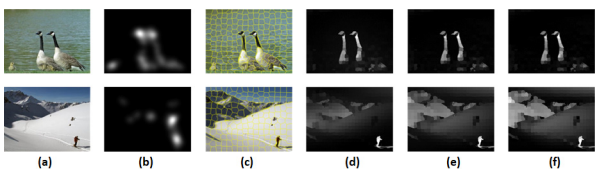

Fig. 3 Contrast priors. (a) Image, (b) ground truth, (c) pixel block division, (d), (e) and (f) contrast prior, σ^2 points are 0.3, 0.5, 0.7 respectively.

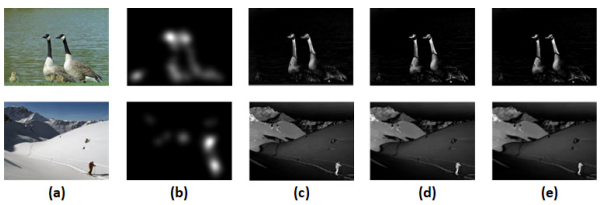

Fig. 4 Frequency priors. (a) Image, (b) ground truth, (c), (d) and (e) are frequency prior, σ_F are 6.0, 6.2 and 6.4 respectively.

as:

$$C_i^{glo} = \sum_{j=1}^{n} d_c^2\left(p_i, p_j\right) W_p\left(p_i, p_j\right) \tag{3}$$

where $w_p(i,j) = e^{-d_p^2(p_i,p_j)/\sigma^2}$, and $w_p(i,j)$ represents the spatial difference between the pixel block p_i and the pixel block p_i.

3.1.2 Frequency Prior

The bandpass filter can well reflect the salient area of the image for color filtering. In addition, according to the measurement of the biological visual system, our cellular response is similar to the Log-Gabor function, which is symmetrical in the logarithmic frequency, and the Log-Gabor filter can be constructed with an arbitrary bandwidth to reduce low-frequency over-representation [37]. We implement a saliency frequency prior to using a 2DLog-Gabor bandpass filter. Through the transfer function of the 2DLog-Gabor filter, the three channels of the image are band-pass filtered in the CIELAB color space, and then the saliency map based on the frequency prior is obtained according to the filtering results of the three channels. The frequency prior saliency S_f is defined as:

$$S_f = \left(L^2 + A^2 + B^2\right)^{1/2} \tag{4}$$

$$L = F^{-1}\left(F\left(I_L\right)^* LG\right) \tag{5}$$

$$A = F^{-1}\left(F\left(I_A\right)^* LG\right) \tag{6}$$

$$B = F^{-1}\left(F\left(I_B\right)^* LG\right) \tag{7}$$

where $F(\bullet)$ and $F^{-1}(\bullet)$ represent Fourier transform and inverse Fourier transform, * represents convolution operation, I_L, I_A, I_B represent the three channels of the image in CIELAB color space, respectively. LG is the transfer function of the 2D Log-Gabor filter, expressed in the frequency domain as:

$$LG(u) = \exp\left(-\left(\log\frac{\|u\|_2}{\omega_0}\right)^2 / 2\sigma_F^2\right) \tag{8}$$

where $u = (u, v) \in R^2$ is the coordinates of the Log-Gabor filter in the frequency domain, ω_0 and σ_F are the center frequency band and bandwidth of the filter, $\omega_0 = 0.002$.

3.1.3 Central Prior

Most saliency detection algorithms use center priors as a complement to contrast or frequency priors to strengthen the impact of saliency spatial locations on detection results. Combining previous research, we let the network learn the prior knowledge of the center, and we constrain each before

a two-dimensional Gaussian function whose mean and covariance matrix are freely learnable. This allows the network to learn its prior knowledge entirely from the data, without relying on assumptions from biological studies. The central bias is modeled using a set of Gaussian functions with a diagonal covariance matrix. The mean and variance of each prior graph are calculated according to the following equations:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\left(\frac{(x-\mu_x)^2}{2\sigma_x^2} + \frac{\left(y-\mu_y\right)^2}{2\sigma_y^2}\right)\right) \tag{9}$$

3.2 Siamese Transformer

Swin Transformer proposes a method that includes sliding window operations and hierarchically constructs transformers. The model introduces the hierarchical construction method commonly used in CNN. With the deepening of the network, downsampling can generate multi-scale features. In the self-attention calculation process, the calculation is constrained to be carried out within the divided local non-overlapping window, so that the algorithm complexity changes from the previous square relationship with the image size to a linear relationship, and the calculation amount is greatly reduced. At the same time, the sliding window is used to make the information of multiple non-overlapping windows interact effectively, that is, translation invariance is maintained without reducing the accuracy rate.

We adopt twin Swin Transformers to extract multi-level feature information of images and multi-prior knowledge respectively. The model first uses Patch Partition to divide the input image into non-overlapping patch sets according to 4×4 adjacent pixels as a patch, and then linearly transforms each pixel channel data through the Linear Embedding module, and the feature dimension is converted to C. Then the data enters multiple Swin Transformer modules and Patch Merging modules to obtain feature maps of different scales. The size of the feature maps is 1/4, 1/8, 1/16, and 1/32 of the input, respectively denoted as $\left\{S_i^c\right\}_{i=1}^4$ and $\left\{S_i^p\right\}_{i=1}^4$. Among them, the Swin Transformer based on the sliding window can be expressed as:

$$\hat{Z}^l = W - MSA\left(LN\left(Z^{l-1}\right)\right) + Z^{l-1} \tag{10}$$

$$Z^l = MLP\left(LN\left(\hat{Z}^l\right)\right) + \hat{Z}^l \tag{11}$$

$$\hat{Z}^{l+1} = SW - MSA\left(LN\left(Z^l\right)\right) + Z^l \tag{12}$$

$$Z^{l+1} = \mathrm{MLP}\left(LN\left(\hat{Z}^{l+1}\right)\right) + \hat{Z}^{l+1} \tag{13}$$

where LN represents the normalized network layer, MLP represents the multi-layer perceptron including the activation function $GELU$, $\hat{Z}^l$ represents the output of the self-attention module, and Z^l represents the output of the MLP module of the l block.

3.3 Attention Calibration Module

Since the position of salient objects in multi-modality image pairs should be the same, the features from different modalities need to be aligned at first to show the common salient position. At the same time, image saliency prediction should not only pay attention to the global context information, but also the local spatial information of the image. Since RGB images show more appearance information, feature extraction contains a large amount of global context information, while prior knowledge features show more local spatial information. The channel attention mechanism aggregates useful feature information by compressing the two-dimensional space of the feature map. However, detailed information is lost through the compressed channel attention weight. The detailed information of the image must be added to the spatial attention mechanism. In this paper, the attention calibration module combines these two different feature information, while paying attention to global contextual information and local spatial information, and guides the cross-modal collaborative learning of global and local information to pay more attention to salient content in each modality.

First, the contextual feature S_i^c and the prior feature S_i^p are fused to perform channel attention calculations to achieve channel alignment of the two modalities:

$$CA_i = M_c\left(S_i^c \times S_i^p\right) \tag{14}$$

Where $\times$ represents element multiplication and $M_c(\bullet)$ represents channel attention calculation [68].

Then, the spatial attention calculation is carried out, and the salient content of each mode is spatially calibrated to achieve detail redirection:

$$F_{\mathrm{i}}^c = M_s\left(S_i^c \times CA_{\mathrm{i}}\right) \times \left(S_i^c \times CA_{\mathrm{i}}\right) \tag{15}$$

$$F_{\mathrm{i}}^p = M_s\left(S_i^p \times CA_{\mathrm{i}}\right) \times \left(S_i^p \times CA_{\mathrm{i}}\right) \tag{16}$$

Where $M_s(\bullet)$ represents channel attention calculation [68].

Finally, the spatial calibration features are fused:

$$F_{\mathrm{i}}^a = F_{\mathrm{i}}^c \times F_{\mathrm{i}}^p \tag{17}$$

After multi-modal information fusion, channel alignment and spatial calibration are realized, and strong characterization ability is obtained.

3.4 Saliency Feature Decoder

After passing the attention calibration module, feature information on multiple scales is obtained. According to the decoding idea of FPN, the high-level features are gradually fused with the shallow features. The multi-scale feature information is up-sampled to obtain a feature map of the size of its adjacent shallow features. To enhance the long-range

and multi-scale information of the feature map, the upsampled feature map is concatenated with the feature information of the corresponding jump connection. The specific operation is as follows:

$$F_i = \begin{cases} \text{Upsample}\left(F_i^a\right) & i = 1 \\ \text{Concat}\left(F_i^a, \hat{F}_{i-1}\right) & i = 2,3,4 \end{cases} \tag{18}$$

$$\hat{F}_i = \text{Re}\,LU\left(BN\left(f^{3\times3}\left(F_i\right)\right)\right) \quad i = 1,2,3,4 \tag{19}$$

Finally, the saliency map with the same input image size is obtained by upsampling, and the sigmoid function is used to activate the saliency map:

$$\hat{y} = \text{sigmoid}\left(U\ \text{psample}\ \left(\hat{F}_4\right)\right) \tag{20}$$

$\hat{y}$ is the significance prediction map.

3.5 Loss Function

Saliency predictions are usually evaluated by different metrics to determine their quality factors [27]. Recent saliency prediction studies [38]–[40] show that using saliency evaluation metrics to define loss functions can significantly improve the performance of saliency prediction models. In this paper, we refer to TranSalNet [30] and use the linear combination of three different indicators as the loss function, which is expressed as follows:

$$\begin{aligned} L\left(\hat{y}, y^{den}, y^{fix}\right) = \alpha L_{KL}\left(\hat{y}, y^{den}\right) + \\ \beta L_{CC}\left(\hat{y}, y^{den}\right) + \lambda L_{NSS}\left(\hat{y}, y^{fix}\right) \end{aligned} \tag{21}$$

where $\hat{y}$ is the saliency prediction map, y^{den} is the continuous saliency distribution map, and y^{fix} is the binary map of the position of human attention. L_{KL}, L_{CC} and L_{NSS} represent kullback-Leibler divergence, linear correlation coefficient, and normalized scanpath saliency, respectively, which are commonly used evaluation indicators in the evaluation of saliency detection models. According to experimental verification, we set α, β and λ to 10, -1 and -1.

4. Experimental Results

4.1 Datasets

The models proposed in this paper were trained and tested on four databases: MIT1003 [41], MIT300 [42], CAT2000 [43] and SALICON [44]. Each database is described in detail below:

MIT1003 [41]: This database contains 1003 images. This database is also the first large database used to measure performance in the field of human eye focus detection. The eye tracker is used to record the eye-focused area in a picture, and the obtained eye-focused area is Gaussian filtered to obtain the final saliency map of the human eye-focused point.

MIT300 [42]: This database contains 300 images, and the processing process is similar to MIT1003. The truth map of the focus of the human eye is not public, but provides an online submission method to allow researchers to submit their models to the prediction map of the human eye focus model of the 300 images in this database, to be able to compare the difference in the performance of the detection model of each human eye focus.

CAT2000 [43]: This database contains 4000 images, 2000 for training and 2000 for testing. It consists of 20 different categories like cartoon, art, satellite and outdoor, etc. Its test set is not public and needs to be submitted online for model evaluation.

SALICON [44]: This dataset contains 20,000 images selected from the Microsoft COCO dataset, which is by far the largest dataset in the field of image human eye focus prediction. It contains 10000 training sets, 5000 validation sets, and 5000 test sets. This dataset does not record eye movement data using an eye tracker, but eye movement data is recorded with a mouse. Its test set is not public, and the prediction results must be submitted to the SALICON challenge website [45] for evaluation.

4.2 Evaluation Metrics

There are many evaluation metrics in the field of saliency prediction research. Previous saliency evaluation studies have shown that using multiple metrics can improve the fairness of evaluation. According to the different assumptions made by evaluation metrics on visual saliency, they can be divided into location-based evaluation metrics and probability distribution-based evaluation metrics: location-based evaluation metrics treat saliency as a random variable, and probability distribution-based evaluation metrics treat saliency as a probability distribution [46]. There are nine commonly used evaluation metrics, but according to the MIT and SALICON benchmarks, the following seven metrics are usually employed now, among which the location-specific evaluation metrics include normalized scanpath saliency (NSS), Area under ROC Curve (ROC curve), shuffled AUC (sAUC) and information gain (IG); location-based evaluation metrics include kullback-leibler divergence (KL), linear correlation coefficient (CC) and similarity metric (SIM) [47].

In addition, according to the measurement methods of different evaluation metrics, they can be divided into similarity measurement metrics and dissimilarity measurement metrics: the larger the similarity metrics, the better the model performance; the smaller the dissimilarity metrics, the better the model performance. Among the metrics used above, the KL is a similarity metrics, and the rest are non-similarity metrics.

4.3 Implementation Details

Currently, most models adopt transfer learning, following a state-of-the-art similar training process, where the parameters of the feature extraction network are initialized with weights trained on ImageNet [48], and other layers are ini-

Table 1 The prior knowledge ablation study of our model on the MIT1003 dataset

Contrast	Frequency	Central	sAUC↑	NSS↑	CC↑	AUC↑	SIM↑
			0.553	2.196	0.892	0.883	0.713
✓			0.569	2.231	0.898	0.887	0.725
	✓		0.559	2.242	0.898	0.885	0.719
		✓	0.557	2.201	0.896	0.887	0.726
✓	✓		0.567	2.251	0.899	0.890	0.736
✓		✓	0.578	2.261	0.897	0.892	0.750
	✓	✓	0.574	2.264	0.900	0.891	0.748
✓	✓	✓	**0.584**	**2.267**	**0.901**	**0.894**	**0.755**

Table 2 Ablation study of our model on MIT1003, CAT2000, and SALICON validation sets.

Dataset	Model	sAUC↑	NSS ↑	CC↑	AUC↑	SIM↑
MIT1003	Base	0.703	**2.930**	0.690	0.897	0.608
	CAS	0.731	2.914	0.783	0.900	0.611
	ME_CAS	**0.763**	**2.930**	**0.787**	**0.913**	**0.629**
CAT2000	Base	0.548	2.115	0.867	0.836	0.743
	CAS	0.553	2.196	0.892	0.883	0.713
	ME_CAS	**0.584**	**2.267**	**0.901**	**0.894**	**0.755**
SALICON	Base	0.698	1.830	0.879	0.847	0.736
	CAS	0.744	1.935	0.903	0.860	0.792
	ME_CAS	**0.746**	**1.985**	**0.916**	**0.869**	**0.804**

tialized randomly. To prevent the model from overfitting, we firstly trained the largest dataset SALICON, then freeze some parameters and fine-tune on MIT1003 and CAT2000. For MIT1003, we randomly divided it into 900 training sets and 103 validation sets; for CAT2000, we randomly sampled 10 images from each category as the validation set, and the rest were training sets.

The size of the input picture is adjusted to 384 × 384, the batch-size of the training network is set to 2, the Adam [49] optimizer is used to train the network, and the initial learning rate is set to 1×10^{-5}. During the training process, the verification frequency is limited to 1, and the verification is performed every time the training is performed. Finally, save the parameters and use the test set to test the final detection ability of the model.

4.4 Ablation Study

4.4.1 Analysis about Prior Knowledge

In order to verify the effectiveness of each prior knowledge, a series of ablation experiments were conducted to compare the performance of the models under different combinations of prior knowledge. We evaluated the contribution of each prior knowledge using the MIT1003 dataset.

Table 1 shows the quantitative results of the multi-prior knowledge ablation experiment. The results show that multiple tests can predict better significance graph. All evaluation metrics are constantly improving. For example, the sAUC aspect achieves a result of 0.553 without a priori. Contrast priors, frequency priors and center priors were added to achieve a relative improvement of 2.8%, 1.0% and 0.7%, respectively. When two priors are used, the results are further improved. For example, when contrast priors and center priors are added, the results are improved by 1.5% compared with only contrast priors. Finally, the multi-priori knowledge is further improved by 1.0% compared with the two priori, and by 5.6% compared with no priori.

4.4.2 Analysis about Module

In order to further analyze the actual gain of the multimodal network and prior knowledge, the ablation experiment of the multimodal network and prior knowledge was conducted in this paper following the same experimental setup. Base represents single-mode network, and the trunk network uses single-branch Swin Transformer to extract features. Meanwhile, the channel space attention module is replaced by cbam module. CAS represents multi-modal network, and the multi-modal input is RGB image, but there is no multi-prior knowledge. ME-CAS represents our cross-modal collaborative network. The input of one branch is RGB image, and the input of the other branch is multi-priori knowledge.

Table 2 shows the results of our model's ablation experiments on three data sets. Through analysis, we can see that our multi-priori knowledge and cross-modal collaboration are very important parts and have indispensable contributions to the improvement of performance. Specifically, on the SALICON dataset, continuous improvement was observed for all indicators. For example, base is 0.879 in terms of CC, but after cross-modal collaboration and multi-priors cross-modal collaboration, the results are 0.903 and 0.916 respectively, an increase of 7.6% and 9.2%. A similar phenomenon can be seen for other measures across all three data sets. The qualitative results are shown in Fig. 5.

4.5 Performance Comparison

To further verify the effectiveness of the saliency prediction algorithm proposed in this paper, a comparative analysis was conducted with existing algorithms on SALICON [44], MIT300 [42], and CAT2000 [43] databases. The methods compared are all part of an excellent paper on human eye attention zone prediction published in top conferences or journals in the field of computer vision in recent years.

For the SALICON database, we submit the prediction results to the official SALICON website, and the results are evaluated on the SALICON challenge website [45]. The competition uses a unified evaluation process, resulting in more fair results. Table 3 shows the results of various evaluation metrics on the SALICON dataset. It can be seen that our model has achieved high performance, and our model has achieved multiple first places. Although some metrics did not make the top three, their performance scores are still impressive.

For the MIT300 database, we use the MIT1003 database to fine-tune the model, generate saliency maps, and submit them to the MIT benchmark for testing. From benchmarks we know that benchmarks evaluate models on different criteria, i.e. a model must be explicitly declared as probabilistic or non-probabilistic and thus can be evaluated

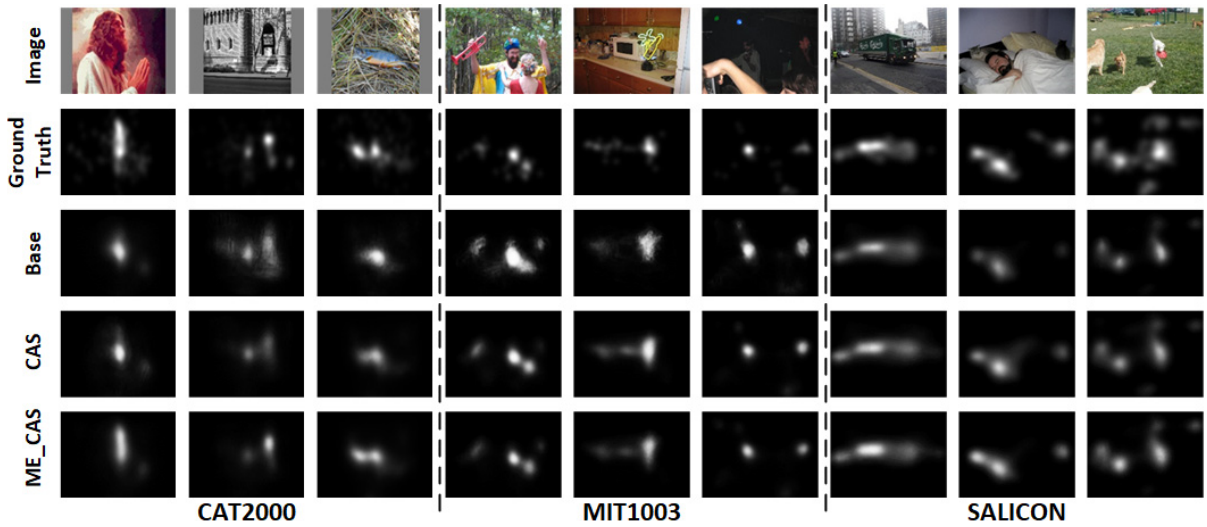

Fig. 5 Comparison of significance prediction performance of three model variants in ablation studies.

Table 3 Performance on test set of LSUN'17 Competition (SALICON-2017-version).

Model	sAUC↑	NSS↑	CC↑	AUC↑	SIM↑	KL↓
SAM-Res [27]	0.741	1.990	0.899	0.865	0.793	0.610
DeepGaze IIE [52]	0.767	1.996	0.872	0.869	0.733	0.285
MD-SEM [53]	0.746	2.058	0.868	0.864	0.774	0.568
GazeGAN [39]	0.736	1.899	0.879	0.864	0.773	0.376
DINet [54]	0.739	1.959	0.902	0.862	0.795	0.864
UNISAL [55]	0.739	1.952	0.879	0.864	0.775	-
MSI-Net [56]	0.736	1.931	0.889	0.865	0.784	0.307
FBNet [57]	0.706	1.687	0.785	0.843	0.694	0.708
TranSalNet [30]	0.747	2.014	0.907	0.868	0.803	0.373
ACNet [50]	0.739	1.948	0.896	0.866	0.786	0.228
SalFBNet [40]	0.740	1.952	0.892	0.868	0.772	0.236
Ours	0.746	1.970	0.909	0.869	0.805	0.378

Table 4 Performance on test set of MIT300.

Model	AUC↑	sAUC↑	CC↑	SIM↑	NSS↑	KL↓
GBVS [11]	0.806	0.629	0.479	0.887	1.245	0.887
CAS [10]	0.758	0.640	0.384	0.431	1.018	1.072
LDS [58]	0.810	0.602	0.517	0.522	1.364	1.063
BMS [59]	0.771	0.691	0.413	0.445	1.151	1.023
ConvSal [60]	0.811	0.589	0.500	0.505	1.336	1.722
DVA [61]	0.843	0.725	0.663	0.584	1.930	0.629
SalGAN [26]	0.849	0.735	0.674	0.593	1.862	0.757
eDN [22]	0.817	0.618	0.451	0.411	1.139	1.136
EML-NET [28]	0.876	0.746	0.789	0.675	2.487	0.843
CASNet II [51]	0.855	0.739	0.705	0.580	1.985	0.585
SAM-Vgg [27]	0.847	0.730	0.663	0.598	1.955	1.274
SAM-Res [27]	0.852	0.739	0.689	0.611	2.062	1.171
ML-Net [62]	0.838	0.739	0.663	0.581	1.974	0.800
TranSalNet [30]	0.873	0.746	0.807	0.689	green2.413	014
GazeGAN [39]	0.860	0.731	0.757	0.649	2.211	1.339
Ours	0.874	0.757	0.819	0.697	2.436	1.281

Table 5 Performance on test set of CAT2000.

Model	AUCJ↑	NSS↑	sAUC↑	CC↑	SIM↑	EMD↓
Itti [12]	0.56	0.25	0.52	0.09	0.34	4.46
GBVS [11]	0.80	1.23	0.58	0.50	0.51	2.99
SUN [65]	0.70	0.77	0.57	0.30	0.43	3.42
LDS [58]	0.83	1.54	0.56	0.62	0.58	2.09
eDN [22]	0.85	1.30	0.55	0.54	0.52	2.64
EYMOL [66]	0.83	1.78	0.51	0.72	0.61	1.91
SDDPM [67]	0.81	1.22	0.54	0.51	0.52	2.31
DeepFix [25]	0.87	2.28	0.58	0.87	0.74	1.15
SAM-Vgg [27]	0.88	2.38	0.58	0.89	0.76	1.07
MSI-Net [56]	0.88	2.30	0.59	0.87	0.75	1.07
Ours	0.88	2.38	0.59	0.90	0.77	0.98

fairly within the category it belongs to. As with MIT significance benchmarks, we do not assume that our models are probabilistic. As suggested by TranSalNet, for the fairness of the results, we only compare the non-probability model with the traditional model. Table 4 shows the results of each evaluation metrics. According to the findings of Bylinskii et al. [46], under the assumption of non-probabilistic modeling, NSS and CC provide the fairest comparison, and if evaluating probabilistic models, KL is recommended; our model ranks first on the metrics CC, while NSS ranks second. At the same time, our metrics also rank first in SIM and sAUC, which shows that our model may be the best probability model.

For the CAT2000 dataset, CAT2000 is used for model fine-tuning, and the saliency map predicted by the model is submitted to the MIT saliency benchmark. The compared model results are from the MIT saliency benchmark website, and the comparison results are more fair. Table 5 shows the results of each evaluation index. It can be seen that the model in this paper has achieved good significance performance, and several evaluation indexes rank first.

To further illustrate the advantages of this chapter's approach, Fig. 6 shows the results of qualitative comparison with other advanced models. The images are from SALICON and MIT1003 data sets. It can be seen that the method in this chapter can more accurately predict significant areas, including indoor, outdoor and track and field scenes. For example, in lines 1-3 of the SALICON dataset and lines 1 and 2 of the MIT1003 dataset, the positions predicted by most methods are too "concentrated" and the prediction ability of details is weak. In lines 4 and 5 of the SALICON dataset,

Fig. 6 Comparison of qualitative results generated by our saliency model with state-of-the-art methods. Images are from SALICON and MIT1003 datasets.

most prediction methods fail to make significant predictions in non-significant positions. In line 3-5 of MIT1003 dataset, most other methods cannot fully predict significant objects, while the model in this chapter successfully fully predicts them. In general, other methods can accurately predict high-level semantic scenes, including faces, objects, and other categories, but cannot predict truly significant areas for complex scenes. However, the method proposed in this paper can effectively cope with challenging scenarios.

5. Conclusion

In this paper, we propose a multi-prior knowledge-based cross-modal saliency prediction model called ME-CAS. In our model, the main novelty is the combination of prior knowledge and neural networks, which enhances the expression of details through multiple prior knowledge. At the same time, the Siamese transformer is applied to saliency prediction. The two branches extract global saliency information and local detail information respectively and learn saliency features across modes. Ablation experiments demonstrate the contribution of multiple prior knowledge to the model while proving the necessity of cross-modal synergy. Extensive experiments demonstrate that our proposed method achieves better predictive performance on public saliency benchmarks compared to other existing models. Our model has achieved good performance on the public natural datasets, but there is still a lack of further exploration in the direction of significance personalization. In the traditional cognitive field, there are many classical theories and models about visual attention, which are more consistent with biological principles. Therefore, it is necessary to combine deep learning with classical cognitive theory to further explore new theories and models.

Acknowledgements

This work was supported by National key R&D projects ”Coal mine disaster fusion monitoring and decision-making digital key technology equipment and demonstration application” subject: Coal mine safety hid danger video image intelligent recognition technology and equipment (2022YFC3004703).

References

[1] R.A. Rensink, “The dynamic representation of scenes,” Visual Cognit., vol.7, no.1-3, pp.17–42, 2000.
[2] H. Huang, M. Cai, L. Lin, L. Lin, J. Zheng, X. Mao, X. Qian, Z. Peng, J. Zhou, Y. Iwamoto, X.-H. Han, Y.-W. Chen, R. Tong, “Graph-based pyramid global context reasoning with a saliency-aware projection for COVID-19 lung infections segmentation,” ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.
[3] S.K. Yarlagadda, D.M. Montserrat, D. Güera, C.J. Boushey, D.A. Kerr, and F. Zhu, “Saliency-aware class-agnostic food image segmentation,” ACM Trans. Comput., vol.2, no.3, pp.1–17, July 2021.
[4] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, “Salient object segmentation via effective integration of saliency and objectness,” IEEE Trans. Multimedia, vol.19, no.8, pp.1742–1756, Aug. 2017.
[5] W. Wang, J. Shen, and F. Porikli, “Saliency-aware geodesic video object segmentation,” IEEE CVPR 2015, 2015.
[6] H. Hadizadeh and I.V. Bajić, “Saliency-aware video compression,” IEEE Trans. Image Process. : a Publication of the IEEE Signal Processing Society, vol.23, no.1, pp.19–33, Jan. 2014.
[7] Q.-G. Ji, Z.-D. Fang, Z.-H. Xie, and Z.-M. Lu, “Video abstraction based on the visual attention model and online clustering,” Signal Process. Image Commun., vol.28, no.3, pp.241–253, March2013.
[8] Q. Jiang, S. Feng, W. Lin, K. Gu, G. Jiang, and H. Sun, “Optimizing multistage discriminative dictionaries for blind image quality assessment,” IEEE Trans. Multimedia, vol.20, no.8, pp.2035–2048, Aug. 2018.
[9] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, “Saliency-guided quality assessment of screen content images,” IEEE Trans. Multimedia, vol.18, no.6, pp.1098–1110, June 2016.
[10] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.10, pp.1915–1926, Oct. 2012.
[11] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” Conference on Advances in Neural Information Processing Systems, 2006.
[12] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol.20, no.11, pp.1254–1259, Nov. 1998.
[13] N. Bruce and J.K. Tsotsos, “Saliency based on information maximization,” International Conference on Neural Information Processing Systems, pp.155–162, Dec. 2005.
[14] D. Gao and N. Vasconcelos, “Discriminant saliency for visual recognition from cluttered scenes,” Advances in Neural Information Processing Systems 17, 2004.
[15] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[17] G. Huang, Z. Liu, L. van der Maaten, and K.Q. Weinberger, “Densely connected convolutional networks,” IEEE Computer Society, 2016.
[18] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, 2012.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Computation and Language, arXiv, 2017.
[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
[21] M.M. Cheng, G.X. Zhang, N.J. Mitra, X. Huang, and S.-M. Hu, “Global contrast based salient region detection,” Computer Vision and Pattern Recognition, 2011.
[22] E. Vig, M. Dorr, and D. Cox, “Large-scale optimization of hierarchical features for saliency prediction in natural images,” 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[23] M. Kümmerer, L. Theis, and M. Bethge, “Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet,” Comput. Sci., 2014.
[24] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” Comput. Sci., 2014.
[25] S. Kruthiventi, K. Ayush, and R.V. Babu “DeepFix: A fully convolutional neural network for predicting human eye fixations,” IEEE Trans. Image Process., vol.26, no.9, pp.4446–4456, Sept˙ 2017.
[26] J. Pan, C. Canton, K. McGuinness, N.E. O’Connor, J. Torres, E. Sayrol, and X. Giro-i-Nieto, “SalGAN: Visual saliency prediction with generative adversarial networks,” Computer Vision and Pattern Recognition, 2017.
[27] C. Marcella, B. Lorenzo, S. Giuseppe, and R. Cucchiara, “Predicting human eye fixations via an LSTM-based saliency attentive model,” IEEE Trans. Image Process., vol.27, no.10, pp.5142–5154, Oct. 2016.
[28] S. Jia and N.D.B. Bruce, “EML-NET:An expandable Multi-Layer NETwork for saliency prediction,” Computer Vision and Pattern Recognition, arXiv, 2018.

[29] A. Linardos, M. Kümmerer, O. Press, and M. Bethge, "DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
[30] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu, "TranSalNet: Towards perceptually relevant visual saliency prediction," Multimedia, arXiv e-prints, 2022.
[31] N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," Computer Vision and Pattern Recognition, 2018.
[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," International Conference on Learning Representations, 2021.
[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-end object detection with transformers," ECCV, 2020.
[34] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," Computer Vision and Pattern Recognition, arXiv, 2021.
[35] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "SwinNet: Swin Transformer drives edge-aware RGB-D and RGB-T salient object detection," IEEE Trans. Circuits Syst. Video Technol., vol.32, no.7, pp.4486–4497, July 2022.
[36] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.11, pp.2274–2282, Nov. 2012.
[37] X. Zhang, Y. Wang, Z. Chen, J. Yan and D. Wang, "Saliency detection via image sparse representation and color features combination," Multimed. Tools Appl., June 2020.
[38] D. Cheng, R. Liu, J. Li, S. Liang, Q. Kou, and K. Zhao, "Activity guided multi-scales collaboration based on scaled-CNN for saliency prediction," Image Vision Comput., vol.114, Oct. 2021.
[39] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. Le Callet, "How is Gaze influenced by image transformations? Dataset and model," IEEE Trans. Image Process. vol.29, pp.29–2287, 2020.
[40] G. Ding, N. Mamolu, A. Caglayan, M. Murakawa, and R. Nakamura, "SalFBNet: Learning pseudo-saliency distribution via feedback convolutional networks," Image Vision Comput., vol.120, April 2022.
[41] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," IEEE International Conference on Computer Vision, 2010.
[42] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," MIT Computer Science and Artificial Intelligence Laboratory Technical Report, 2012.
[43] A. Borji and L. Itti, "CAT2000: A large scale fixation dataset for boosting saliency research," Computer Vision and Pattern Recognition, arXiv, 2015.
[44] H. Xun, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," 2015 IEEE International Conference on Computer Vision (ICCV), 2015.
[45] L.Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang and H.-Q. Zhou, "A visual attention model for adapting images on small displays," Multimedia Systems, Oct. 2003.
[46] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," IEEE Trans. Pattern Anal. Mach. Intell., vol.41, no.3, pp.740–757, March 2017.
[47] A. Borji, H.R. Tavakoli, D.N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," IEEE International Conference on Computer Vision, 2014.
[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg and L. Fei-Fei, "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis., April 2015.
[49] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," Comput. Sci., 2014.
[50] P. Li, X. Xing, X. Xu, B. Cai, and J. Cheng, "Attention-aware concentrated network for saliency prediction," Neurocomputing, vol.429, pp.199–214, March 2021.
[51] S. Fan, Z. Shen, J. Ming, B.L. Koenig, J. Xu, M.S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[52] A. Linardos, M. Kümmerer, and O. Press, M. Bethge, "DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
[53] C. Fosco, A. Newman, P. Sukhum, Y. Bin Zhang, N. Zhao, A. Oliva, and Z. Bylinskii, "How much time do you have? modeling multi-duration saliency," Computer Vision and Pattern Recognition, 2020.
[54] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A dilated inception network for visual saliency prediction," IEEE Trans. Multimedia, vol.22, no.8, pp.2163–2176, Aug. 2019.
[55] R. Droste, J. Jiao, and J.A. Noble, "Unified image and video saliency modeling," ECCV 2020, pp.419–435, Oct. 2020.
[56] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder-decoder network for visual saliency prediction," Neural Netw.vol.129, pp.261–270, Sept. 2019.
[57] G. Ding, N. Imamoglu, A. Caglayan, M. Murakawa, and R. Nakamura, "FBNet: FeedBack-recursive CNN for saliency detection," 17th International Conference on Machine Vision Applications (MVA), 2021.
[58] S. Fang, J. Li, Y. Tian, T. Huang, and X. Chen, "Learning discriminative subspaces on random contrasts for image saliency analysis," IEEE Trans. Neural Netw. Learn. Syst., vol.28, no.5, pp.1095–1108, May 2017.
[59] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," 2013 IEEE International Conference on Computer Vision, 2013.
[60] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," J. Vis., vol.13, no.11, March 2013.
[61] W. Wang and J. Shen, "Deep Visual attention prediction," IEEE Trans. Image Process., vol.27, no.5, pp.2368–2378, May 2018.
[62] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A Deep multi-level network for saliency prediction," International Conference on Pattern Recognition, 2016.
[63] M. Jiang and Q. Zhao, "Learning visual attention to identify people with autism spectrum disorder," 16th IEEE International Conference on Computer Vision (ICCV), 2017.
[64] Y. Xu, S. Gao, J. Wu, N. Li, and J. Yu, "Personalized saliency and its prediction," IEEE Trans. Pattern Anal. Mach. Intell., vol.41, no.12, pp.2975–2989, Dec. 2019.
[65] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," J. Vis., vol.8, no.7, Dec. 2008.
[66] D. Zanca and M. Gori, "Variational laws of visual attention for dynamic scenes," Proc. NIPS, pp.3826–3835, Dec. 2017.
[67] N. Rabbani, B. Nazari, S. Sadri, and R. Rikhtehgaran, "Efficient Bayesian approach to saliency detection based on Dirichlet process mixture," IET Image Proc., vol.11, no.11, pp.1103–1113, Nov. 2017.
[68] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon, "CBAM: Convolutional block attention module," the 15th European Conference on Computer Vision (ECCV), pp.3–19, Oct. 2018.

Fazhan Yang received the B.S. degree from Xu Hai College of China University of Mining and Technology, Xuzhou, China, in 2020. He has been pursuing the M.S degree in the School of Information and Control Engineering, China University of Mining and Technology, from 2020 year to now. His research interests in Internet of Things in Mines and perceptual image processing.

Xingge Guo received the master's and Ph.D. degrees from the University of Mining and Technology, China, in 2006 and 2013, respectively. 2018–2020 Visiting Scholar at Columbia University and University of South Carolina. He is currently a associate professor of China University of Mining and Technology, China. His research interest covers mobile communication, information processing and industrial intelligence.

Song Liang received the B.S. degree from China University of Mining and Technology, Xuzhou, China, in 2013. He has been pursuing the Ph.D. degree in the School of Information and Control Engineering, China University of Mining and Technology, from 2018 year to now. His research interests include visual affective computing and perceptual image processing.

Peipei Zhao is currently a associate professor of China University of Mining and Technology, China. Her research interests include information processing and coal mine communication.

Shanhua Li received the B.S. degree from Linyi University, Linyi, China, in 2020. He has been pursuing the M.S degree in the School of Information and Control Engineering, China University of Mining and Technology, from 2020 year to now. His research interests include action recognition and image processing.