Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19404-4

Transformer-based multi-level attention integration network for video saliency prediction

Rui Tan[1,3] · **Minghui Sun**[2,3] · **Yanhua Liang**[2,3]

Received: 1 November 2023 / Revised: 8 April 2024 / Accepted: 14 May 2024

Abstract
Most existing models for video saliency prediction heavily rely on 3D convolutional operations to extract spatio-temporal features. However, it is worth noting that 3D convolution produces a local receptive field, which may struggle to capture long-range spatio-temporal dependencies effectively. To compensate for such shortage, this paper introduces a novel approach called the Transformer-based Multi-level Attention Integration Network (TMAI-Net) for video saliency prediction. TMAI-Net is designed as a two-stream encoder-decoder model, carefully integrating multi-level features of semantic information. Our model incorporates a Multi-level Interactive Attention(MLIA) module and a Transformer, both implemented based on self-attention mechanism, which are placed at different levels of the model to capture long-range spatio-temporal feature dependencies. Additionally, our model operates on input video frames and attentional patches, allowing the Transformer module to capture structural similarities between related objects in global features and attention features. This, in turn, enables the model to allocate increased attention to salient areas. The efficacy of our proposed approach is validated through extensive experiments conducted on three widely recognized benchmark datasets.

Keywords Video saliency prediction · Transformer · Spatio-temporal feature · Self-attention

1 Introduction

One of the most enduring research problems in computer vision is the video saliency prediction. It aims to find the most noticeable regions of a dynamic scene. Video saliency prediction

✉ Minghui Sun
smh@jlu.edu.cn

Rui Tan
tanrui22@mails.jlu.edu.cn

Yanhua Liang
yhliang@jlu.edu.cn

1 Software College, Jilin University, Qianjin Street, Changchun 130012, Jilin, China

2 College of Computer Science and Technology, Jilin University, Qianjin Street, Changchun 130012, Jilin, China

3 Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Qianjin Street, Changchun 130012, Jilin, China

Published online: 25 May 2024

models have been extensively adopted in a variety of video processing applications, including video compression [1, 2], video captioning [3], and video surveillance [4, 5].

Historically, video saliency prediction methodologies have traditionally relied on the amalgamation of visual cues, including intensity, color, motion, and spatial frequency. These cues are woven together to generate a saliency map, offering insights into regions of visual saliency. However, earlier methods exhibited limitations, manifesting in their inability to effectively incorporate temporal dependencies and the indiscriminate treatment of individual pixels, often leading to suboptimal results. In recent years, the advent of deep learning has been combined with the proliferation of high-quality video saliency prediction datasets, which has catalyzed significant advances in video saliency prediction and the emergence of several different model categories. These models include two-stream models [6–9], LSTM-based models [10–12], and 3D convolutional models [13–15]. The two-stream model acquires temporal and spatial information separately, and then fuses them into the spatio-temporal information to obtain the final saliency map. However, the two-stream model often extracts temporal information based on optical flow, so it only considers the temporal information between adjacent frames. The emergence of the LSTM-based model alleviates this limitation because LSTM can extend temporal perception. However, since the LSTM-based model processes spatial information and temporal information with convolutional networks and LSTMs respectively, the model cannot use spatial information and temporal information at the same time, which is important in the field of saliency prediction. To address this issue, Min et al. [13] proposed TASED-Net, a model founded on a 3D convolutional network, specifically devised for the joint processing of spatio-temporal information. While there has been substantial progress in 3D convolution-based models, they are still unable to overcome the inherent limitations of local receptive fields. This is where our research strives to reach a solution. Since the long and short term memory of the human visual system affects visual attention processes, we take inspiration from the self-attention mechanism [16]. We note that the dot product attention inside the self-attention mechanism can be used to establish long-range spatio-temporal interactions between features at different time steps. As a result, we added MLIA and Transformer, both implemented on a self-attention mechanism, to the model, placed at different levels to generate a global spatiotemporal context. The reason for choosing Transformer is that due to its inherent attention mechanism and Multilayer Perceptron (MLP) structure, Transformer can split input object features into patch tokens to mine the patch structure similarity of related objects. It makes the model pay more attention to the saliency region while establishing long-range dependencies.

We propose a novel approach, the Transformer-based Multi-level Attention Integration Network (TMAI-Net). This innovative model is designed to collectively address the previously mentioned limitations. The encoder for TMAI-Net is a 3D fully convolutional network from S3D [17] that was pre-trained on the Kinetics dataset [18]. 3D convolutional layers have the ability to encode hierarchical spatio-temporal information, which can encode not only low-level information such as colour contrasts, but also high-level semantic information such as persons. In our model, the 3DCNN encoder extracts four branches from different levels from shallow to deep, which produce different features from low to high levels. We place MLIA and Transformer at the shallowest and deepest levels of the four branches respectively to construct long-distance spatial-temporal interactions through the self-attention mechanism they both have. Besides, we use a two-stream input strategy in our model [19], where the model receives the stacked original video frames together with their appropriate attentional patches as input. We use Transformer to mine the structural similarity of related objects in

the extracted global features and attention features to make the model focus more on saliency regions. Overall, the following are our main contributions:

- We propose a Transformer-based Multi-level Attention Integration Network (TMAI-Net) for video saliency prediction, which introduces the self-attention mechanism to compensate for the limitations of existing 3D CNN-based models.
- We present a Multi-Level Interactive Attention (MLIA) module to capture long-range spatio-temporal relationships between time steps. The MLIA module utilize the self-attention mechanism at the pixel level to predict human visual attention.
- The MLIA module and Transformer module are placed in the shallowest and deepest levels of the four branches extracted from the model's encoder, respectively, and directly establish the global spatio-temporal context at most levels. Besides, the structural similarity among related objects in global and attention features is mined through the Transformer, which helps the model to focus more on saliency regions.

2 Related work

2.1 Recent video saliency prediction models

The traditional video saliency prediction model combines both static and motion information and uses hand-designed spatio-temporal features for saliency modeling [20, 21]. Given the swift development of deep learning methodologies and the accessibility of extensive dynamically annotated datasets, such as DHF1K [11], deep learning-based video saliency prediction models have prevailed, demonstrating notable superiority over conventional models. Bak et al. [6] proposed a two-stream convolutional neural network that takes video frames and corresponding optical flow maps as input, merging spatial and temporal streams to produce salieny maps. Li et al. [22] presented a precise end-to-end learning framework for video saliency prediction, which collects motion information through optical flow and enhances temporal coherence by employing LSTM networks to encode sequence features. Wang et al. [23] proposed ACLNet, which enhances the CNN-LSTM architecture with an attention mechanism, facilitating rapid end-to-end saliency learning and enabling temporal saliency representation across consecutive frames via LSTM. Liu et al. [24] designed a novel saliency detection algorithm that effectively transfers image reconstruction knowledge to the learning process of saliency detection. Liu et al. [25] designed a new scene-guided two-branch network for salient object detection that allows cross-task knowledge distillation from scene classification. Liu et al. [26] extended residual pose routing to saliency prediction, which improved the computational efficiency while reducing the parameters. In addition, some methods introduce the nature of the study of part-whole relationships into salient object detection and use multi-flow strategies and confidence scores to improve the model's ability to segment salient objects [27–29]. Moreover, 3D convolutional architectures have been investigated for video saliency prediction tasks. Min et al. [13] proposed TASED-Net, a 3D convolutional encoder-decoder model designed to concurrently manage temporal information while extracting spatial features. Furthermore, Xue et al. [19] proposed a novel 3D convolutional encoder-decoder network named ECANet, which proposes explicit cyclic attention for temporal modeling and pixel emphasis. ViNet is an innovative visual architecture that includes an auditory module to investigate the fusion of audiovisual cues in video saliency prediction tasks [30]. STA3D is a spatio-temporal attention 3D network that selectively propagates saliency temporal features and refines spatial features for video saliency prediction [31]. The

above video saliency prediction model is constructed based on 3DCNN. However, the 3D convolution operation only produces local receptive fields and ignores feature dependencies at long distances.

In our work, our model uses MLIA and Transformer to model long-range spatio-temporal relationships at different levels, directly constructing global context and enabling interactive attention across spaces and scales.

2.2 Attention mechanism

In a variety of computer vision [32, 33] and natural language processing tasks [34, 35], attention mechanisms have recently shown exceptional effectiveness. The input labels are initially transformed into queries, keys, and values at the embedding layer in the conventional attention mechanism. Then, dot product attention is used to compute the long-range relationships among the labels in the input sequence. In the domain of computer vision, Oh et al. [36] proposed a temporal memory network utilizing an attention mechanism for video image segmentation, effectively capturing long-range dependencies between current and past frames while maintaining memory updates for enhanced performance. Yuan et al. [16] proposed a novel feature pyramidal interactive attention network for egocentric gaze prediction, which leverages an attention mechanism to facilitate interactive attention across space and scale, effectively capturing long-term relationships among spatio-temporal features at various time steps. Wang et al. [37] proposed the STSANet to model long-range spatio-temporal relations separately for different levels of information extracted, and then fuse the spatio-temporal features of different levels to output saliency maps. Zhang et al. [38] designed an attention-guided mechanism that adaptively learns adjacent feature fusion weights to perform better learning of multiscale spatio-temporal features.

The above models use self-attention mechanism to model long-range spatio-temporal dependencies, which compensates for the limitations of 3DCNN in modeling local spatio-temporal features. Besides, given the different perceptual fields and the different richness of semantic information at each level, it is also necessary to consider the differences in the way of modeling the global dependencies at different levels. In our module, in order to better utilize multi-level features, maximize the accuracy of the model, and reduce the complexity of the model, we place MLIA and Transformer in the shallowest and deepest levels of the four branches extracted by the model encoder, respectively. We leverage the self-attention mechanism in MLIA and Transformer to model long-range spatio-temporal dependencies.

2.3 Visual transformer

The advent of the Transformer architecture has considerably accelerated development in the areas of computer vision and natural language processing (NLP). It has gained widespread adoption in tasks encompassing classification [39–42], segmentation [43], and detection [44–47], consistently delivering outstanding performance. Compared with the limited CNN modeling of local spatio-temporal feature, the Transformer has been used for the modeling of global spatio-temporal feature due to its inherent self-attention mechanism. Liu et al. [48] proposed a Video Swin Transformer based on spatio-temporal local sensing bias, which is a network designed based on Swin Transformer [49], which computes self-attention correlations by processing spatio-temporal inputs through a window shifting mechanism. Ma et al. [50] applied the Transformer to video saliency prediction and obtained the temporal

dependence of input past frames and target future frames by the Transformer and achieved excellent performance. Wang et al. [51] proposed a novel saliency detection algorithm applied to optical remote sensing images, which exploits the advantages of both CNN and Transformer to better extract local and global contextual information in complex scenes. Su et al. [52] proposed a unified Transformer framework for group-based segmentation, through which long-range dependencies between image features are captured and patch-structural similarities between related objects are explored. Zhou et al. [53] used Video Swin Transformer as a model backbone to generate multilevel spatio-temporal features with rich contextual cues.

In our module, Transformer is used to find out how structurally related the global features and attention features are. The Transformer module's inherent ability to interact with global features allows it to be used in combination with our proposed MLIA module to simulate long-range spatio-temporal dependencies at various information levels, significantly enhancing the model's performance.

3 Approach

3.1 Overall architecture

TMAI-Net is a two-stream structured model based on the Transformer module and MLIA which built by 3DCNN to model the human visual attention processes. Our model predicts the frame-by-frame saliency map by means of a sliding window. For a video with a total number of frames N, the saliency prediction of any frame in the video is achieved by considering a fixed number of consecutive past frames, which is referred to as T in our work. The input to this model is the original video frame F and the corresponding attentional patches A. In other words, S_t, a saliency map at t, is predicted given an input clip $(F_{t-T+1}, \ldots, F_t)$ and attentional patch $(A_{t-T+1}, \ldots, A_t)$ for any $t \in \{T, \ldots, N\}$ where F_t is the frame at time step t and A_t is the attention patch at time step t. S_t can be calculated using the following equation:

$$S_t = \begin{cases} Q\left(\{F_{t+T-i}\}_{i=1}^{T}, \{A_{t+T-i}\}_{i=1}^{T}\right) \cdots t \in (0, T) \\ Q\left(\{F_{t-T+i}\}_{i=1}^{T}, \{A_{t-T+i}\}_{i=1}^{T}\right) \cdots t \in [T, N] \end{cases} \tag{1}$$

where Q represents our TMAI-Net model.

The proposed model's structure and design are illustrated in Fig. 1. Our model employs the full convolutional component of the S3D network [17], which has been pre-trained on Kinetics data[18], as its foundational architecture. In the S3D network, our model implements the encoding of multi-level features by 3DCNN. For the four branches output from the backbone part of the model corresponds to low-level features and high-level features, respectively. At the lowest level (closest to the pixel) of the model we add the Multi-level Interactive Attention (MLIA) module based on a self-attention mechanism that implements long-range spatio-temporal relationships modeling. In addition, we add the Transformer module to mine the structural similarity associated with each video frame and the corresponding attention patch of the model input, which helps the model to focus more on the salient regions. Besides, the Transformer works together with the MLIA module to implement global spatio-temporal feature modeling at different levels. Lastly, the features output by the MLIA module and the Transformer module are combined and decoded in the decoder to output the saliency map.

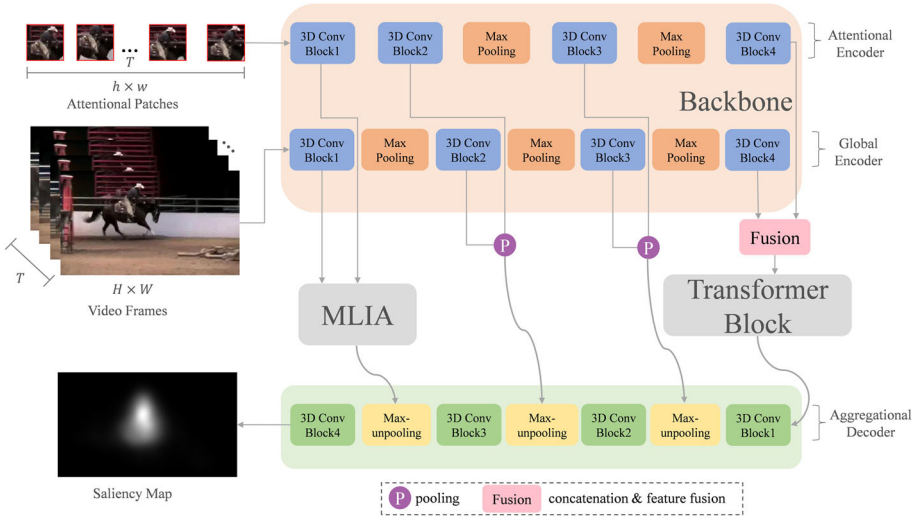

Fig. 1 Detailed structure of the TMAI-Net. The model contains two encoders (global encoder and attentional encoder) and a decoder, the Multi-level Interactive Attention (MLIA) module, and the Transformer module. The encoder encodes the input of the model to generate multi-level spatio-temporal features corresponding to four branches. MLIA and Transformer are placed at the shallowest and deepest levels of the four branches respectively to establish long-term spatio-temporal dependencies with different time steps. Then, the saliency map is generated by the decoder

3.2 Encoders and decoder

The encoder and decoder architecture used in [13] provides the basis of our TMAI-Net. To implement temporal modeling and pixel emphasis we introduce the explicit circular attention mechanism proposed by [19]. Both encoders of TMAI-Net use the S3D network as the basic structure. After that, we fuse the global features and attention features and input them into the decoder. The decoder performs spatial decoding of features while jointly aggregating temporal information to produce saliency maps.

The global encoder of TMAI-Net uses the S3D network as the underlying architecture, removing the final pooling, convolutional and fully connected layers to extract spatio-temporal features from multiple levels. The output $\left\{C_g^i\right\}_{i=1}^4$ of the four branches of the global encoder can be calculated using the following formula.

$$C_g^i = \begin{cases} R\left(F * w_g^i + b_g^i\right) & i = 1 \\ R\left(C_g^{i-1} * w_g^i + b_g^i\right) & i \neq 1 \end{cases}, \tag{2}$$

where $i \in (0, 4]$; $R(\cdot)$ represents the ReLU activation function, and w and b represent the weight and bias of the ith branch, respectively.

The attentional encoder of TMAI-Net reduces the max-pooling layer next to and inside the first convolutional layer compared to the global encoder. For the specific convolutional layers in the S3D network see [17].The output $\left\{C_a^i\right\}_{i=1}^4$ of the four branches of the attentional encoder can be calculated using the following formula.

$$C_a^i = \begin{cases} R\left(A * w_a^i + b_a^i\right) & i = 1 \\ R\left(C_a^{i-1} * w_a^i + b_a^i\right) & i \neq 1 \end{cases}, \tag{3}$$

The parameters in the formula for the attentional encoder are similar to those of the global encoder. At the lowest and highest level of the four branches extracted from the encoder we placed the MLIA and Transformer architecture for establishing long-range spatio-temporal interactions, respectively.

3.3 Multi-level Interactive Attention (MLIA) module

Video saliency prediction entails the challenge of emulating human visual attention in dynamic scenes, necessitating a comprehensive grasp of the contextual information present within the video. Such a task requires not only combining multiple levels of semantic information, but also capturing long-range relationships between visual features at different moments. In our model, after input pass through the model's backbone, the temporal channel dimension of the multilevel features is compressed to 4 by the 3D convolutional layer. Given the persistent limitations in existing video saliency prediction models concerning the learning of spatio-temporal feature correlations and saliency region identification, we propose a novel approach. In our model, we incorporate the MLIA module at the most granular level within the four branches originating from the model's backbone. This integration aims to effectively model long-range dependencies among time steps within the temporal channel at the pixel level.

We propose a Multi-level Interactive Attention module (MLIA) as shown in Fig. 2. Our MLIA module references the SATA module proposed in [37]. In method [37], the SATA module is placed for use on four branches on the backbone, enhancing the visual characteristics between the different time steps at different levels. However, the SATA module does not work in our approach because we have two feature streams and it does not make sense to consider the update of long-range spatio-temporal relationships for only one feature stream. In this case, it is necessary to consider how the fused auxiliary feature stream is processed by MLIA.

As shown in Fig. 2, MLIA contains a total of three sub-modules including a feature stream fusion module and two self-attention sub-layers. The inputs to the MLIA module are the outputs of the lowest level in the encoder of the TMAI-Net. In our model, the two input feature streams go through the feature stream fusion module in MLIA and enter into the two self-attention sublayers, which directly capture the long-range relationship between spatio-temporal features at different time steps through the dot product attention. The input

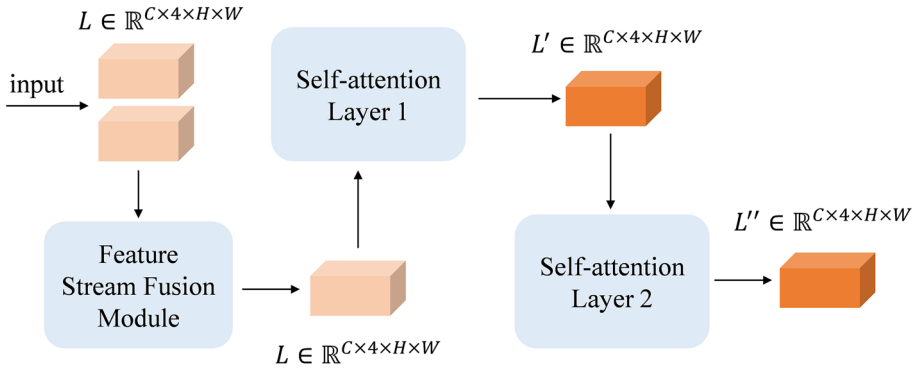

Fig. 2 The overall structure diagram of the Multi-level Interactive Attention (MLIA) module

of the MLIA module can be denoted as $L \in \mathbb{R}^{C\times 4\times H\times W}$, where C represents the semantic channel, 4 represents the temporal channel, and H and W denote the dimensions for height and width, respectively. In Self-attention Layer 1, as illustrated in Fig. 3(a), the input feature L is split into 4 sub-features along the time channel $\left\{L^1, L^2, L^3, L^4\right\} \in \mathbb{R}^{C\times 1\times H\times W}$. We convert the obtained sub-features into queries, keys and values and measure the two-by-two relationships by dot product attention. The global dependencies of spatio-temporal features are captured based on these relational aggregation information. The dot product attention is depicted in Figure 4. The specific computation is as follows:

$$DP - \text{Att}\left(L_q, L_k, L_v\right) = \text{Softmax}\left(\left(L_q\right)^T L_k\right)\left(L_v\right)^T \tag{4}$$

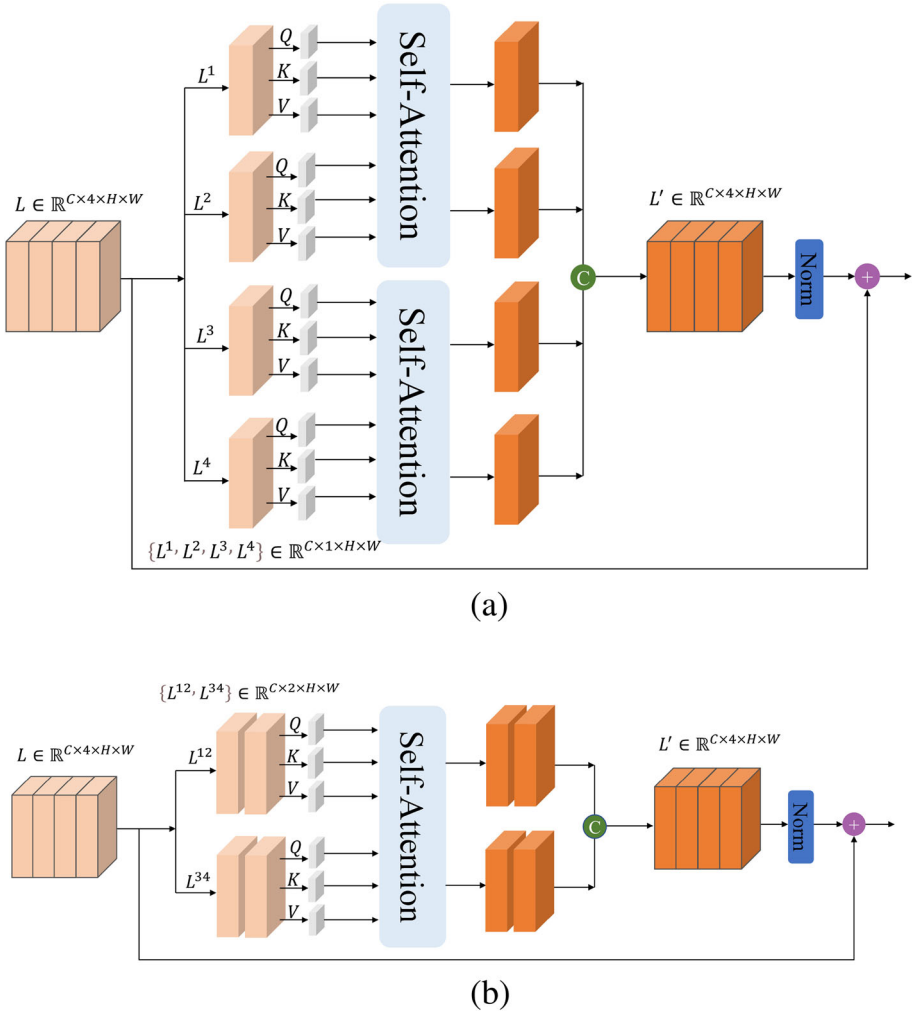

Fig. 3 The self-attention sub-layers of the Multi-level Interactive Attention (MLIA) module. (a) Self-attention Layer 1 of MLIA. (b) Self-attention Layer 2 of MLIA

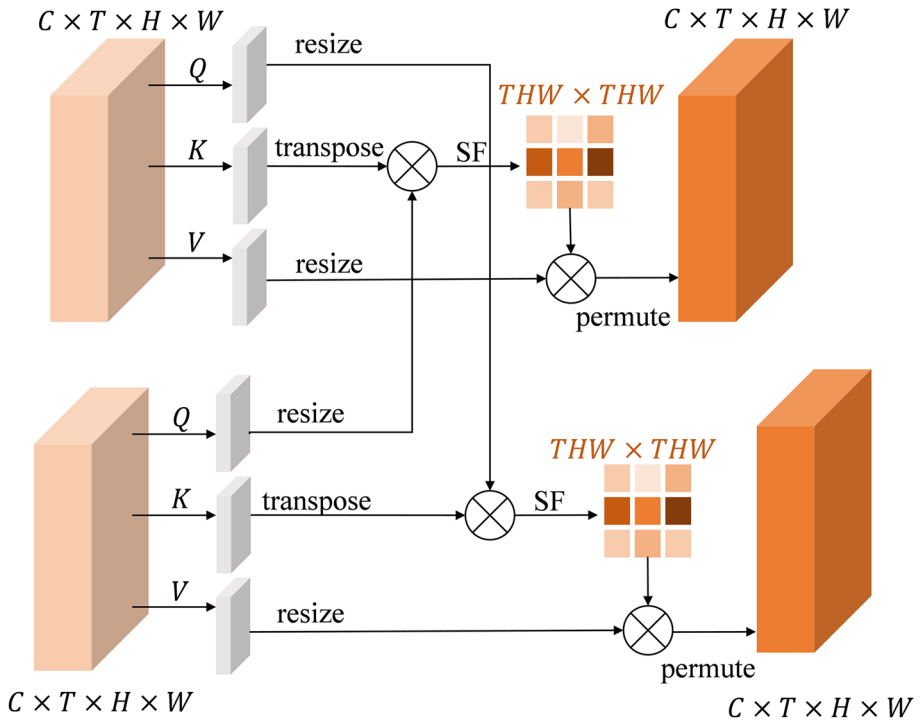

Fig. 4 The details of the dot product attention in the Self-attention layer. SF stands for the softmax operation

Where $DP-\mathrm{Att}(\cdot)$ stands for dot product attention, L_q, L_k, and L_v represent the query, key, and value after sub-feature transformation, respectively. $Softmax(\cdot)$ is the softmax activation function. Subsequently, they are recombined along the temporal channel. Unlike the Self-attention Layer 1, the Self-attention Layer 2, as depicted in Fig. 3(b), the input features are partitioned into two sub-features $\left\{L^{12}, L^{34}\right\} \in \mathbb{R}^{C\times 2\times H\times W}$. The subsequent operation is the same as the Self-attention Layer 1, where the long-range dependence of spatio-temporal features is established by dot product attention. By computing the dot product attention in two sub-layers of MLIA, it is possible to realize that any one of the four features $\left\{L^1, L^2, L^3, L^4\right\}$ can be updated directly by the long-range spatio-temporal relationship with the other three features.

In addition, the fusion of the two input feature streams needs to be implemented in the MLIA. The fused feature stream C_f can be computed using the following equation:

$$C_f = R\left(\text{Concate}\ (C_1, C_2) * w_f + b_f\right) \tag{5}$$

Where $R(\cdot)$ is the ReLU activation function, w and b represent the parameters corresponding to the weight and bias of the fusion module, $Concate(\cdot)$ represents concatenation operation of two feature streams, C_1 and C_2 represent the input video stream and the explicit cyclic attention stream, respectively. The MLIA module enables the modeling of spatio-temporal features with different time steps, which greatly improves the performance of the model.

3.4 Transformer block

The Transformer [32] has garnered substantial interest within the realm of computer vision due to its proven ability to capture long-range relationships. In our model, we include the Transformer block at the deepest level of the four branches drawn from encoder, which helps to leverage structural similarities between related objects in the attention and global features to better achieve pixel emphasis. Transformer is able to capture the long-range dependence of spatio-temporal features thanks to the self-attention mechanism. It and MLIA can model the long-term spatio-temporal relationship at different levels, which significantly improves the model performance.

The module of our Transformer block is shown in Fig. 5. Conventional Transformer models use an encoder and decoder architecture with stacked self-attention layers and point-wise, completely connected layers in both the encoder and the decoder. The encoder in our module consists of a stack of M identical layers, with two sub-layers in each layer. The first sub-layer involves a multi-head self-attention mechanism (MSA), while the second sub-layer consists of a fully connected feedforward network (FFN) equipped with a multi-layer perceptron (MLP). Two sub-layers are joined using residuals and then layer normalized (LN). For the

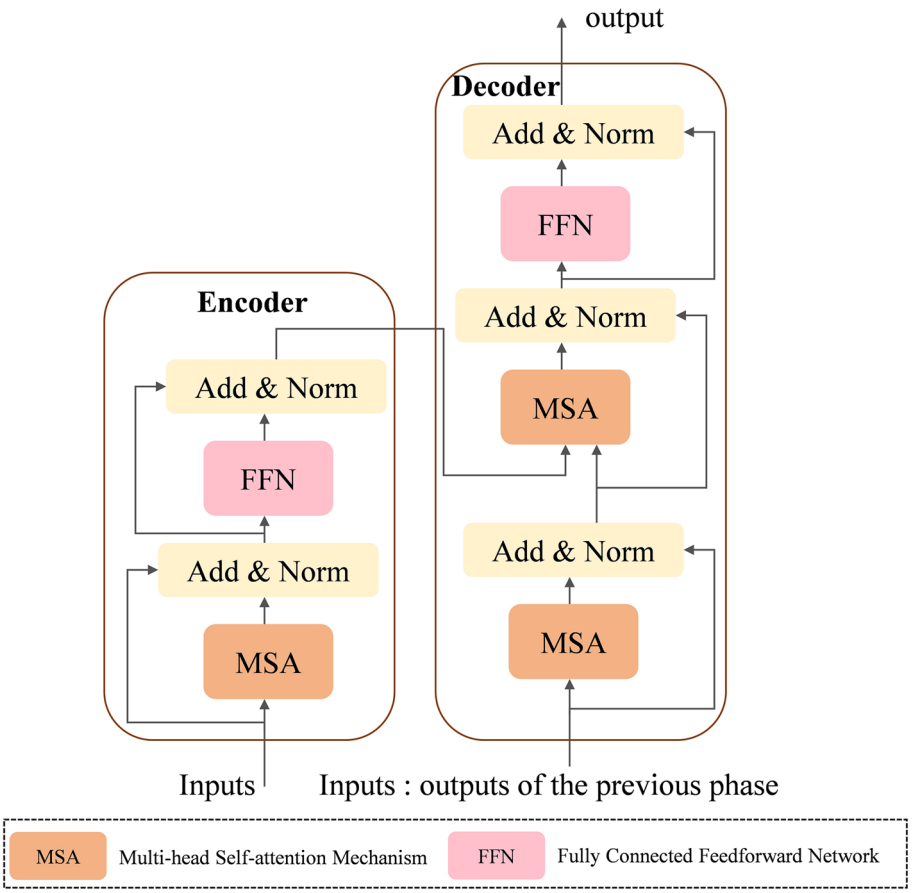

Fig. 5 Transformer-based feature interaction module

input token $X \in \mathbb{R}^{C \times L}$, the Transformer processes it through position encoding and two sub-layers of the encoder, as follows:

$$x_0 = [u_1 + p_1, u_2 + p_2, \ldots, u_L + p_L] \tag{6}$$

$$x'_m = \mathrm{MSA}\left(\mathrm{LN}\left(x_{m-1}\right)\right) + x_{m-1} \tag{7}$$

$$x_m = \mathrm{MLP}\left(\mathrm{LN}\left(x'_m\right)\right) + x'_m \tag{8}$$

$$y = LN\left(x_m\right) \tag{9}$$

where p_i represents the position embeddings for input tokens u_i, y refers to the output. M represents the number of encoders and $m \in \{1, \ldots, M\}$. MSA builds on the self-attention mechanism, which maps queries and sets of key-value pairs to outputs. Within the MSA, each set of quiries' attention function is calculated in parallel and combined into a matrix called Q. Similarly, matrices K and V, containing the keys and values, are packed accordingly. The MSA is calculated as follows:

$$\mathrm{Self-Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

$$H_i = \mathrm{Self-Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{11}$$

$$\mathrm{MSA}(Q, K, V) = \mathrm{Concat}\left(H_1, \ldots, H_h\right)W^O \tag{12}$$

where the parameter matrices $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$, and $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$. d_k represents the dimension of the query, d_v represents the dimension of the value, and d_{model} signifies the number of dimensions in the output generated by all sub-layers and embedding layers within the model. The decoder consists of M identical layers stacked like the encoder. The decoder adds a third sub-layer, which is utilized for multi-head attention on the encoder stack's output, in contrast to the encoder. Through the incorporation of the attention mechanism, the Transformer effectively captures long-range dependencies and alleviates the computational challenges associated with calculating self-attention relationships for extensive image and video entities.

In our module, both encoders use the S3D network as the basic structure and generate four distinct branches $\left\{C_g^i\right\}_{i=1}^4$ and $\left\{C_a^i\right\}_{i=1}^4$, respectively. We fuse the C_g^4 and C_a^4 from the 4-th branch of the two encoders into C_f. For $C_f \in \mathbb{R}^{B \times C \times 4 \times H \times W}$, in the Transformer module, where B represents the batch size and C, 4, H, and W stand for the semantic channel, temporal channel, height, and width, in that order. Concretely, we first reshape C_f into a sequence flattened tokens $C_t \in \mathbb{R}^{U \times B \times C}$, where $U = 4 * H * W$. All these token C_t are then entered into the Transformer block.

$$C_y = \mathrm{Transformer}\ (C_t) \tag{13}$$

where C_y represents the output of the Transformer block. Afterward, we reshape C_y back into an feature map $C'_y \in \mathbb{R}^{B \times C \times 4 \times H \times W}$.

4 Experiments

4.1 Datasets

DHF1K [23]: DHF1K is a comprehensive dataset tailored for dynamic free gaze prediction. It consists of videos with a 1000 frame-per-second frame rate and 640 × 360 resolution that feature a wide variety of scenes, actions, activities, and more. Each frame of the video was viewed from 17 observers. In the DHF1K dataset, 1000 videos are divided into 600 training sets, 100 validation sets and 300 test sets.

Hollywood-2 [54]: Hollywood-2 stands as one of the most extensive and demanding datasets accessible in this domain. It contains 1,707 videos featuring human behavior in Hollywood movies and the corresponding 19 viewers labeled with ground-truth saliency maps. These videos encompass a wide range of categories, including activities such as phone calls, driving, exiting a vehicle, handshakes, and running. The Hollywood-2 dataset has been divided into two sets for training and testing: a test set containing 884 sequences and a training set with 823 sequences.

UCFSports [54]: UCFSports contains 150 videos of various sports action categories such as diving, golf swing, weightlifting, horseback riding, walking, etc. Annotations for these videos have been gathered through a task-driven approach. Following the segmentation of UCFSports, 103 videos in total were assigned to the training set, while 47 videos constituted the test set.

4.2 Metrics

To ensure a more precise and equitable assessment of the model's performance, we employ five widely recognized metrics, in line with established prior research [13, 19, 54]: NSS, SIM, CC, AUC-J, and s-AUC. Specifically, NSS is used as a simple correspondence measure between the salient and true graphs to estimate the linear correlation between the anticipated result and the true gaze graph. CC is used to assess the degree of correlation or dependence between the significant plot and the gaze-point plot, also known as the linear correlation coefficient. SIM is a tool for comparing how similar two distributions are. AUC-J and s-AUC are both variants of AUC, they are both computed using binary maps of gaze points, and both are commonly used metrics for assessing saliency maps. Higher scores for the aforementioned measures indicate that the model is doing better.

4.3 Experimental setup

TMAI-Net is implemented based on Python and Pytorch [55] framework. For the input of the model, we selected a sliding window T size of 32. Every input frame has a size of 224 × 384, and the attention patches that match the input frames have a size of 28 × 48. The entire model is trained using the SGD optimizer. We set the model's learning rate to 0.001 and use an early stop strategy on the validation set to prevent overfitting during the training phase. The total number of iterations during training is set at 4000, and the batch size is set at 30.

First, we train our model using the DHF1K training set. We must benchmark the model's results on the test set online because DHF1K retains the test set's annotations. Then, we trained the models on Hollywood-2 and UCFSports, respectively. Since DHF1K contains more types

of objects and scenes compared to Hollywood-2 and UCFSports, it has advantages in diversity, scalability, and generality of datasets, so we mainly evaluate the model on DHF1K and use the remaining two datasets as supplements.

In video saliency prediction tasks, the Kullback-Leibler divergence (KLD) [7] has demonstrated its efficacy as a loss function and is extensively employed in the field [13, 19, 54, 56]. Following are the calculations for the KL loss function:

$$KL(S, G) = \sum_i G_i \log\left(\epsilon + \frac{G_i}{\epsilon + S_i}\right) \tag{14}$$

Where S signifies the predicted saliency map, G represents the ground truth. ϵ denotes the regularization constant.

4.4 Comparison with the state-of-the-art models

We perform a quantitative comparative analysis of our TMAI-Net against 12 state-of-the-art video saliency prediction models, which encompass STSConvNet [6], SALICON [7], OM-CNN [10], ACLNet [11], STRA-Net [56], TASED-Net [13], SALSAC [12], UNISAL [15], ECANet [19], ViNet [30], STSANet [37], and TMFI-Net [53]. This evaluation is carried out using the DHF1K [23] dataset as well as the test sets of Hollywood-2 [54] and UCFSports [54]. We selected the aforementioned models after analyzing their structure. TASED-Net was chosen because we adopted this model as the basic structure of TMAI-Net. ECANet is chosen because we use its proposed two-stream input approach based on explicit cyclic attention. STSConvNet,OM-CNN, and SALICON were chosen because they are all two-stream structures like TMAI-Net. STSANet and TMFI-Net were chosen because they are currently the two best-performing models in the field of video saliency prediction. The other selected models are also cutting-edge models in the field of video saliency prediction at present. By evaluating our models with the above models, we can fairly, comprehensively, and accurately assess the efficacy of TMAI-Net.

Quantitative. Table 1 presents the quantitative results for these models across the five metrics. Comparing our TMAI-Net against the other 12 state-of-the-art models, it obtained the top 3 results in 4 out of 5 metrics, suggesting that it is competitive in the DHF1K dataset. Although our model is not currently the best performing in the field of video saliency prediction, we far outperform the most advanced models in terms of computational efficiency (model size and running speed). Table 2 shows specific experiments on computational efficiency. The experimental results obtained from the Hollywood-2 and UCFSports datasets illustrate that our TMAI-Net outperforms both TASED-Net and ECANet across all metrics. This illustrates the value of our model.

To assess the model's performance on the Hollywood-2 and UCFSports datasets more fairly and accurately, we also need to make some additional adjustments. Our model simulates the behavior of the human visual attention mechanism in modeling virtual memories, which requires sufficient temporal information. Nonetheless, it's important to note that a few samples within these two datasets lacked an adequate number of frames for our model to effectively model virtual memories. Additionally, certain samples within Hollywood-2 and UCFSports consisted of only one or two images, which resulted in outcomes that didn't meet our anticipated performance levels. To address this issue, we selected samples from both datasets with more than 64 frames to reevaluate our model so that our model has enough frames to learn saliency patterns and model virtual memories in human visual scenes. The

Table 1 Results of a quantitative comparison between TMAI-Net and other state-of-the-art models using the DHF1K, Hollywood-2, and UCFSports test sets

Model	DHF1K					Hollywood-2					UCFSports				
	AUC-J	SIM	s-AUC	CC	NSS	AUC-J	SIM	s-AUC	CC	NSS	AUC-J	SIM	s-AUC	CC	NSS
STSConvNet	0.834	0.197	0.581	0.325	1.632	0.863	0.276	0.710	0.382	1.748	0.832	0.264	0.685	0.343	1.753
SALICON	0.857	0.232	0.590	0.327	1.901	0.586	0.321	0.711	0.425	2.013	0.848	0.304	0.738	0.375	1.838
OM-CNN	0.856	0.256	0.583	0.344	1.911	0.887	0.356	0.693	0.446	2.313	0.870	0.321	0.691	0.405	2.089
ACLNet	0.890	0.315	0.601	0.434	2.354	0.913	0.542	0.757	0.623	3.086	0.897	0.406	0.744	0.510	2.567
STRA-Net	0.895	0.355	0.663	0.458	2.558	0.923	0.536	0.774	0.662	3.478	0.910	0.479	0.751	0.593	3.018
TASED-Net	0.895	0.361	0.712	0.470	2.667	0.918	0.507	0.768	0.646	3.302	0.899	0.469	0.752	0.582	2.920
SALSAC	0.896	0.357	0.697	0.479	2.673	0.931	0.529	0.712	0.670	3.356	0.926	0.534	0.806	0.671	3.523
UNISAL	0.901	0.390	0.691	0.490	2.776	0.934	0.542	0.759	0.673	3.901	0.918	0.523	0.775	0.644	3.381
ECANet	0.903	0.385	0.717	0.500	2.814	0.929	0.526	**0.806**	0.673	3.380	0.917	0.498	0.797	0.636	3.189
ViNet	0.908	0.381	**0.728**	0.510	2.870	0.930	0.550	**0.813**	0.693	3.730	0.924	0.522	**0.810**	0.673	3.620
STSANet	**0.913**	0.383	0.723	**0.529**	**3.010**	**0.938**	**0.579**	–	**0.721**	**3.927**	**0.936**	**0.560**	–	**0.705**	**3.908**
TMFI-Net	**0.915**	**0.407**	**0.731**	**0.546**	**3.146**	**0.940**	**0.607**	–	**0.739**	**4.095**	**0.936**	**0.565**	–	**0.707**	**3.863**
TMAI-Net	0.908	**0.391**	0.714	0.511	2.879	0.923	0.542	0.797	0.668	3.434	0.919	0.494	0.797	0.636	3.197
TMAIlong	–	–	–	–	–	0.923	0.551	**0.806**	0.678	3.480	**0.929**	0.521	**0.807**	0.689	3.542

The top two have been highlighted in **bold**. The 3rd rank in the TMAI-Net results has been underlined. The Hollywood-2 and UCFSports dataset comparison results are indicated by TMAIlong

Table 2 Model size (MB) and average prediction time comparison of video saliency methods

Model	Size(MB)	Time
ACLNet	250	0.02
STRA-Net	641	0.02
TASED-Net	82	0.06
SalSAC	93	0.02
UNISAL	15	0.009
ViNet	124	0.016
STSANet	643	0.035
TMFI-Net	234	0.033
TMAI-Net	207	0.018

results of the model re-evaluation based on this method are shown in Table 1, and "TMAI-long" is the result of the experiment. The outcomes demonstrate that, in both the Hollywood-2 and UCFSports datasets, our model can rank in the top 3 for a wide range of metrics.

Qualitative. In Fig. 6, we have selected some representative advanced video saliency prediction models for qualitative comparison with our proposed model, including STRA-Net [56], TASED-Net [13], and ViNet [30]. We use the DHF1K dataset's validation set to qualitatively evaluate our model. The aforementioned procedures are carried out on identical hardware setups to guarantee the impartiality of the evaluation.

The experimental results show that our model performs more accurately than other models. In Fig. 6 (a), a woman is pushing a floating ball with a child into the pool. In this case, due to the interaction between the woman and the ball, attention will transition from the woman to the floating ball, and the ground-truth maps record this attention shift. Among all models, only TMAI-Net and TASED-Net predicted the attentional shift from the woman to the floating ball. Compared with TASED-Net, the prediction result of TMAI-Net is more accurate. In Figure 6 (b), a man is cutting a stone with a machine. As he moves the tray to cut the stone, the ground-truth map shows that the viewer's attention shifts from the machine to the man. Among all the models, only TMAI-Net provides the most accurate prediction. In Figure 6 (c), a large dog is walking toward a small dog in the shot, and as time passes, the ground-truth map shows the viewer's attention moving from the large dog to the small dog. Among all the models, the prediction result of TMAI-Net is closest to the ground-truth map. In Fig. 6 (d), two men are fishing and one of the men casts his rod into the water. As time passes, attention transitions from the men to the rod and subsequently to the fish. In this case, only TMAI-Net accurately predicts this process, ViNet ignores the attention transfer process, and TASED-Net and STRA-Net pay unnecessary attention to the fishing boat.

Computational load. We compare the computational efficiency of our model to a number of state-of-the-art models, such as ACLNet [11], STRA-Net [56], TASED-Net [13], SalSAC [12], UNISAL [15], ViNet [30], STSANet [37], and TMFI-Net [53], in order to assess the computational efficiency of the model. Table 2 makes it evident that our model exhibits competitiveness. In addition, our model uses a two-stream input, and for a 224×384 video frame and its corresponding 28×48 attention patch. our model takes about 0.018s. The computational efficiency of our model is significantly better than that of the two most advanced video saliency prediction models STSANet and TMFI-Net.

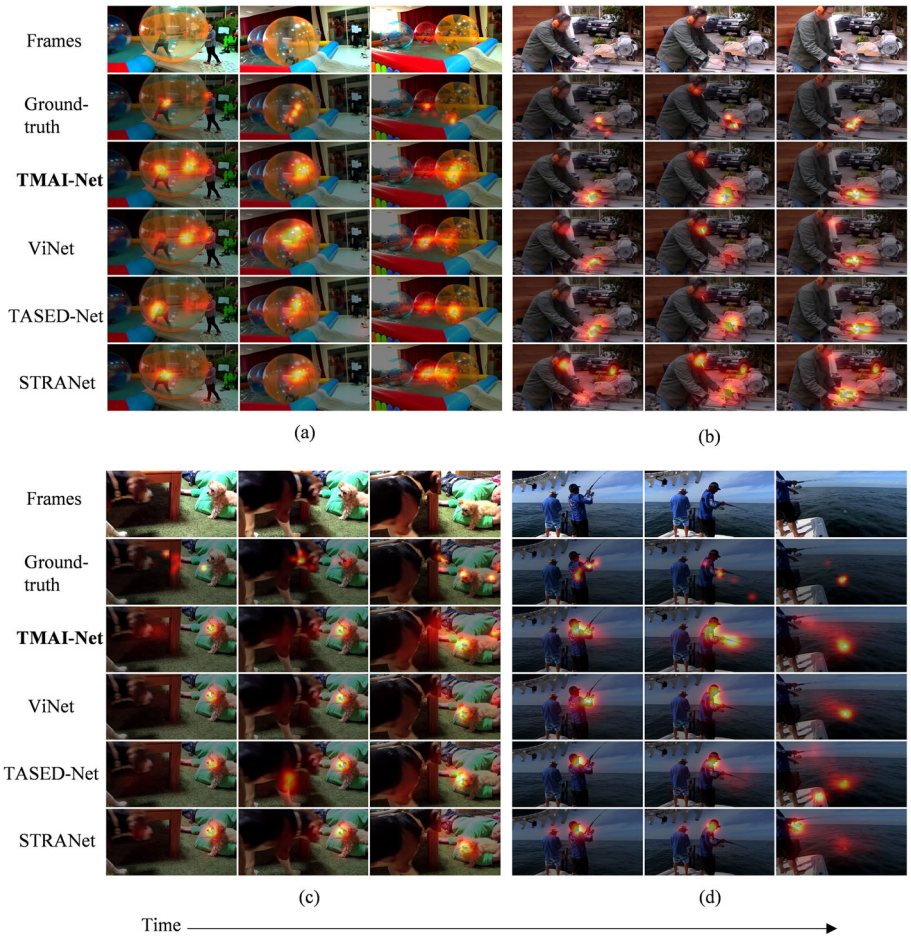

Fig. 6 Qualitative comparisons on several video categories, each sampling three frames for display, between the TMAI-Net and other state-of-the-art video saliency models

4.5 Ablation studies

In this section, we use the DHF1K dataset to conduct an ablation analysis of TMAI-Net. First, various variants of the model were built in order to more thoroughly examine the applicability of each component. Specifically, our model offers three settings, including "Two-stream", "Two-stream+MLIA" and "Two-stream+Transformer". "Two-Stream" means that a two-stream network is constructed using the 3D backbone as a baseline. "Two-stream+MLIA" means adding the MLIA to "Two-stream". "Two-stream+Transformer" means to add Transformer module to "Two-stream". Table 3 displays the outcomes of the ablation experiments. "Ours" refers to our model TMAI-Net.

The Contributions of the MLIA module. The difference in results between "Two-stream" and "Two-stream+MLIA" in Table 3 shows that the addition of the MLIA to "Two-stream" improves the performance of four (AUC-J, s-AUC, CC, NSS) of the five

Table 3 Ablation study on MLIA and and Transformer module

Model	AUC-J	SIM	s-AUC	CC	NSS
Two-stream	0.910	0.394	0.725	0.515	2.877
Two-stream+MLIA	0.912	0.388	0.726	0.519	2.912
Two-stream+Transformer	0.913	0.386	0.724	0.520	2.906
Ours	0.914	0.399	0.723	0.524	2.937

metrics. This demonstrates that the MLIA can indeed bring performance improvement to our model.

The Contributions of the Transformer module. Comparing the performance of "Two-stream" and "Two-stream+Transformer", we can see that the performance of three(AUC-J, CC, NSS) of the five metrics improves when "Two-stream" is added with the Transformer module. This finding implies that the Transformer module in the model does bring performance gains.

The Contributions of the MLIA module and Transformer module. Comparing the performance of "Ours" and "Two-stream+MLIA", we can see that four (SIM, s-AUC, CC, NSS) of the five metrics have improved when the Transformer module is added to "Two-stream+MLIA". In addition, comparing the performance of "Ours" and "Two-stream+Transformer", we can see that four (SIM, s-AUC, CC, NSS) of the five metrics have performance gains when the MLIA is added to "Two-stream+Transformer". This proves that both the MLIA and the Transformer module bring performance gains to the model.

Ablation Study on MLIA. Several variants of the MLIA module were created in order to further validate its contribution to the model. In our model, the MLIA module is placed at the lowest level (closest to the pixel) of the four branches extracted from the encoder to establish long-range spatio-temporal dependencies at the pixel level. In order to study and demonstrate the importance of pixel-level spatiotemporal dependence, we designed three Settings, as shown in Table 4, including "Setting 1"," Setting 2 "and" Setting 3", which represent placing the MLIA at levels 2, 3 and 4 of the four branches of the model, respectively. "setting4" stands for our MLIA. The experimental results showed that the metrics of "setting4" were significantly better than those of other MLIA module variants. This demonstrates the superiority of placing the MLIA module at the lowest level of the model to establish long-range spatio-temporal relationships at the pixel level.

The visual comparative analyses in ablation studies. In order to more intuitively illustrate the effectiveness of each component in the model, we conducted a visual comparative analysis experiment on the variants of each TMAI-Net designed in the ablation experiment,

Table 4 Ablation study on MLIA module

Model	AUC-J	SIM	s-AUC	CC	NSS
setting1	0.909	0.395	0.727	0.515	2.890
setting2	0.909	0.390	0.728	0.515	2.891
setting3	0.910	0.359	0.704	0.493	2.733
setting4	0.912	0.387	0.726	0.519	2.912

and the experimental results are shown in Fig. 7. The experimental results show that the structure in TMAI-Net is more accurate than other variants. As shown in Fig. 7(a) and (b), only TMAI-Net correctly indicates that the saliency region is on the big dog's head and not its belly in frame 3 in (a), and in (b), all variants except TMAI-Net incorrectly labelled the saliency region on the fishing boat. The same difference in results is shown in (c) (d) as in (a) (b).

4.6 Failure cases and analyses

While our model generally demonstrates strong performance, there are instances where it encounters challenges. Here, we present a few scenarios that TMAI-Net was unable to handle, along with their resolutions.

The failure cases illustrated in Fig. 8 highlight instances in which our model did not yield satisfactory prediction results. 1) The area of saliency is particularly similar to the background in terms of color or texture. In Fig. 8(a), since the swimming fish and the deep sea's blue backdrop resemble one another quite a bit, it is challenging for TMAI-Net to determine the exact saliency region. 2) Accurate prediction of saliency results for human postures. In Fig. 8(b), for the dancer's dancing limbs or head, TMAI-Net cannot accurately predict the

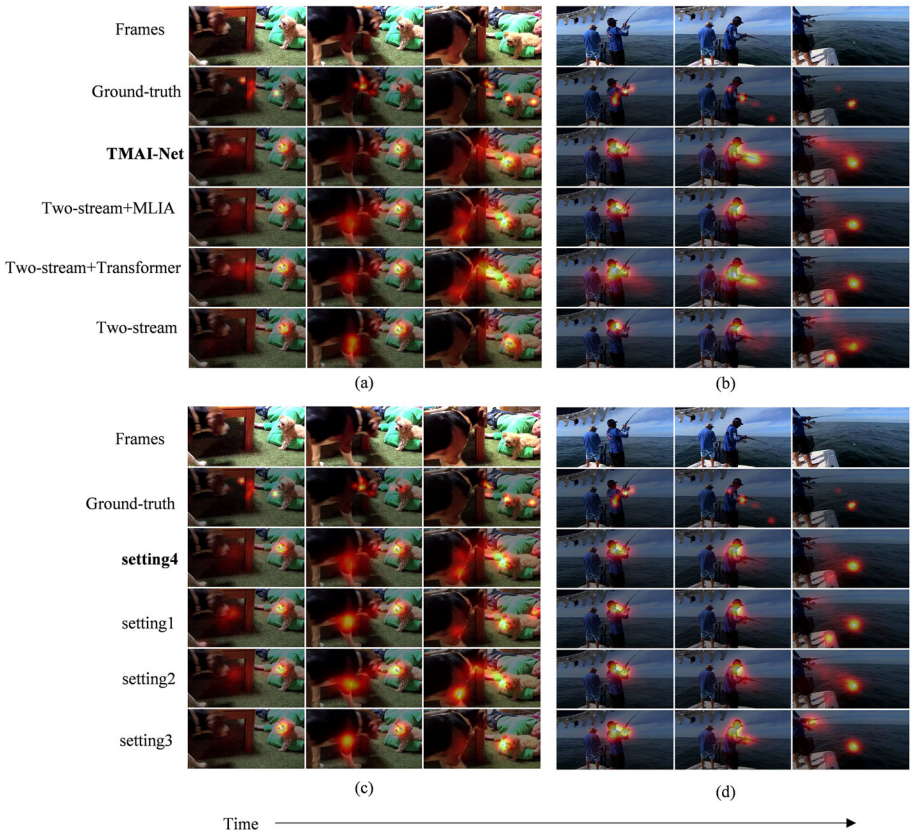

Fig. 7 The visual comparative analyses in ablation studies

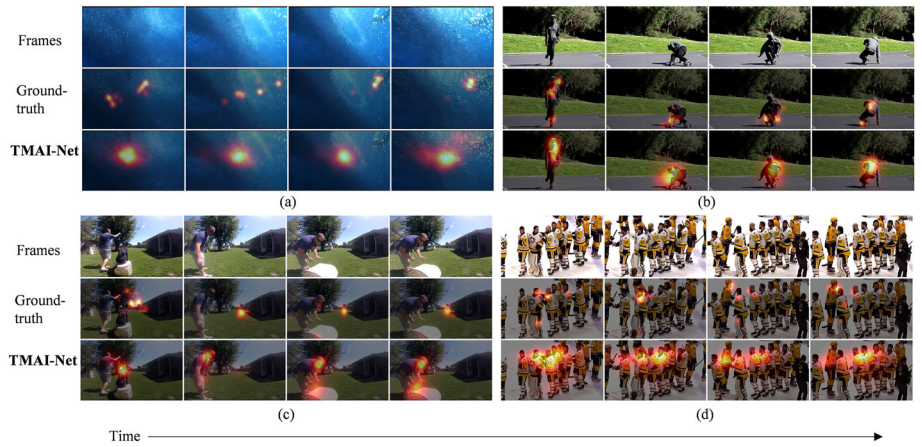

Fig. 8 Several cases of our method failing on the DHF1K dataset

viewer's area of attention while watching the dancer dance. 3) Blurred saliency areas in the shot. In Fig. 8(c), a man is teasing a dog with a dog teaser. When the dog teaser is thrown, the ground-truth map shows that the viewer's attention should be drawn to the distant dog teaser. The ground-truth map reveals that the viewer's focus is diverted to the far-off dog when it goes to pick up the faraway stick. However, TMAI-Net's prediction results in the attention being drawn to the man in the near distance. 4) There are a lot of things that are moving in the video. There are a lot of moving persons in the movie in Fig. 8(d). The person whose motion is most obvious draws the attention of the human eye. However, the results of the computational model are scattered to the others. The large number of moving objects in the video makes it difficult for the model to generate accurate predictions.

By analysing the above failure cases, we found that these scenarios are not common in the dataset. We can add more similar cases in the training set to improve the accuracy of the model in these scenarios. In addition to this, for cases (a) and (b), there exist specialised directions in object detection to be researched, i.e., video camouflage object detection and human pose estimation. In order to improve the accuracy of case (a) and case (b), we can introduce the solution of these two directions in our model in the future. For cases (c) and (d), the complex and changing scene and multiple moving objects make it difficult for the model to accurately localise the saliency region. This is because TMAI-Net can only infer saliency results from a video clip and cannot understand the contextual information in the video as humans do, and it is difficult for the model to capture the interactions between moving objects. How to make the model understand the contextual information conveyed by the video more accurately is a worthwhile research problem in our future work.

5 Conclusion

In this paper, a novel Transformer-based Multi-level Attention Integration Network (TMAI-Net) for video saliency prediction is proposed. We propose a Multi-level Interactive Attention (MLIA) module that can capture long-range dependencies among the temporal channel's time steps. In the MLIA module, the long-range spatio-temporal dependency update is achieved by computing dependencies at different time steps using the self-attention mechanism. Based on

Transformer's inherent global feature interaction capabilities, we add MLIA and Transformer to the shallowest and deepest of the model's four branches extracted from the encoder, directly establishing the global context at different levels. Furthermore, for the model's two-stream input, we add the Transformer module to mine their structural similarity between related objects and make the model focus more on saliency regions in the video. Comprehensive experiments have demonstrated that TMAI-Net is competitive with current state-of-the-art approaches. The ablation experiment's findings offer convincing proof of the efficiency of each TMAI-Net component.

Acknowledgements This study has been partially supported by Program of Science and Technology Development Plan of Jilin Province of China (20220201147GX,20240101374JC) and Shenzhen Technology R&D Program (JCYJ20230807150300001) and the Graduate Innovation Fund of Jilin University(2023CX206).

Author Contributions Rui Tan: Writing - original draft, Writing - review & editing, Methodology, Software. Minghui Sun: Conceptualization, Resources, Supervision, Project administration. Yanhua Liang: Writing - review & editing.

Data Availability No datasets were generated in this study. The dataset analyzed in this study is publicly available.
following public domain resources:
https://github.com/wenguanwang/DHF1K DHF1K DataSets
https://www.di.ens.fr/~ laptev/actions/hollywood2 Hollywood-2 DataSets
https://www.crcv.ucf.edu/data/UCF_Sports_Action.php UCF Sports

Declarations

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

References

1. Hadizadeh H, Bajić IV (2013) Saliency-aware video compression. IEEE Trans Image Process 23(1):19–33
2. Zhu S, Liu C, Xu Z (2019) High-definition video compression system based on perception guidance of salient information of a convolutional neural network and hevc compression domain. IEEE Trans Circuits Syst Video Technol 30(7):1946–1959
3. Guo C, Zhang L (2009) A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans Image Process 19(1):185–198
4. Guraya FFE, Cheikh FA, Tremeau A, Tong Y, Konik H (2010) Predictive saliency maps for surveillance videos. In: 2010 Ninth international symposium on distributed computing and applications to business, engineering and science, pp 508–513. IEEE
5. Yubing T, Cheikh FA, Guraya FFE, Konik H (2011) Trémeau A (2011) A spatiotemporal saliency model for video surveillance. Cognitive Computation 3:241–263
6. Bak C, Kocak A, Erdem E, Erdem A (2017) Spatio-temporal saliency networks for dynamic saliency prediction. IEEE Trans Multimed 20(7):1688–1698
7. Huang X, Shen C, Boix X, Zhao Q (2015) Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: Proceedings of the IEEE international conference on computer vision, pp 262–270
8. Kocak A, Erdem E, Erdem A (2021) A gated fusion network for dynamic saliency prediction. IEEE Trans Cogn Dev Sys 14(3):995–1008
9. Zhang K, Chen Z (2018) Video saliency prediction based on spatial-temporal two-stream network. IEEE Trans Circuits Syst Video Technol 29(12):3544–3557
10. Jiang L, Xu M, Wang Z (2017) Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm. arXiv:1709.06316

11. Wang W, Shen J, Guo F, Cheng M-M, Borji A (2018) Revisiting video saliency: A large-scale benchmark and a new model. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4894–4903
12. Wu X, Wu Z, Zhang J, Ju L, Wang S (2020) Salsac: A video saliency prediction model with shuffled attentions and correlation-based convlstm. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 12410–12417
13. Min K, Corso JJ (2019) Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2394–2403
14. Chang Q, Zhu S (2021) Temporal-spatial feature pyramid for video saliency detection. arXiv:2105.04213
15. Droste R, Jiao J, Noble JA (2020) Unified image and video saliency modeling. In: computer vision–ECCV 2020: 16th european conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pp 419–435. Springer
16. Yuan M, Xu D (2023) Spatio-temporal feature pyramid interactive attention network for egocentric gaze prediction. IEEE Transactions on Circuits and Systems for Video Technology
17. Xie S, Sun C, Huang J, Tu Z, Murphy K (2018) Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the european conference on computer vision (ECCV), pp 305–321
18. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308
19. Xue H, Sun M, Liang Y (2022) Ecanet: Explicit cyclic attention-based network for video saliency prediction. Neurocomputing 468:233–244
20. Mahadevan V, Vasconcelos N (2009) Spatiotemporal saliency in dynamic scenes. IEEE Trans Pattern Anal Mach Intell 32(1):171–177
21. Fang Y, Wang Z, Lin W, Fang Z (2014) Video saliency incorporating spatiotemporal cues and uncertainty weighting. IEEE Trans Image Process 23(9):3910–3921
22. Li G, Xie Y, Wei T, Wang K, Lin L (2018) Flow guided recurrent neural encoder for video salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3243–3252
23. Wang W, Shen J, Xie J, Cheng M-M, Ling H, Borji A (2019) Revisiting video saliency prediction in the deep learning era. IEEE Trans Pattern Anal Mach Intell 43(1):220–237
24. Liu Y, Xiong Z, Yuan Y, Wang Q (2023) Distilling knowledge from super resolution for efficient remote sensing salient object detection. IEEE Transactions on Geoscience and Remote Sensing
25. Liu Y, Xiong Z, Yuan Y, Wang Q (2023) Transcending pixels: Boosting saliency detection via scene understanding from aerial imagery. IEEE Transactions on Geoscience and Remote Sensing
26. Liu Y, Cheng D, Zhang D, Xu S, Han J (2024) Capsule networks with residual pose routing. IEEE Transactions on Neural Networks and Learning Systems
27. Liu Y, Zhang D, Zhang Q, Han J (2021) Part-object relational visual saliency. IEEE Trans Pattern Anal Mach Intell 44(7):3688–3704
28. Liu Y, Zhou L, Wu G, Xu S, Han J (2023) Tcgnet: Type-correlation guidance for salient object detection. IEEE Transactions on Intelligent Transportation Systems
29. Liu Y, Dong X, Zhang D, Xu S (2024) Deep unsupervised part-whole relational visual saliency. Neurocomputing 563:126916
30. Jain S, Yarlagadda P, Jyoti S, Karthik S, Subramanian R, Gandhi V (2021) Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In: 2021 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 3520–3527. IEEE
31. Zou W, Zhuo S, Tang Y, Tian S, Li X, Xu C (2021) Sta3d: Spatiotemporally attentive 3d network for video saliency prediction. Pattern Recogn Lett 147:78–84
32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Advances in Neural Information Processing Systems 30
33. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
34. Nawaz HS, Shi Z, Gan Y, Hirpa A, Dong J, Zheng H (2022) Temporal moment localization via natural language by utilizing video question answers as a special variant and bypassing nlp for corpora. IEEE Trans Circuits Syst Video Technol 32(9):6174–6185
35. Huang J, Zhou W, Li H, Li W (2018) Attention-based 3d-cnns for large-vocabulary sign language recognition. IEEE Trans Circuits Syst Video Technol 29(9):2822–2832
36. Oh SW, Lee J-Y, Xu N, Kim SJ (2019) Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9226–9235

37. Wang Z, Liu Z, Li G, Wang Y, Zhang T, Xu L, Wang J (2021) Spatio-temporal self-attention network for video saliency prediction. IEEE Transactions on Multimedia
38. Zhang Y, Zhang T, Wu C, Tao R: Multi-scale spatiotemporal feature fusion network for video saliency prediction. IEEE Transactions on Multimedia (2023)
39. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929
40. Srinivas A, Lin T-Y, Parmar N, Shlens J, Abbeel P, Vaswani A (2021) Bottleneck transformers for visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16519–16529
41. Xu C, Makihara Y, Li X, Yagi Y, Lu J (2020) Cross-view gait recognition using pairwise spatial transformer networks. IEEE Trans Circuits Syst Video Technol 31(1):260–274
42. Yuan L, Chen Y, Wang T, Yu W, Shi Y, Jiang Z-H, Tay FE, Feng J, Yan S (2021) Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 558–567
43. Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, Fu Y, Feng J, Xiang T, Torr PH, et al (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6881–6890
44. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision, pp 213–229. Springer
45. Yuan Z, Song X, Bai L, Wang Z, Ouyang W (2021) Temporal-channel transformer for 3d lidar-based video object detection for autonomous driving. IEEE Trans Circuits Syst Video Technol 32(4):2068–2078
46. Zhao L, Guo J, Xu D, Sheng L (2021) Transformer3d-det: Improving 3d object detection by vote refinement. IEEE Trans Circuits Syst Video Technol 31(12):4735–4746
47. Sun Z, Cao S, Yang Y, Kitani KM (2021) Rethinking transformer-based set prediction for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3611–3620
48. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H (2022) Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3202–3211
49. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
50. Ma C, Sun H, Rao Y, Zhou J, Lu J (2022) Video saliency forecasting transformer. IEEE Trans Circuits Syst Video Technol 32(10):6850–6862
51. Wang Q, Liu Y, Xiong Z, Yuan Y (2022) Hybrid feature aligned network for salient object detection in optical remote sensing imagery. IEEE Trans Geosci Remote Sens 60:1–15
52. Su Y, Deng J, Sun R, Lin G, Su H, Wu Q (2023) A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. IEEE Transactions on Multimedia
53. Zhou X, Wu S, Shi R, Zheng B, Wang S, Yin H, Zhang J, Yan C (2023) Transformer-based multi-scale feature integration network for video saliency prediction. IEEE Transactions on Circuits and Systems for Video Technology
54. Mathe S, Sminchisescu C (2014) Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(7):1408–1424
55. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al (2019) Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 32
56. Lai Q, Wang W, Sun H, Shen J (2019) Video saliency prediction using spatiotemporal residual attentive networks. IEEE Trans Image Process 29:1113–1126