Cognitive Computation (2023) 15:856–868
https://doi.org/10.1007/s12559-023-10114-x

Human Vision Attention Mechanism-Inspired Temporal-Spatial Feature Pyramid for Video Saliency Detection

Qinyao Chang[1] · Shiping Zhu[1]

Received: 28 April 2022 / Accepted: 14 January 2023 / Published online: 22 February 2023

Abstract

Inspired by the human vision attention mechanism, the human vision system uses multilevel features to extract accurate visual saliency information, so multilevel features are important for saliency detection. On the basis of the numerous biological frameworks for visual information processing, we find that better combination and use of multilevel features with time information can greatly improve the accuracy of the video saliency model. The proposed TSFP-Net has the advantages of much higher prediction precision, simple structure, second smallest size, and the third fastest running time compared to the state-of-the-art methods. The encoder extracts multiscale temporal-spatial features from the input continuous video frames and then constructs a temporal-spatial feature pyramid through temporal-spatial convolution and top-down feature integration. The decoder performs hierarchical decoding of temporal-spatial features from different scales and finally produces a saliency map from the integration of multiple video frames. Our model is simple yet effective and can run in real time. We perform abundant experiments, and the results indicate that the well-designed structure can significantly improve the precision of video saliency detection. Experimental results on three purely visual video saliency benchmarks demonstrate that our method outperforms the existing state-of-the-art methods.

Keywords Video saliency · 3D convolution · Temporal-spatial feature pyramid

Introduction

Video saliency detection aims to predict the point of fixation for the human eye while watching videos freely. Visual saliency detection imitates the visual attention mechanism of the human visual system and is a typical computer vision task that is motivated by a biologically inspired basis. Visual saliency information acting as the interest region can be integrated with the original image and video information. Video saliency detection is an important and fundamental mechanism in computer vision tasks, such as intelligent important person/scene capturing and tracking, photo salient region enhancement, salient object segmentation, and video compression. It is widely applied in many areas, such as video compression [1, 2], video surveillance [3, 4], and video captioning [5].

Most existing video saliency detection models employ the encoder-decoder structure and rely on temporal recurrence to predict video saliency. For example, Wang et al. proposed ACLNet [6], which encodes static saliency features through an attention mechanism and then learns dynamic saliency through ConvLSTM [7]. Linardos et al. proposed SalEMA [8], which uses an exponential moving average instead of LSTM to extract temporal features for video saliency detection. Wu et al. proposed SalSAC [9], which proposes a correlation-based ConvLSTM to balance the alteration of saliency caused by the change in image characteristics of the past frame and current frame. However, such a saliency modeling approach has the following problems.

First, the spatial saliency model is pretrained on the static image saliency datasets before fine-tuning on the video saliency datasets. However, the effectiveness of this transfer learning mechanism may be limited since the resolutions of the two datasets are different, while saliency is greatly influenced by the image shape. Second, restricted by memory, the training of the video saliency model requires extracting continuous video frames from the datasets randomly. However, the approach based on LSTM needs to utilize

✉ Shiping Zhu
spzhu@163.com

[1] Department of Measurement Control and Information Technology, School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China

backpropagation through time to predict the video saliency of each frame. In this way, the state of LSTM of the first frame for the selected clip must be void, while during the test, only the state of the LSTM of the first frame of the video is void; such discrepancy makes the modeling of the method based on LSTM insufficient. Third, as mentioned by Min [10], all the methods based on LSTM overlay the temporal information on top of the spatial information and fail to utilize both kinds of information at the same time, which is crucial for video saliency detection.

To alleviate the above problems, some methods employ 3D convolutions to continuously aggregate the temporal and spatial cues of videos [10–12]. While they achieve outstanding performance, there still remains an important issue, that is, the lack of utilization of multilevel features. Multilevel features are essential for the task of saliency detection since the human visual mechanism is complicated and the concerned region is determined by various factors and from multiple levels. For example, some large objects may be salient, which are captured from deeper layers with relatively large receptive fields. Some small but moving at high-speed objects are also salient, which are captured from shallower layers holding more low-level information. Although the use of multilevel features such as FPN has already shined in the field of 2D object detection, there are currently few methods to fully verify that multilevel features are effective for video saliency [47]. Jain et al. proposed ViNet [34], which proves that multilevel features are effective for video saliency and achieve excellent performance. However, there is still room for research on how to better use and combine multilevel features and build a fully convolutional model to maximize the accuracy of the model.

To solve these problems, we propose a new 3D fully convolutional encoder-decoder architecture for video saliency detection. The generated saliency maps of video frames by the proposed method are shown in Fig. 1.

In the "Related Works" section, we summarize the related works of video saliency detection. In the "The Proposed Novel TSFP-Net" section, we present the proposed novel TSFP-Net. In the "Experimental Results" section, the experimental results are given. In the "Conclusion" section, the conclusion is summarized.

Related Works

Video saliency detection consists of multiple directions, which can mainly be divided into two categories: fixation prediction and salient object detection. Fixation prediction aims to model the probability that the human eye pays attention to each pixel while watching video images. The preparation of such a dataset usually needs to recruit many volunteers, and an eye tracker is used to freely record the gaze position of each volunteer when they watch videos. Salient object detection aims to segment the accurate contours of the objects of interest of the human eyes in the video images, and the dataset shall be manually marked to obtain the accurate segmentation edges of the salient objects. We focus on fixation prediction in this paper.

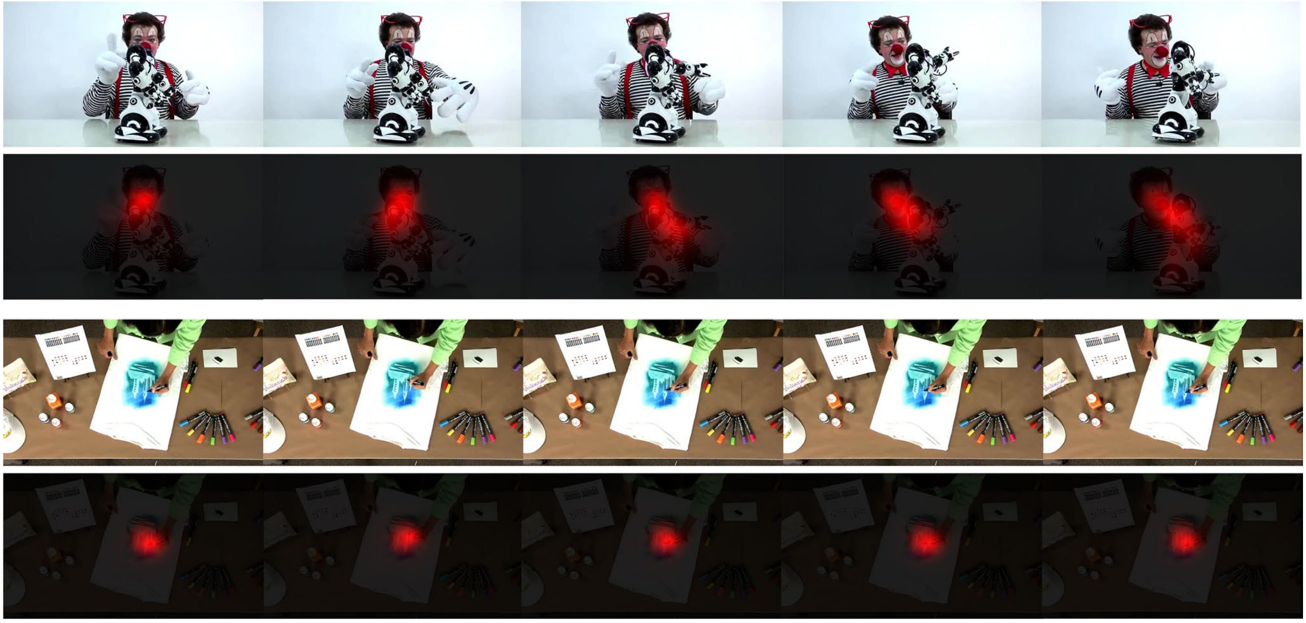

Fig. 1 Visualization of video saliency results of two different videos (interval of 30 frames)

The Latest 2D Video Saliency Detection Networks

In the past, most video saliency detection methods predicted the saliency map by adding a temporal recurrence module to the static network. Jiang et al. proposed DeepVS [22], which establishes a subnetwork of objects through YOLO [23], builds up a subnetwork of motion through FlowNet [24], and then conveys the obtained spatial–temporal features to the double-layer ConvLSTM for prediction. Wang et al. proposed ACLNet [6], which adopts an attention module and a ConvLSTM module to construct the network, among which the attention module is trained on the large static saliency dataset SALICON [25] and the ConvLSTM module is trained on the video saliency dataset. The final model is obtained through the alternating training of static and dynamic saliency. Linardos et al. proposed SalEMA [8], which discusses the performance of the exponential moving average (EMA) and ConvLSTM for video saliency modeling and discovers that the former can acquire a close or even better effect than ConvLSTM.

Lai et al. proposed STRA-Net [13], which proposes a kind of two-stream model in which the motion flow and appearance can couple through dense residual cross-connections at various layers; meanwhile, multiple local attentions can be utilized to enhance the integration of the temporal-spatial features and then conduct the final prediction of the saliency map through ConvGRU and global attention. Wu et al. proposed SalSAC [9], which improves the robustness of the network through a shuffled attention module, and the correlation-based ConvLSTM is employed to balance the change in static image features for the previous frame and current frame. Chen et al. proposed ESAN-VSP [26], which adopts a multiscale deformable convolutional alignment network (MDAN) to align the features of adjacent frames and then predicts the video motion information through Bi-ConvLSTM. Droste et al. proposed UNISAL [27], which is a unified image and video saliency detection model that can extract static features through MobileNet v2 [28], and determined whether to predict temporal information through the Con-vGRU connected by the residual of the controllable switch. In addition, it also adopts domain adaptation technology to realize the high-precision saliency detection of various video datasets and image datasets. Bellitto et al. proposed a deep learning network architecture for video saliency via spatiotemporal reasoning that consists of three parts: a high-level representation module [29], an attention module, and a memory and reasoning module. Recently, Zheng et al. proposed progressive real-time video salient object detection via cascaded fully convolutional networks with motion attention [30].

The Latest 3D Video Saliency Detection Networks

Bazzani et al. proposed RMDN [31], which utilizes C3D [32] to extract the temporal-spatial features and then aggregates time information through LSTM. Min et al. proposed TASED-Net [10], which adopts an S3D network [33] as an encoder, and the decoder uses 3D deconvolution and unpooling to continuously enlarge the image to obtain the saliency map. The unpooling layer adopts auxiliary pooling to fill the feature acquired from the decoder to the activated position corresponding to the maxpooling layer of the encoder. Bellitto et al. proposed HD^{2}S [12], which delivers the multiscale feature output by a 3D encoder to a conspicuity net for decoding separately and then combines all the decoded feature maps to obtain the final saliency map.

Jain et al. proposed ViNet [34], which adopts a 3D encoder-decoder structure in a 2D U-Net-like fashion so that the decoding features of various layers can be constantly concatenated with the corresponding feature of the encoder in the temporal dimension. Then, the video saliency detection results can be obtained through continuous 3D convolution and trilinear upsampling.

Audio–Video Saliency Prediction

Some recent studies have begun to explore the impact of the combination of vision and hearing on saliency. Aytar et al. proposed SoundNet [35], which uses a large amount of unlabeled sound data and video data and uses a pretrained visual model for self-supervised learning to obtain an acoustic representation. Tsiami et al. proposed STAVIS [11], which performs spatial sound source localization through SoundNet combined with visual features in SUSiNet [36] and concatenates the feature maps obtained through sound source localization and visual output feature maps to merge and output the saliency map. Jain et al. proposed ViNet [34], which uses three different methods to fuse the advanced features of the SoundNet output with the deepest features of the ViNet encoder and then performs audio–video saliency prediction.

Chen et al. proposed a multisensory framework of audio and visual signals for video saliency prediction. It mainly includes four modules: auditory feature extraction, visual feature extraction, semantic interaction between auditory features and visual features, and feature fusion [37].

The Proposed Novel TSFP-Net

We fully consider the influence of time, space, and scale and establish a temporal-spatial feature pyramid. Meanwhile, the temporal-spatial semantic features of the deep layer are aggregated to each layer of the pyramid. In view of the different receptive fields of the temporal dimension for the features of various layers, we separately perform independent hierarchical decoding on different levels of the feature pyramid to fully consider the effect of temporal-spatial saliency features with various scales. Since the unpooling layer is bound together with the max-pooling layer, the decoder network cannot be designed freely. Referring to recent studies on the semantic segmentation of 2D networks, convolution with upsampling in decoders [14–18] can obtain better results than the previous method, which adopted deconvolution or unpooling [19–21]. We remove the previous deconvolution and unpooling operation of the 3D fully convolutional encoder-decoder [10] and completely adopt 3D convolution and trilinear upsampling.

We design a 3D fully convolutional encoder-decoder architecture for video saliency detection since the huge defect existed in the model designed in the 2D network described in the preceding part of the paper. Different from the abovementioned 3D network, our network completely utilizes the 3D convolutional layer and trilinear upsampling layer. Our network is the first to build a temporal-spatial feature pyramid in the field of video saliency and aggregate deep semantic features in each layer of feature maps in the feature pyramid. Through the hierarchical decoding of temporal-spatial features at different scales, we obtain the detection results of video saliency that are significantly superior to existing networks.

Temporal-Spatial Feature Pyramid Network

The overall architecture of the proposed temporal-spatial feature pyramid network is shown in Fig. 2.

The main steps of the proposed TSFP-Net are as follows:

Algorithm: The proposed method of TSFP-Net.

Input: T frames at one time of a video.

Step 1: The S3D encoder performs temporal-spatial feature aggregation to obtain the temporal-spatial features of different scales.

Step 2: The top-down path module integrates deep temporal-spatial semantic features with shallow feature of different scales to establish the temporal-spatial feature pyramid.

Step 3: The temporal-spatial features with multiscale semantic information are decoded hierarchically.

Output: A saliency map of the last frame of a T frame video clip is generated.

The implementation details of TSFP-Net are as follows:

Step 1. For TSFP-Net, since the saliency of any frame is determined by several frames in the past, the network inputs T frames at one time and finally outputs a saliency map of the last frame of a T frame video clip. Given the input video clip $\left\{I_{t-T+1}, ..., I_t\right\}$, the S3D encoder performs temporal-spatial feature aggregation through 3D convolution and maxpooling to obtain the temporal-spatial features of different scales [31].

Step 2. The top-down path enhancement integrates deep temporal-spatial semantic features into shallow feature maps of different scales to establish the temporal-spatial feature pyramid. Then, we provide the specific structure of TSFP-Net. In addition, the S3D backbone includes the neck of building the temporal-spatial feature pyramid and the hierarchical convolutional decoder. The overall architecture of the temporal-spatial feature pyramid is shown in Fig. 3, where UP(tri) refers to trilinear upsampling, and the thickness of the cube refers to the channel dimension.

The shallow features have smaller receptive fields, which are utilized to detect small salient objects. The deep layer features have larger receptive fields, which are utilized to detect large salient objects. As a result, the features of different levels are continuously decoded and upsampled to obtain features with the same temporal-spatial and channel dimensions. These features are summed element by element, and the time and channel dimensions are reduced through the 3D convolution of the output layer. The saliency map S_t at time t is obtained through the sigmoid activation function.

In this way, in the form of a sliding window, each time we insert a new frame and delete the first frame, leaving the length of the video clip in the window as T. We can perform frame-by-frame video saliency detection; by doing so, all saliency results of the T frames and subsequent frames of each video can be detected. For the first $T-1$ frames, we can obtain the saliency maps by roughly reversely playing the video frame of the first $2\,T-1$ frames and putting them into the sliding window.

Only the deep layer of the multiscale temporal-spatial features output by the S3D encoder contains advanced semantic features that can be utilized for video saliency detection. Consequently, we add top-down path enhancement to continuously integrate deep high-level semantic features into shallow feature maps. The feature dimensions output by the S3D encoder are, $192 \times \frac{T}{2} \times \frac{H}{4} \times \frac{W}{4}$, $480 \times \frac{T}{2} \times \frac{H}{8} \times \frac{W}{8}$, $832 \times \frac{T}{4} \times \frac{H}{16} \times \frac{W}{16}$, and $1024 \times \frac{T}{8} \times \frac{H}{32} \times \frac{W}{32}$; H and W represent the height and width of the convolution kernel. First, we compress the channel dimensions of the 4 temporal-spatial features to 192 through the $1 \times 1 \times 1$ convolutional layer.

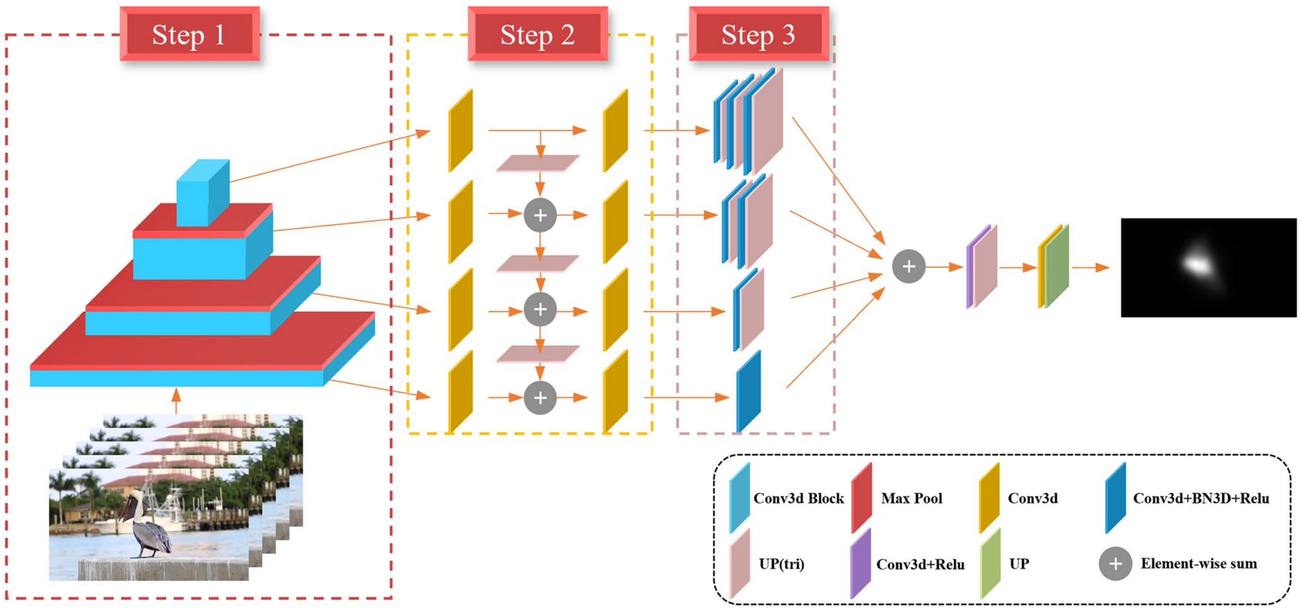

Fig. 2 The overall architecture of the TSFP-Net. (Notes: UP(tri) means trilinear upsampling, UP means bilinear upsampling)

Second, through trilinear upsampling, the deep features are continuously integrated into the shallow features. Third, the output layer adopts a 3 × 3 × 3 convolution to output the multiscale temporal-spatial features integrated with semantic information.

Since the module only integrates the semantic information at the deep layer to the shallow layer, we do not use any activation function and normalized layer.

Step 3. The temporal-spatial features with multiscale semantic information are decoded hierarchically. The structure of the hierarchical convolutional decoder is displayed in Fig. 4.

The temporal-spatial features of different scales all contain semantic information, and the receptive fields of different features are different. Hence, there is no need to interact with each other between features of different scales, and the features of each level can be decoded independently. Finally, the saliency detection results of different receptive fields can be integrated. The decoder at each layer adopts the combination of 3D convolution, 3D batch normalization, and trilinear upsampling for model structure design. To reduce computational complexity, the first 3D convolution of each layer compresses the channel dimension to 96. To finally merge the decoding features of different levels, the final feature

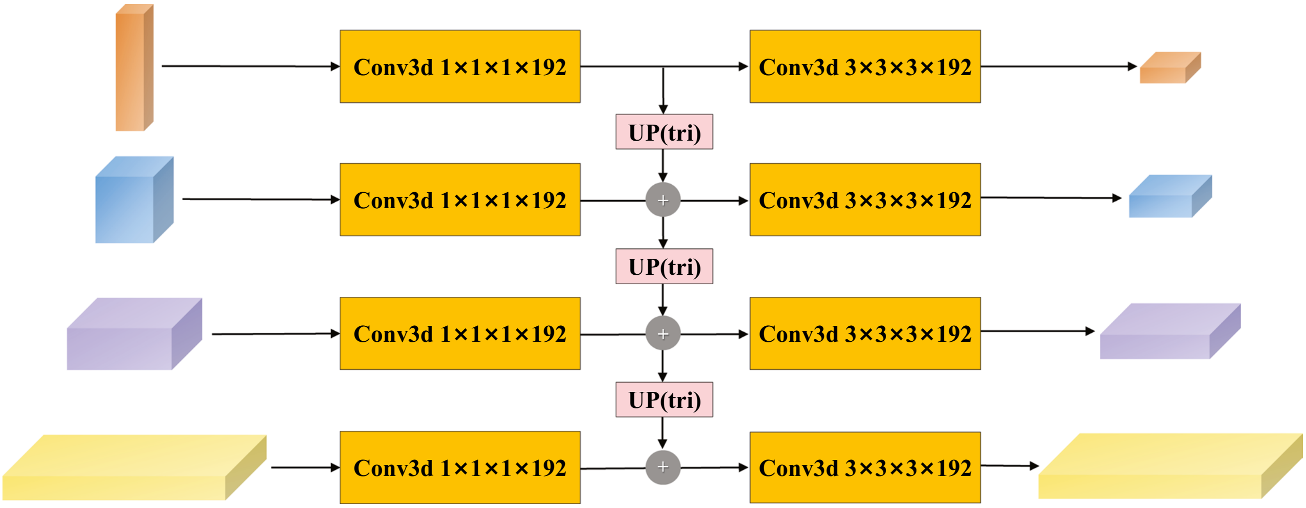

Fig. 3 Building module of the temporal-spatial feature pyramid

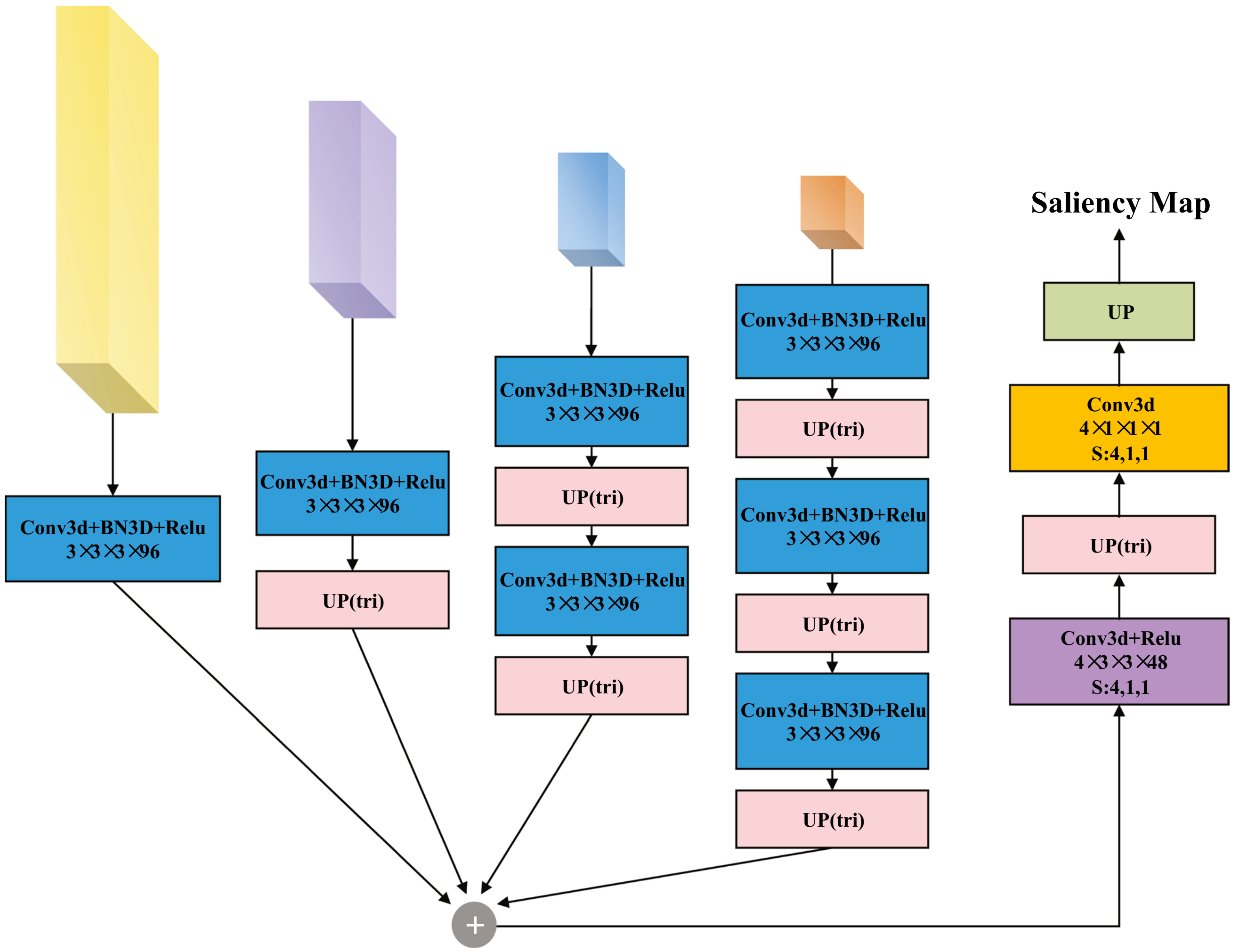

Fig. 4 The structure of the hierarchical convolutional decoder (Note: UP refers to bilinear upsampling, UP(tri) refers to trilinear upsampling)

dimensions of the output of different decoders should be exactly the same.

Therefore, when setting the last trilinear upsampling layer of all levels, we only expand the width and height by 2 times. The time dimension remains unchanged, but the other trilinear upsampling layers simultaneously expand the width, height, and time dimensions by 2 times. In this way, the dimensions of the feature maps output by the four decoders are obtained as $96 \times \frac{T}{2} \times \frac{H}{4} \times \frac{W}{4}$. Then, the final saliency map can be obtained through two 3D convolutional layers, two upsampling layers, and a final sigmoid activation function.

Loss Function

The training of the video saliency network is a regression problem that aims to make the distribution of the output saliency map consistent with the ground truth. In the past, a large number of video saliency models adopted Kullback–Leibler (KL) divergence as a loss function to train the model and achieved good results [49]. However, there are multiple metrics that evaluate the saliency from different aspects; among them, the linear correlation coefficient (CC) and the normalized scanpath saliency (NSS) seem to be more reliable for evaluating the quality of the saliency map [49]. We take the weighted summation of the above KL, CC, and NSS to represent the final loss function, and the subsequent ablation studies prove that the weighted summation of the three losses achieves better results than just using the KL loss.

Assuming that the predicted saliency map is S∈[0,1], the labeled binary fixation map is F∈{0,1}, and the ground truth saliency map generated by the fixation map is G∈[0,1], the final loss function can be expressed as

$$L(S,F,G) = L_{KL}(S,G) + \alpha_1 L_{CC}(S,G) + \alpha_2 L_{NSS}(S,F) \qquad (1)$$

We set $\alpha_1 = 0.5$ and $\alpha_2 = 0.1$ according to the value range of each item. L_{KL}, L_{CC}, and L_{NSS} signify the loss of Kullback–Leibler

(KL) divergence, the linear correlation coefficient (CC), and the normalized scanpath saliency (NSS), respectively. Their calculation formulas are as follows:

$$L_{KL}(S,G) = \sum\nolimits_x G(x) \ln \frac{G(x)}{S(x)} \tag{2}$$

$$L_{CC}(S,G) = -\frac{\mathrm{cov}(S,G)}{\sigma(S)\sigma(G)} \tag{3}$$

$$L_{NSS}(S,F) = -\frac{1}{N} \sum\nolimits_x s(x)F(x), \left(s(x) = \frac{S(x) - \mu(S(x))}{\sigma(S(x))} \right) \tag{4}$$

where $\sum_x(\cdot)$ represents summing all the pixels, cov(·) represents the covariance, $\mu(\cdot)$ represents the mean, and $\rho(\cdot)$ represents the variance.

Experimental Results

Datasets

Similar to most video saliency studies, we evaluate our method on the three most commonly used video saliency datasets, which are DHF1K [6], Hollywood-2 [38], and UCFsports [38]. At the same time, we evaluate our model on six audio–video saliency datasets: DIEM [39], Coutrot1 [40, 41], Coutrot2 [40, 41], AVAD [42], ETMD [43], and SumMe [44].

The DDF1K dataset contains 1000 videos collected from spanning a large range of scenes, motions, object types, and complex backgrounds and is the largest and most diverse video saliency dataset to date. It consists of 600 videos for training, 100 videos for validation, and 300 videos for testing. For a fair comparison, the first 700 videos publicly provide ground truth for training and validation, while the remaining 300 videos do not provide ground truth; therefore, the experimental results shall be submitted to the evaluation server for blind assessment, which is different from the other datasets. Since the variety of this dataset is the most complicated, we conduct our experiments and ablation studies mainly based on it.

The Hollywood-2 dataset contains 1707 videos, which can be divided into 6659 short video clips for training and testing; among them, the training set consists of 3100 clips, and the test set consists of 3559 clips. The dataset is a task-driven video saliency dataset, mainly focusing on human actions in movie scenes. The UCF-sports dataset contains 150 video clips taken from the UCF Sport Action Dataset [45], mainly emphasizing human actions in sports. It is divided into 103 video clips for training and 47 video clips for testing. DIEM consists of 81 movie clips of varying genres. They are sourced from publicly accessible repositories and consist of 64 training videos and 17 test videos. Coutrot datasets are split into Coutrot1 and Coutrot2. Coutrot1 contains 60 clips with dynamic natural scenes split into 4 visual categories. Coutrot2 contains 15 clips of 4 persons in a meeting and the corresponding eye-tracking data from 40 persons. The AVAD dataset contains 45 short clips of 5–10-s duration with several audio-visual scenes. The ETMD dataset contains 12 videos from six different Hollywood movies. The SumMe dataset contains 25 unstructured videos, which are acquired in a controlled psychological experiment.

Experimental Setup

To train TSFP-Net, we first initialize our encoder using the S3D model pretrained on Kinetics. In the DHF1K dataset, we adopt the standard division of the training set and validation set to train our model. T continuous video frames are randomly selected from each video each time, each frame is resized to 192 × 352, and the batch size is set to 16 videos during the training. Restricted by the memory, we can only deal with 4 videos each time, so we accumulate the gradient and update the model parameters every other 4 steps. We use the Adam optimizer [48], the initial learning rate is set to 0.0001, and the learning rate is reduced by 10 times at the 22nd, 25th, and 26th epochs. We train 26 epochs in total and use early stopping in the DHF1K validation set to save the model parameters corresponding to the largest NSS result on the validation set. Due to the excessive number of images in the validation set, we only use the first 80 frames of each video for validation during the training process.

For the Hollywood-2 and UCF-sports datasets, we use the models trained on DHF1K to fine-tune the models separately. Since these two datasets contain a large number of video clips that are less than T, for all video clips less than T in the training set, we first repeat the first frame $T-1$ times in front, and we adopt early stopping on the test set of these two datasets.

For six audio–video saliency datasets, we use the model pretrained in DHF1K to initialize the model and fine-tune it on six audio–video saliency datasets without audio. The three different splits used in the datasets are the same as in [11], and we evaluate the average metrics of different splits.

We use the most commonly used evaluation metrics in the DHF1K benchmark to evaluate our model for the DHF1K dataset. These include (i) normalized scanpath saliency (NSS); (ii) linear correlation coefficient (CC); (iii) similarity (SIM); (iv) area under the curve by Judd (AUC-J); and (v) shuffled AUC (s-AUC) [50]. For all these metrics, the larger the value is, the better. For other datasets and ablation studies, we use AUC-J, SIM, CC, and NSS metrics.

The definitions of NSS, CC, SIM, AUC-J, and s-AUC are as follows [49]:

Table 1 The experimental results of the DHF1K validation set while training at different clip lengths (*T*) (The best scores are shown in red)

Clip length (T)	AUC-J	SIM	CC	NSS
16	0.916	0.392	0.500	2.876
32	0.919	0.397	0.529	3.009
48	0.917	0.398	0.526	2.990

$$NSS(P,R) = \frac{1}{N}\sum_i \overline{P}_i \times R_i,\ N = \sum_i R_i,\ \overline{P} = \frac{P-\mu(P)}{\sigma(P)} \qquad (5)$$

$$CC(P,Q) = \frac{\mathrm{cov}(P,Q)}{\sigma(P)\sigma(Q)} \qquad (6)$$

The similarity metric (SIM) considers the saliency prediction result *P* and the continuous human attention truth distribution *Q* as probability distributions. Then, *P* and *Q* are normalized, and the minimum value on each pixel is calculated and finally added to obtain the SIM.

$$SIM(P,Q) = \sum_i \min\left(P'_i, Q'_i\right),\ \sum_i P'_i = 1,\ \ \sum_i Q'_i = 1 \qquad (7)$$

The AUC is the area under the receiver operating characteristic (ROC) curve. The ROC curve is drawn with the false positive rate (FPR) as the horizontal axis and the true positive rate (TPR) as the vertical axis. The FPR and TPR are calculated as follows:

$$\begin{cases} FPR = \frac{FP}{FP+TN} \\ TPR = \frac{TP}{TP+FN} \end{cases} \qquad (8)$$

In the calculation of the area under the curve by Judd (AUC-J), the true positive probability is the pixel ratio predicted accurately on all true value concerns, and the false positive probability is the pixel ratio predicted as significant on nonconcerns.

The shuffled AUC (s-AUC) reduces the sensitivity of the original AUC index to the center offset. When sampling nonsignificant points, the s-AUC index takes samples from the distribution of concerns on multiple other images instead of randomly sampling nonsignificant points on the original image.

Evaluation on DHF1K

The DHF1K dataset is currently the largest and most diverse video saliency dataset; thus, DHF1K is adopted as the preferred dataset for ablation study and evaluation of the test set. We change the length of *T* to 16, 32, and 48 to train our model and observe the results on the DHF1K validation set. The experimental results are shown in Table 1. We discover that when *T* is 32, the performance is the best because it obtains the highest AUC-J, CC, and NSS.

We discover that our model is significantly better than other state-of-the-art methods, especially NSS, CC, and AUC-J, which make remarkable gains. Although s-AUC and SIM fail to rank first in Table 2, they make up the top three. In particular, according to [46], the AUC is more suitable to evaluate the performance of the video saliency model. SIM penalizes models with false negatives significantly more than false positives; in terms of evaluation, it is inferior to NSS and CC, which treat false positives and false negatives symmetrically. Consequently, NSS and CC are believed to be most related to the human eye's visual attention and are recommended to evaluate the saliency model [46]. Compared with other methods, we make a huge breakthrough in terms of NSS and CC.

Table 2 Comparison of the saliency metrics on the DHF1K test set for TSFP-Net and other state-of-the-art methods (The best scores are shown in red, and the second-best scores are shown in blue)

Metrics Method	NSS	CC	SIM	AUC-J	s-AUC
DeepVS [22]	1.911	0.344	0.256	0.856	0.583
ACLNet [6]	2.354	0.434	0.315	0.890	0.601
SalEMA [8]	2.574	0.449	0.466	0.890	0.667
STRA-Net [13]	2.558	0.458	0.355	0.895	0.663
TASED-Net [10]	2.667	0.470	0.361	0.895	0.712
SalSAC [9]	2.673	0.479	0.357	0.896	0.697
UNISAL [27]	2.776	0.490	0.390	0.901	0.691
HD^{2}S [12]	2.781	0.497	0.406	0.901	0.699
ViNet [34]	2.872	0.511	0.381	0.908	0.729
TSFP-Net	2.966	0.517	0.392	0.912	0.723

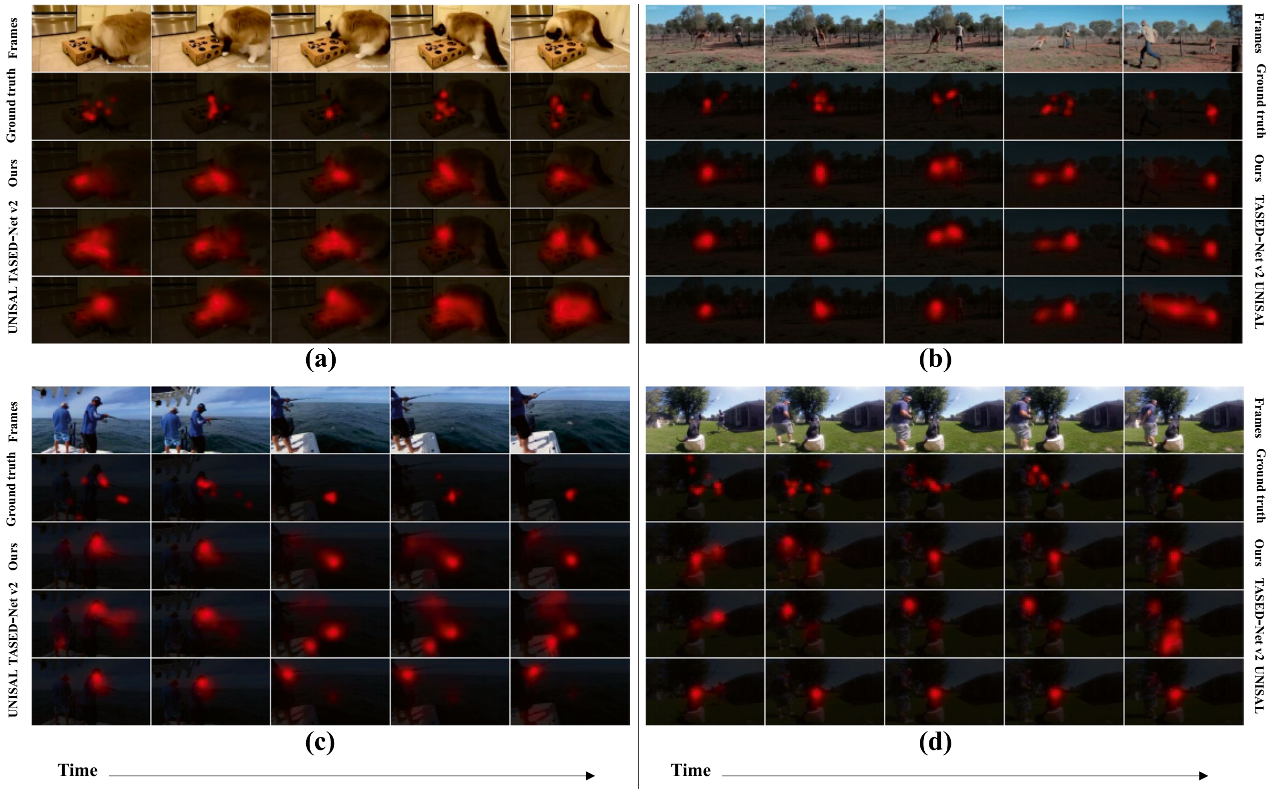

Fig. 5 **a–d** Comparison of the visualization results of saliency maps for TSFP-Net and two other state-of-the-art methods. TSFP-Net is significantly superior to TASED-Net v2 and UNISAL; the generated saliency maps are denser, and there are basically no false detections and missed detections, while the other two methods have obvious false detections and missed detections

Next, we submit the results of our model to the evaluation server of the DHF1K test set. The results for TSFP-Net and all other state-of-the-art methods [6, 8–10, 12, 13, 22, 27, 34] on the DHF1K test set are shown in Table 2.

Meanwhile, as shown in Table 2, the models based on the 3D fully convolutional encoder-decoder are mostly superior to the 2D models based on LSTM [6, 8, 9, 13, 22, 27], which are related to the defects of the 2D network that we analyzed previously and the simultaneous temporal-spatial aggregation of 3D convolution. Our model is currently the most powerful 3D fully convolutional encoder-decoder and video saliency network, which proves the effectiveness of our method.

We also visualize the saliency maps generated through TSFP-Net from the DHF1K validation set and compare it with other state-of-the-art methods, which is shown in Fig. 5. Since TASED-Net has been updated to TASED-Net v2 and the code is open source, the NSS on the DHF1K test set can be up to 2.797. As a result, we compare TSFP-Net with the two most powerful models recently published: TASED-Net v2 and UNISAL. It can be seen that our model has great advantages. First, the saliency maps generated by our method are more concentrated and have a smaller area that is closer to the ground truth, while the saliency maps of the other two methods are more scattered. Second, our method usually does not produce false detections and missed detections, while the other two methods have more obvious false detections and missed detections.

As shown in Fig. 5a, the other two methods produce redundant detections. In Fig. 5c, only our model can accurately detect fishing hooks. TASED-Net v2 produces redundant detections, and UNISAL is completely wrong.

Table 3 Runtime comparison for TSFP-Net and other state-of-the-art methods

Method	Runtime (s)	Model sizes (MB)
DeepVS [22]	0.05	344
ACLNet [6]	0.02	250
SalEMA [8]	0.01	364
STRA-Net [13]	0.02	641
TASED-Net [10]	0.06	82
SalSAC [9]	0.02	93.5
UNISAL [27]	0.009	15.5
HD^{2}S [12]	0.03	116
ViNet [34]	0.016	124
TSFP-Net	0.011	58.4

Table 4 Comparison of saliency metrics for TSFP-Net and other state-of-the-art methods on the Hollywood-2 test set and UCF-sports test set (The best scores are shown in red, and the second-best scores are shown in blue)

Dataset / Method	Hollywood-2				UCF-sports			
	AUC-J	SIM	CC	NSS	AUC-J	SIM	CC	NSS
DeepVS [22]	0.887	0.356	0.446	2.313	0.870	0.321	0.405	2.089
ACLNet [6]	0.913	0.542	0.623	3.086	0.897	0.406	0.510	2.567
SalEMA [8]	0.919	0.487	0.613	3.186	0.906	0.431	0.544	2.638
STRA-Net [13]	0.923	0.536	0.662	3.478	0.910	0.479	0.593	3.018
TASED-Net [10]	0.918	0.507	0.646	3.302	0.899	0.469	0.582	2.920
SalSAC [9]	0.931	0.529	0.670	3.356	0.926	0.534	0.671	3.523
UNISAL [27]	0.934	0.542	0.673	3.901	0.918	0.523	0.644	3.381
HD^{2}S [12]	0.927	0.558	0.668	3.426	0.913	0.493	0.594	3.001
ViNet [34]	0.930	0.550	0.693	3.730	0.924	0.522	0.673	3.620
TSFP-Net	0.936	0.571	0.711	3.910	0.923	0.561	0.685	3.698

We also compare the runtime and the model size of our model with other state-of-the-art methods. We test our model on an Intel Core i7-820QM CPU@3.06 GHz with 64 GB RAM and an NVIDIA RTX 2080Ti GPU, which takes approximately 0.011 s to generate a saliency map. The comparison of running time and model size with other methods is shown in Table 3. As shown in Table 3, TSFP-Net is the second smallest model in all models (UNISAL is the first), while the accuracy of TSFP-Net has huge gains compared to other models.

As seen, not only does the accuracy of our model greatly exceed the state-of-the-art methods, but the speed of generating the saliency map is the third fastest, and the model size is the second smallest but enough to obtain the highest accuracy.

Evaluation on Other Datasets

We also evaluate the performance of our model on Hollywood-2 and UCF-sports. We observe that these two datasets are task-driven video saliency datasets, and there are a large number of video clips with less than 32 frames. Even Hollywood-2 has many video clips with only 1 or 2 frames, and the difference between two adjacent frames of all video clips is very obvious. Reverse playback of the video itself can change the video saliency on the large-scale DHF1K dataset. Since the length of video clips is long enough (several hundred frames), the images are extracted according to the appropriate frame rate, and the types of videos are diverse, the impact of the reverse playback can be mitigated to produce normal saliency results in the previous frames of the video.

Table 5 Comparison results on the DIEM, Coutrot 1, and Coutrot 2 test sets (The best scores are shown in red)

Dataset / Method	DIEM				Coutrot1				Coutrot2			
	AUC-J	SIM	CC	NSS	AUC-J	SIM	CC	NSS	AUC-J	SIM	CC	NSS
ACLNet [6]	0.869	0.427	0.522	2.02	0.850	0.361	0.425	1.92	0.926	0.322	0.448	3.16
TASED-Net [10]	0.881	0.461	0.557	2.16	0.867	0.388	0.479	2.18	0.921	0.314	0.437	3.17
STAVIS [11]	0.883	0.482	0.579	2.26	0.868	0.393	0.472	2.11	0.958	0.511	0.734	5.28
ViNet [34]	0.898	0.483	0.626	2.47	0.886	0.423	0.551	2.68	0.950	0.466	0.724	5.61
AViNet(B) [34]	0.899	0.498	0.632	2.53	0.889	0.425	0.560	2.73	0.951	0.493	0.754	5.95
TSFP-Net	0.905	0.529	0.649	2.63	0.894	0.451	0.570	2.75	0.957	0.516	0.718	5.30

Table 6 Comparison results on the AVAD, ETMD, and SumMe test sets (The best scores are shown in red)

Method \ Dataset	AVAD				ETMD				SumMe			
	AUC-J	SIM	CC	NSS	AUC-J	SIM	CC	NSS	AUC-J	SIM	CC	NSS
ACLNet [6]	0.905	0.446	0.580	3.17	0.915	0.329	0.477	2.36	0.868	0.296	0.379	1.79
TASED-Net [10]	0.914	0.439	0.601	3.16	0.916	0.366	0.509	2.63	0.884	0.333	0.428	2.10
STAVIS [11]	0.919	0.457	0.608	3.18	0.931	0.425	0.569	2.94	0.888	0.337	0.422	2.04
ViNet [34]	0.928	0.504	0.694	3.82	0.928	0.409	0.569	3.06	0.898	0.345	0.466	2.40
AViNet(B) [34]	0.927	0.491	0.674	3.77	0.928	0.406	0.571	3.08	0.897	0.343	0.463	2.41
TSFP-Net	0.931	0.530	0.688	3.79	0.932	0.433	0.576	3.09	0.894	0.362	0.463	2.28

However, we reveal that in these two datasets, the saliency results of the previous frame obtained from reverse playback are very poor.

First, as a result, we do not adopt reverse playback to predict the saliency of the previous frame during the test set. Second, in terms of the video frames that are less than T, we supplement $T-1$ frames in front and obtain the saliency frame by frame through the order of play. Third, in terms of the clip length that is between T and 2 $T-1$, we repeat the first frame and supplement the video clips to 2 $T-1$. Fourth, we predict the saliency frame by frame after the T frames. Fifth, for the clip length that is greater than or equal to 2 $T-1$, we directly predict the saliency frame by frame of all frames after the T frames.

The comparison results of our method on the Hollywood-2 and UCF-sports test sets obtained in this way and other state-of-the-art methods are shown in Table 4. It can be seen that our model is also highly superior to other methods on these two datasets.

We also evaluated the results of TSFP-Net on six audio–video saliency datasets, and the performance comparisons with other methods are shown in Tables 5 and 6. Although our model does not contain audio, it is much better than all the state-of-the-art methods on most datasets.

Ablation Studies

We first prove that the multiscale temporal-spatial feature pyramid constructed by top-down path enhancement and hierarchical decoding is effective and important for video saliency prediction.

Table 7 Performance comparison for TSFP-Net with different network structures on the validation set of DHF1K

Different architecture	NSS	CC	AUC-J	SIM
TSFP-Net (only final-level)	2.7868	0.5010	0.9121	0.3860
TSFP-Net (only multilevel)	2.8857	0.5097	0.9156	0.3819
TSFP-Net	3.0086	0.5290	0.9188	0.3975

First, we only use the hierarchical decoder and do not build the temporal-spatial feature pyramid. We only change the channel dimensions of the output multiscale temporal-spatial features through a $1 \times 1 \times 1$ convolution to make the feature channels input into the hierarchical decoder consistent. After that, the features directly input the hierarchical decoder and are integrated to obtain the saliency map. This configuration is TSFP-Net (only multilevel). Second, we delete the hierarchical decoder and only adopt the deepest features of the encoder for decoding to obtain saliency; the configuration is TSFP-Net (only final-level). The results on the validation set of DHF1K for different network structures are shown in Table 7.

We observe that the results of hierarchical decoding for different layers are significantly better than those obtained using only the deepest layer's features, and adding top-down path enhancement to construct a semantic temporal-spatial feature pyramid combined with hierarchical decoding has the best effect. Compared to TASED-Net [10], which adopts 3D deconvolution and unpooling, our TSFP-Net (only final-level) only adopts 3D convolution and trilinear upsampling. The NSS result on the validation set of DHF1K is 2.787, which is better than that of TASED-Net, which is 2.706. This indicates that deconvolution and unpooling not only rely too much on the maxpooling layer in the encoder, which leads to the inability to freely design the network structure, but also limit the learning ability of the network to some extent.

We also compare the effects of different loss functions on network performance, and the results are shown in Table 8. We prove that the adoption of the weighted summation of three losses can obtain better performance than using the KL loss alone.

Table 8 Performance comparison for TSFP-Net with different loss functions on the validation set of DHF1K

Different loss	NSS	CC	AUC-J	SIM
TSFP-Net (only KL loss)	2.9876	0.5287	0.9186	0.3927
TSFP-Net	3.0086	0.5290	0.9188	0.3975

Conclusion

Compared to the existing video saliency detection models, we put forward a novel 3D fully convolutional multiscale temporal-spatial feature pyramid network of TSFP-Net consisting of 3D convolution and trilinear upsampling, which is the first to build a temporal-spatial feature pyramid and aggregate deep semantic features in each layer of feature maps in the feature pyramid.

The main contributions of the paper are as follows: First, we develop a new 3D fully convolutional temporal-spatial feature pyramid network called TSFP-Net, which completely consists of 3D convolution and trilinear upsampling and obtains very high accuracy in the case of a small model size. Second, we construct a feature pyramid of different scales containing rich temporal-spatial semantic features and build a hierarchical 3D convolutional decoder for decoding. We prove that such an approach can significantly improve the detection performance of video saliency. Third, we evaluate our model on three purely visual large-scale video saliency datasets. Compared with the state-of-the-art methods, our model can achieve large gains.

We test our model on an Intel Core i7-820QM CPU@3.06 GHz with 64 GB RAM and an NVIDIA RTX 2080Ti GPU for the comparison of TSFP-Net and other state-of-the-art methods on three purely visual video saliency benchmarks to prove the effectiveness of our method.

The experimental results show that the proposed model has the second smallest size and much higher prediction precision, and the running time is real-time and third fastest. The proposed video saliency detection model is obviously different and significantly superior to all state-of-the-art methods.

The fusion mechanism of the video and audio information should be further researched to continually improve the video saliency prediction precision. Video saliency detection for 4 K or 8 K video should also be researched to reveal the saliency information in ultrahigh resolution video.

In the next step, the vision transformer-based architecture will be incorporated into the video saliency prediction field. We will also extend the proposed human vision attention mechanism-inspired temporal-spatial feature pyramid for video saliency detection to video saliency forecasting by forecasting the saliency of future frames.

Funding This study was funded by the National Natural Science Foundation of China (NSFC) under grants No. 61375025, 61075011, and 60675018 and the Scientific Research Foundation for the Returned Overseas Chinese Scholars from the State Education Ministry of China.

Data Availability The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest The authors declare no competing interests.

References

1. Hadizadeh H, Bajic´ IV. Saliency-aware video compression. IEEE Trans Image Process. 2013;23(1):19–33.
2. Zhu S, Liu C, Xu Z. 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. 2019;30(7):1946–59.
3. Guraya FFE, Cheikh FA, Tremeau A, Tong Y, Konik H. Predictive saliency maps for surveillance videos. Ninth Int Symp Distrib Comput App to Bus Engr Sci IEEE. 2010;2010:508–13.
4. Lyu C, Liu Y, Wang X, Chen Y, Jin J, Yang J. Visual early leakage detection for industrial surveillance environments. IEEE Trans Industr Inf. 2022;18(6):3670–80.
5. Nguyen TV, Xu M, Gao G, Kankanhalli M, Tian Q, Yan S. Static saliency vs. dynamic saliency: a comparative study. Proc of the 21st ACM Int Conf on Multimed. 2013:987–996.
6. Wang W, Shen J, Guo F, Cheng MM, Borji A. Revisiting video saliency: a large-scale benchmark and a new model. Proc IEEE Conf Comput Vis Pattern Recognit. 2018:4894–4903.
7. Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-C. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. arXiv preprint. arXiv:1506.04214, 2015.
8. Linardos P, Mohedano E, Nieto JJ, O'Connor NE, Giro-i-Nieto X, McGuinness K. Simple vs complex temporal recurrences for video saliency prediction. arXiv preprint.arXiv:1907.01869, 2019.
9. Wu X, Wu Z, Zhang J, Ju L, Wang S. Salsac: a video saliency prediction model with shuffled attentions and correlation-based convlstm. Proc AAAI Conf Artif Intel. 2020;34(07):12410–7.
10. Min K, Corso JJ. Tased-net: temporally aggregating spatial encoder-decoder network for video saliency detection. Proc IEEE/CVF Int Conf Comput Vis. 2019:2394–2403.
11. Tsiami A, Koutras P, Maragos P. Stavis: spatiotemporal audiovisual saliency network. Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2020:4766–4776.
12. Bellitto G, Salanitri FP, Palazzo S, Rundo F, Giordano D, Spampinato C. Hierarchical domain-adapted feature learning for video saliency prediction. arXiv preprint. arXiv:2010.01220v4, 2021.
13. Lai Q, Wang W, Sun H, Shen J. Video saliency prediction using spatiotemporal residual attentive networks. IEEE Trans Image Process. 2019;29:1113–26.
14. Chen L-C, Papandreou G, Kokki I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint. arXiv:1412.7062, 2014.
15. Zhu L, Ji D, Zhu S, Gan W, Wu W, Yan J. Learning statistical texture for semantic segmentation. Proc IEEE Conf Comput Vis Pattern Recognit. 2021:12532–12541.
16. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv preprint. arXiv:1706.05587, 2017.
17. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proc European Conf Comput Vis (ECCV). 2018:801–818.
18. Lin G, Milan A, Shen C, Reid I. Refinenet: multipath refinement networks for high-resolution semantic segmentation. Proc IEEE Conf Comput Vis Pattern Recognit. 2017:1925–1934.

19. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(12):2481–95.
20. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proc IEEE Conf Comput Vis Pattern Recognit 2015:3431–3440.
21. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Int Conf Med Image Comput Computer-assisted Intervention Springer. 2015:234–241.
22. Jiang L, Xu M, Liu T, Qiao M, Wang Z. Deepvs: a deep learning based video saliency prediction approach. Proc European Conf Comput Vis (ECCV). 2018:602–617.
23. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. Proc IEEE Conf Comput Vis Pattern Recognit. 2016:779–788.
24. Dosovitskiy A, Fischer P, Ilg E, Hausser P, Hazirbas C, Golkov V, Van Der Smagt P, Cremers D, Brox T. Flownet: learning optical flow with convolutional networks. Proc IEEE Int Conf Comput Vis. 2015:2758–2766.
25. Huang X, Shen C, Boix X, Zhao Q. Salicon: reducing the semantic gap in saliency prediction by adapting deep neural networks. Proc IEEE Int Conf Comput Vis. 2015:262–270.
26. Chen J, Song H, Zhang K, Liu B, Liu Q. Video saliency prediction using enhanced spatiotemporal alignment network. Pattern Recogn. 2021;107615:1–12.
27. Droste R, Jiao J, Noble JA. Unified image and video saliency modeling. European Conf Comput Vis Springer. 2020:419–435.
28. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. Mobilenetv2: inverted residuals and linear bottlenecks. Proc IEEE Conf Comput Vis Pattern Recognit. 2018:4510–4520.
29. Bellitto G, Proietto Salanitri F, Palazzo S, Rundo F, Giordano D, Spampinato C. Hierarchical domain-adapted feature learning for video saliency prediction. Int J Comput Vis 2021;129:3216–3232.
30. Zheng Q, Li Y, Zheng L, Shen Q. Progressively real-time video salient object detection via cascaded fully convolutional networks with motion attention. Neurocomputing. 2022;467:465–75.
31. Bazzani L, Larochelle H, Torresani L. Recurrent mixture density network for spatiotemporal visual attention. arXiv preprint arXiv:1603.08199, 2016.
32. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. Proc IEEE Int Conf Comput Vis. 2015:4489–4497.
33. Xie S, Sun C, Huang J, Tu Z, Murphy K. Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification. Proc European Conf Comput Vis (ECCV). 2018:305–321.
34. Jain S, Yarlagadda P, JyotiS, Karthik S, Subramanian R, Gandhi V. Vinet: pushing the limits of visual modality for audio-visual saliency prediction. arXiv preprint. arXiv:2012.06170v2, 2021.
35. Aytar Y, Vondrick C, Torralba A. Soundnet: learning sound representations from unlabeled video. arXiv preprint. arXiv:1610.09001, 2016.
36. Koutras P, Maragos P. Susinet: see, understand and summarize it. Proc IEEE/CVF Conf Comput Vis Pattern Recognit Workshops. 2019:809–819.
37. Chen J, Li Q, Ling H, Ren D, Duan P. Audiovisual saliency prediction via deep learning. Neurocomputing. 2021;428:248–58.
38. Mathe S, Sminchisescu C. Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2014;37(7):1408–24.
39. Mital PK, Smith TJ, Hill RL, Henderson JM. Clustering of gaze during dynamic scene viewing is predicted by motion. Cogn Comput. 2011;3(1):5–24.
40. Coutrot A, Guyader N. How saliency, faces, and sound influence gaze in dynamic social scenes. J Vis. 2014;14(8):5–5.
41. Coutrot A, Guyader N. Multimodal saliency models for videos. From Human Attention to Computational Attention Springer. 2016:291–304.
42. Min X, Zhai G, Gu K, Yang X. Fixation prediction through multimodal analysis. ACM Trans Multimed Comput Commun Appl (TOMM). 2016;13(1):1–23.
43. Koutras P, Maragos P. A perceptually based spatiotemporal computational framework for visual saliency estimation. Signal Process: Image Commun. 2015;38:15–31.
44. Gygli M, Grabner H, Riemenschneider H, Van Gool L. Creating summaries from user videos. European Conf Comput Vis (ECCV) Springer. 2014:505–520.
45. Rodriguez MD, Ahmed J, Shah M. Action mach a spatiotemporal maximum average correlation height filter for action recognition. IEEE Conf Comput Vis Pattern Recognit. 2008;2008:1–8.
46. Bylinskii Z, Judd T, Oliva A, Torralba A, Durand F. What do different evaluation metrics tell us about saliency models? IEEE Trans Pattern Anal Mach Intell. 2018;41(3):740–57.
47. Lin T-Y, Dollar P, Girshick R, He KM, Hariharan B, Belongie S. Feature pyramid networks for object detection. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR). 2017:2117–2125.
48. Kingma DP, Ba J. Adam: a method for stochastic optimization. 3rd Int Conf Learning Rep San Diego. 2015:1–15.
49. Riche N, Duvinage M, Mancas M, Gosselin B, Dutoit T. Saliency and human fixations: state-of-the-art and study of comparison metrics. Proc IEEE Conf Comput Vis. 2013:1153−1160.
50. Borji A, Tavakoli HR, Sihite DN, Itti L. Analysis of scores, datasets, and models in visual saliency prediction. Proc IEEE Conf Comput Vis. 2013:921−928.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.