IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

OFF-ViNet: Optical Flow-Based Feature Warping ViNet for Video Saliency Prediction Considering Future Prediction

REITA IKENOYA[1], TOMONORI TASHIRO[1], and GOSUKE OHASHI[1], (Member, IEEE)
[1]Department of Electrical and Electronic Engineering, Shizuoka University, Hamamatsu, Shizuoka 432-8561, Japan

Corresponding author: Reita Ikenoya (e-mail: ikenoya.reita.18@shizuoka.ac.jp).

This work was supported by JSPS KAKENHI Number 23K11117.

ABSTRACT Active studies have been conducted on video saliency prediction, which predicts human visual attention toward videos. Most deep learning-based video saliency prediction models implicitly learn features that contribute to video saliency prediction, greatly improving accuracy. This study proposes a model called optical flow-based feature warping ViNet (OFF-ViNet). This model explicitly adds a Warping module, which is a mechanism that considers future predictions based on object motion in addition to implicitly learned features. The Warping module spatially warps the hierarchical features extracted by the 3D convolutional backbone based on the optical flow to obtain a feature representation that predicts the future. Compared with exisiting models, OFF-ViNet achieves better and competitive accuracy with state-of-the-art models on video saliency prediction datasets, particularly on UCF-Sports, which contains several videos with moving objects.

INDEX TERMS Video saliency prediction, Optical flow, 3D convolutional neural network

I. INTRODUCTION

HUMAN visual attention selectively processes information in regions of high visual importance. Consequently, humans process a vast amount of complex information through vision and quickly recognize the external world. Therefore, certain characteristic areas in the human visual field tend to be focused on. Visual saliency is defined as the degree to which gaze tends to be concentrated, and a saliency map is an image that emphasizes visual saliency. Fig. 1 shows the video frames and overlay saliency maps corresponding to the video frames.

Since the model was proposed by Itti et al. in 1998 [1], various methods [2]–[12] have been proposed for saliency prediction, which predicts the visual saliency of an image. Saliency prediction includes image and video saliency predictions. Video saliency prediction has been applied in various areas of computer vision, such as robot camera control [13], video subtitling [14], video compression [15], [16], and video segmentation [17], [18].

Much research has been conducted on deep learning-based saliency prediction [19]–[52]. Spatial features such as color, brightness, orientation, and object-like features are important for image saliency prediction [53], [54]. Video saliency prediction is based on temporal features, such as motion, in addition to the aforementioned spatial features [53]. However, cases in which the gaze is directed ahead of the moving target rather than on the target itself have been reported [55]. This is achieved by predicting future motions based on past object motions. Deep learning-based video saliency prediction models train models to extract spatiotemporal features from large datasets. The authors confirmed that the observer's prediction of the target's future motion increases the saliency that appears ahead of the target, which existing video saliency prediction models cannot accurately predict [56]. This suggests that in video saliency prediction, an effective way to properly predict the saliency that appears in the targets is

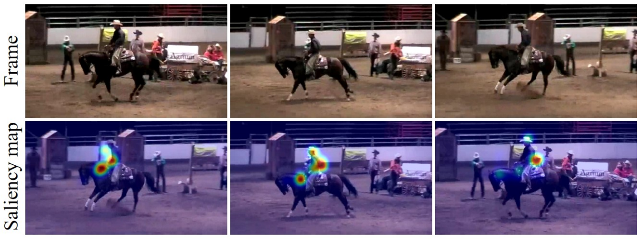

FIGURE 1. Examples of video frames and saliency maps. Saliency maps are overlaid on the frame, with red and blue regions being more salient and less salient, respectively.

to predict the current location of the target based on its past location and movement.

This study proposes an optical flow-based feature warping ViNet (OFF-Net), a video saliency prediction model that explicitly considers future predictions based on optical flow. A Warping module is introduced into the 3D convolution-based model that achieved high accuracy with existing video sailency prediction models.

The main contributions of this study are as follows:

1) An optical flow-based feature warping ViNet (OFF-ViNet), which adds a new module, namely, Warping module to the video saliency prediction model as a mechanism to obtain feature representations that predict the future. OFF-ViNet explicitly predicts the future using an optical flow.
2) OFF-ViNet achieves a competitive performance with the state-of-the-art models on multiple benchmark datasets for video saliency prediction. In particular, OFF-ViNet outperforms the state-of-the-art models on the UCF-Sports dataset, which contains many videos with moving targets.
3) Feature representations that predict the future are effective for predicting video saliency.

II. RELATED WORK

Saliency prediction includes image and video saliency predictions.

A. IMAGE SALIENCY PREDICTION

In 1998, Itti et al. [1] proposed a model that integrates color, intensity, and orientation features and models the human visual system of images. Since then, numerous handcrafted feature-based methods have been proposed [2]–[12]. Recently, with the development of deep learning, various models such as eDN [19], and DeepGaze IIE [20], which apply deep learning to saliency map prediction, have been proposed [19]–[23]. In static scenes, the saliency maps predicted by deep models are close to those created by multiple observer fixations. However, in dynamic scenes, the image saliency prediction model does not consider temporal information such as object motion, making proper prediction difficult.

B. VIDEO SALIENCY PREDICTION

Similar to image saliency prediction, a method based on handcrafted features has been proposed for video saliency prediction [57]–[70]. However, it is difficult to represent video saliency, which is dominated by various factors, in terms of handcrafted features. Therefore, deep learning-based models have been proposed for video saliency prediction [24]–[52]. In addition to spatial features, it is necessary to consider temporal features in video saliency prediction. Four structures were developed to capture spatiotemporal features for video saliency predictions. The first is a two-stream model that represents spatiotemporal information in two streams [24]–[28]. The second is a recurrent neural network (RNN)-based model, in which spatial features are accumulated by a convolutional neural network (CNN) and temporal features are extracted by an RNN, such as a convolutional long short-term memory (ConvLSTM) or convolutional gated recurrent unit (ConvGRU) [29]–[37]. The third is a 3D convolution-based model that enables simultaneous processing of spatiotemporal information using 3D convolution [38]–[49]. The last is a transformer-based model that processes long-range spatiotemporal features [50]–[52].

1) Two-stream model

Bak et al. proposed a two-stream network [24] to apply deep models to video saliency prediction. Two-stream networks fuse the outputs of the two CNNs, each with five layers. One network uses red, green, and blue (RGB) images, the other network uses optical flows as input, and the spatiotemporal features extracted from each are fused and used for prediction. Zhang et al. [25], Wu et al. [26], and Kocak et al. [27] studied how to fuse spatiotemporal information to improve the performance of a video saliency prediction model consisting of two streams. Fu et al. [28] proposed UVANet which merges two streams through transfer learning. UVANet is fast because it uses a student network based on knowledge distillation. However, it is difficult for the two-stream model to consider a long-time context because it uses optical flow to capture the motion between adjacent frames and a short-time sequence of RGB images as input.

2) RNN-based model

RNN-based models have been proposed to consider long-term temporal relationships. Bazzani et al. [29] proposed a method that inputs 16-frame clips into a 3D convolution, aggregates clip-level features using LSTM, a type of RNN, and outputs parameters for a mixed Gaussian model. Wang et al. proposed ACLNet [37], which introduced convLSTM into a CNN to account for long-term temporal relationships. ACLNet learns spatial saliency using the CNN and temporal features between frames using convLSTM. Droste et al. proposed UNISAL [36], which utilizes the features aggregated by MobileNetV2 [71] to predict saliency in a unified model for images and videos. UNISAL uses domain adaptation techniques, such as Domain-Adaptive Priors, Domain-Adaptive Fusion, Domain-Adaptive Smoothing and Bypass-RNN, to achieve highly accurate predictions for different datasets containing images and videos. UNISAL's Bypass-RNN models temporal features using convGRU, a type of RNN, when predicting saliency maps for videos. However, these models using RNNs cannot simultaneously process spatiotemporal information.

3) 3D convolution-based model

Models based on 3D convolution, which simultaneously process spatiotemporal information, exhibit high performance. In particular, methods using S3D [72] as the backbone, which have been pretrained on the Kinetics [73] dataset for action classification, have achieved high accuracy [40]–[46], [48],

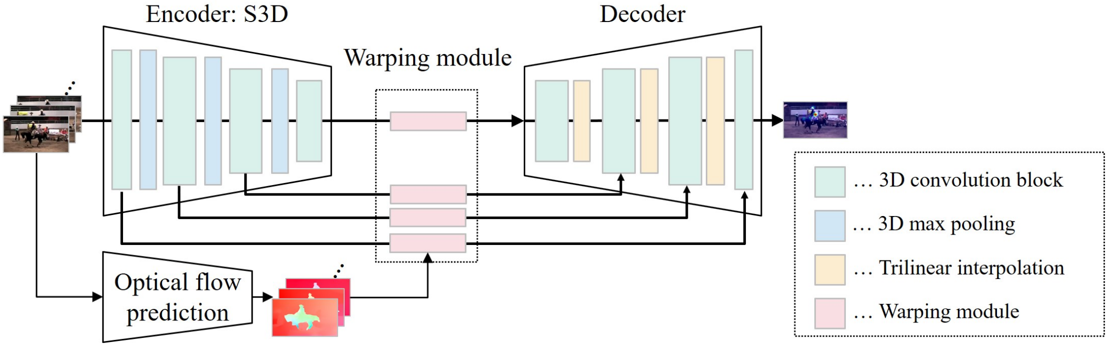

FIGURE 2. Structure of OFF-ViNet, which comprises S3D, an encoder of 3D convolution, a Warping module based on optical flow, and a decoder of 3D convolution. The Warping module acquires feature representations that predict the future.

[49]. Min et al. proposed TASED-Net, which uses pretrained S3D on the Kinetics dataset to simultaneously use spatiotemporal features [41]. Jain et al. proposed ViNet [42], which fuses features extracted by S3D with 3D convolution and trilinear interpolation in a U-Net-like manner [74]. Wang et al. proposed STSANet, which uses self-attention to consider long-range spatiotemporal features extracted by S3D [45]. STSANet connects a spatiotemporal self-attention (STSA) module to each of the four low to high-order blocks of S3D that have been pretrained on the Kinetics dataset and upsamples and integrates the features to output a final saliency map. The STSA module captures long-range spatiotemporal features by aggregating global relationships from local features accumulated by the 3D convolutional neural network.

4) Transformer-based model

Recently, transformers that efficiently compute long-range dependencies have been presented in the area of image recognition [75], [76]. In video recognition, various video transformers [77]–[79] inspired by the success of image transformers have been proposed. In video saliency prediction, attempts have been made to capture long-range spatiotemporal features using a transformer as the backbone, instead of a 3D convolution-based backbone with a local receptive field [50]–[52]. Ma et al. proposed a pure transformer-based VSFT that uses blocks from a Video Swin Transformer [78] as the backbone of the model [50]. Zhou et al. proposed TMFI-Net [51] that decodes multiscale features captured by the backbone Video Swin Transformer [78]. The above model achieves an accurate prediction by implicitly learning feature representations that are useful for video saliency prediction from a large dataset in a data-driven manner. The video frames switch to the next frame in a very small amount of time per frame. Therefore, it is assumed that a human gazing at a video frame determines the current gaze position based on the past frames. Thus, it is useful to explicitly introduce feature representations that predict the future of video saliency prediction.

III. PROPOSED MODEL

This section describes the proposed OFF-ViNet method in detail. OFF-ViNet is based into ViNet [42], which is a state-of-the-art model for video saliency prediction. First, ViNet is modified to incorporate a Warping module. OFF-ViNet introduces Warping modules to the modified ViNet to explicitly obtain feature representations that predict the future. It then provides appropriate video saliency prediction for videos with moving objects. Sections III-A, III-B and III-C present an overview of the architecture, Warping module, and implementation details, respectively.

A. ARCHITECTURE OVERVIEW

Fig. 2 shows the OFF-ViNet architecture. OFF-ViNet takes a sequence of frames $X_{in} \in \mathbb{R}^{3\times T_{in}\times H\times W}$ from time t to time $t - T_{in} + 1$, where $T_{in} = 32$ and predicts the saliency of the frame at time t. A modified ViNet is used as the baseline. ViNet [42] uses S3D as the backbone, which has been pretrained on the Kinetics dataset, as in existing 3D convolution-based models [40]–[46], [48]. Low- to high-order features extracted from the four blocks in the backbone are skip-connected in a U-Net-like manner and used for decoding. The ViNet decoder uses a concatenation method along the time dimension with some parameters to fuse hierarchical features. However, OFF-ViNet uses feature representations that predict the future; therefore, it changes to concatenation along the channel dimension, preserving the temporal relationships and fusing the hierarchical features.

A Warping module based on optical flow is used as a mechanism to obtain feature representations that predict the future of OFF-ViNet. The prediction mechanism is incorporated in the skip connection part of the encoder-decoder structure and applies each of the multi-time optical flows $\mathbf{flow} \in \mathbb{R}^{2\times T\times H\times W}$ to the hierarchical features $\mathbf{F^1}, \mathbf{F^2}, \mathbf{F^3}, \mathbf{F^4} \in \mathbb{R}^{C\times T\times H\times W}$ extracted from S3D to obtain a feature representation that predicts the future, where C, T, H, and W denote the number of channels, temporal length, height, and width, respectively. RAFT [80] is used as the optical flow model.

B. WARPING MODULE

Fig. 3 shows the structure of the Warping module. The Warping module obtains a feature representation that predicts the future from time t_w $(t - T_{in} < t_w \leq t)$ when the model predicts the saliency at time t. The Warping module applies pointwise

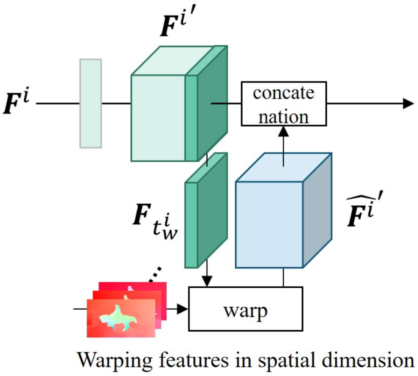

FIGURE 3. Structure of Warping module. A Warping module warps the features with half the channel dimensionality by 3D pointwise convolution based on the optical flow.

convolution to the feature $\mathbf{F}^i \in \mathbb{R}^{C^i \times T^i \times H^i \times W^i}(1 \leq i \leq 4)$ to obtain $\mathbf{F}^{i'} \in \mathbb{R}^{\frac{C^i}{2} \times T^i \times H^i \times W^i}$, where C^i, T^i, H^i, and W^i denote the number of channels, temporal length, height, and width in $\mathbf{F}^i$, respectively. $\mathbf{F}_{t_w^i} \in \mathbb{R}^{\frac{C^i}{2} \times T^i \times H^i \times W^i}$ is used to obtain the feature $\mathbf{F}^{i'} \in \mathbb{R}^{\frac{C^i}{2} \times T^i \times H^i \times W^i}$ that predicts the future, where $\mathbf{F}_{t_w^i}$ is the element in $\mathbf{F}^{i'}$ at temporal position $t_w^i = \lfloor (t - t_w) \times T^i / T_{in} \rfloor (0 \leq t_w^i < T^i)$ corresponding to t_w, which is the time to predict the future. The feature $\mathbf{F}_{t_w^i}$ is warped by the multi-time optical flow $\mathbf{flow} \in \mathbb{R}^{2 \times T^i \times H^i \times W^i}$ to obtain a feature $\mathbf{F}^{i'}$ that predicts the future. Finally, $\mathbf{F}^{i'}$ and $\hat{\mathbf{F}}^{i'}$ are concatenated over the channel dimension.

Fig. 4 shows as overview of the warp. Warp is an operation that determines the value of each pixel at $\mathbf{F}_{t_w}$ from the pixel value at the end of the optical flow at that pixel. Forward-warping [81] is performed to obtain a feature representation that predicts the feature. Forward-warping is an operation to warp from the optical flow of the frame at time t_w and time $t_w - n$ to the optical flow of the frame at time $t_w + n$ and time t_w, from the optical flow of frames at time t_w and time $t_w - n$. Let $\mathbf{flow}_{t_w \to t_w - n}(\mathbf{x})$ be the optical flow at position $\mathbf{x}$ from time t_w to time $t_w - n$. The optical flow for forward-warp $\hat{\mathbf{flow}}_{t_w+n \to t_w}(\mathbf{x})$ follows the following equation:

$$\hat{\mathbf{flow}}_{t_w+n \to t_w}(\mathbf{x} - \mathrm{round}(\mathbf{flow}_{t_w \to t_w - n}(\mathbf{x}))) = \mathbf{flow}_{t_w \to t_w - n}(\mathbf{x}) \quad (1)$$

The feature $\mathbf{F}_{t_w \to t_w + n}(\mathbf{x})$ at $t_w + n$ predicted in the future then obeys the following equation:

$$\mathbf{F}_{t_w \to t_w + n}(\mathbf{x}) = \mathbf{F}_{t_w^i}(\mathbf{x} + \hat{\mathbf{flow}}_{t_w+n \to t_w}(\mathbf{x})) \quad (2)$$

where $\mathbf{F}_{t_w^i}$ is the feature of $\hat{\mathbf{F}}^{i'}$ at position $\mathbf{x}$ at any time $t_w - i$. The features $\mathbf{F}_{t_w \to t_w+1}(\mathbf{x})$, $\mathbf{F}_{t_w \to t_w+2}(\mathbf{x})$,. . ., $\mathbf{F}_{t_w \to t_w+T^i}(\mathbf{x}) \in \mathbb{R}^{\frac{C^i}{2} \times 1 \times H^i \times W^i}$ obtained by varying $n(1 \leq n \leq T^i)$ by one in the above warp are concatenated over the time dimension to obtain $\mathbf{F}^{i'} \in \mathbb{R}^{\frac{C^i}{2} \times T^i \times H^i \times W^i}$.

A multiscale optical flow is needed to apply the warp to multiscale features $\mathbf{F}^i \in \mathbb{R}^{\frac{C^i}{2} \times T^i \times H^i \times W^i}$. Therefore, the optical flow $flow_H$ estimated at the same scale as the input

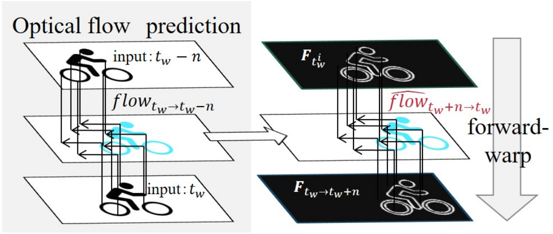

FIGURE 4. Overview of warping. Forward-warping is performed on features to obtain a feature representation that predicts the future.

$\mathbf{X}_{in}$ is downsampled to obtain an optical flow $flow_L(\mathbf{x})$ with the same spatial resolution as the feature $\mathbf{F}^i$ to be warped. The downsampling of the optical flow by a factor of $1/k$ follows the following equation:

$$flow_L(\mathbf{x}) = flow_H\big(\operatorname*{argmax}_{\mathbf{m} \in \mathbb{R}^{k \times k}}(\|flow_H(k\mathbf{x} + \mathbf{m})\|)\big) \quad (3)$$

where $\| \cdot \|$ denotes the Euclidean norm.

C. IMPLEMENTATION DETAILS

OFF-ViNet uses 32 consecutive $T_{in} = 32$ frames as input sequences. The temporal dimensions of the feature output from the four blocks are 16, 16, 8, and 4, respectively, since the temporal dimensions is compressed by half in S3D block1, block3, and block4 of the backbone. The time t_w is set to 3, corresponding to 0.1s in a 30 fps video. Therefore, the features to be warped in each block are $t_w^i(1 \leq i \leq 4) = 1, 1, 0$ and 0 from the beginning.

In the video, the number of input pictures $T_{in} = 32$ is ≤ 32 when $t \leq 31$. In DHF1K, which is a typical dataset for video saliency prediction, a black screen was presented to the observer before starting the video when collecting the fixation points. In OFF-ViNet, when the number of input images is $\leq T_{in} = 32$, the zero padding constructs the input sequence in the temporal direction, which is equivalent to inputting black frames, according to the DHF1K data collection conditions. The loss function used in training is the same as that in STSANet [45] and TMFI-Net [51], using Kullback-Leibler (KL) Divergence and Pearson's correlation coefficient (CC), and is expressed as follows:

$$\mathrm{Loss}(P, Q) = \mathrm{KL}(P, Q) - \mathrm{CC}(P, Q) \quad (4)$$

where P is the predicted saliency map, Q is the ground truth, and KL is the dissimilarity between the two distributions, calculated as follows:

$$\mathrm{KL}(P, Q) = \sum_i Q_i \log(\epsilon + \frac{Q_i}{\epsilon + P_i}) \quad (5)$$

where ϵ is the regularization constant. CC represents the correlation between the two distributions and is calculated as follows:

$$\mathrm{CC}(P, Q) = \frac{\mathrm{cov}(P, Q)}{\mathrm{sd}(P) \times \mathrm{sd}(Q)} \quad (6)$$

where sd and cov are the standard deviation and the covariance, respectively.

IV. EXPERIMENT

This section presents the experimental results and analysis of the proposed OFF-ViNet. Sections IV-A, IV-B, IV-C, IV-D, IV-E, and IV-F present the dataset used in the experiments, experimental conditions, metrics used to evaluate the performance of the model, the comparison of the experimental results with other state-of-the-art models, ablation studies, and the discussion, respectively.

A. DATASETS

This section presents the following representative datasets for video saliency prediction: DHF1K, Hollywood-2, and UCF-Sports. These datasets include a video, the positions of multiple observer fixations on the video, and the ground truth of the saliency map created from the fixations.

1) DHF1K [37]

This dataset consists of 1000 videos with various scenes, movements, and objects provided by Wang et al. and the fixation points of 17 observers on the videos. The dataset is divided into 600, 100, and 300 training, validation, and test dataset, respectively. No ground truth is provided for the test data and prediction results are submitted to the benchmark website for evaluation.

2) Hollywood-2 [82]

This dataset contains videos and gazing points created from 1707 movies provided by Mathe et al. Videos with multiple resolutions are presented to the observer and resized to fit the display while maintaining the aspect ratio of the video. Similar to Wang et al. [37], 823 and 884 movies are used in this study for training and testing, respectively.

3) UCF-Sports [82]

This dataset contains 150 videos and gazing point data provided by Mathe et al. The videos were extracted from sports videos of nine different actions, many of which have moving objects. Similar to Wang et al. [37], 103 and 47 videos are used in this study for training and testing, respectively.

B. EXPERIMENTAL SETUP

The proposed model was implemented on an NVIDIA A100 using PyTorch [83]. The S3D of the backbone was pretrained on the Kinetics dataset. The RAFT for the optical flow estimation model was a pretrained model provided by PyTorch. Default PyTorch settings were used to initialize the weights of the other parameters. The Adam optimizer [84] was used as the optimization function, with an initial learning rate of 1.0×10^{-3}. In the case of training stagnation, the proposed model was optimized by reducing the learning rate 0.1 times. All input videos were resized to resolution of 224×384, and the batch size was 8.

C. METRICS

There are several methods for evaluating saliency, depending on the format of the ground truth data and treatment of the negative sample [85]. This experiment used the similarity (SIM), Pearson's correlation coefficient (CC), which uses the ground truth as the saliency map, and the Normalized Scanpath Saliency (NSS), and two types of area under the curve (AUC), i. e., AUC-Judd (AUC-J) and shuffled AUC (sAUC), which use the ground truth as the location of the fixations of the observer.

D. COMPARISON RESULTS

OFF-ViNet was compared with state-of-the-art video saliency prediction models, such as DeepVS [32], ACLNet [37], STRANet [34], SalEMA [33], TASED-Net [41], Chen et al [31]. SalSAC [35], UNISAL [36], ECANet [46], HD2S [43], ViNet [42], TSFP-Net [44], STSANet [45], TinyHD [47], HFTR-Net [48], TMFI-Net [51], UniST [52], and MSFF-Net [49]. Table 1 presents quantitative evaluations on the test datasets of DHF1K, Hollywood-2, and UCF-Sports.

On the DHF1K column, the proposed model achieved an accuracy second only to TMFI-Net in AUC-J, CC, and NSS and surpassed TMFI-Net in SIM, placing it second overall, after SalEMA. SalEMA, achieving the highest accuracy in SIM, is less accurate than the other state-of-the-art models in terms of the other four metrics. On the Hollywood-2 column, the proposed model achieved the second-best accuracy in AUC-J, following UniST, and the third-best accuracy in SIM, CC, and NSS, following UniST and TMFI-Net. In the UCF-Sports column, the proposed model achieved the highest overall accuracy in the SIM, CC, and NSS evaluation metrics and the third-best accuracy in AUC-J. The UCF-Sports dataset contains many videos with moving objects, which are considered easy to predict by OFF-ViNet. Table 2 shows quantitative comparison of our model with SalEMA [33], TASED-Net [41], Chen et al [31]. SalSAC [35], UNISAL [36], ECANet [46], HD2S [43], ViNet [42], TSFP-Net [44], STSANet [45], TinyHD [47], HFTR-Net [48], TMFI-Net [51], and UniST [52], and MSFF-Net [49] on the validation dataset of DHF1K. The proposed model achieved the second-best accuracy in AUC-J and NSS. Fig. 5 shows a qualitative comparison of proposed OFF-ViNet and the code-available, state-of-the-art STSANet [45], ViNet [42], TASED-Net [41], UNISAL [36] and OFF-ViNet in terms of video saliency prediction. OFF-ViNet predicts saliency in appropriate regions, particularly in videos with moving objects.

E. ABLATION STUDIES

This section verifies the effectiveness of the proposed OFF-ViNet component, that is, the Warping module, on the DHF1K dataset. The training and validation data of the DHF1K dataset were used for the training and evaluation, respectively. Fig. 6 shows the structure of OFF-ViNet without the Warping module, and this model is referred to as w/o Warping module. The Warping module uses pointwise convolution to obtain a feature representation that predicts the future from half of the dimensions of the channel. The dimensions of the features that are skip-connected from the encoder to the decoder do not change between the input

TABLE 1. Quantitative comparison of various models of video saliency prediction results on the test datasets of DHF1K, Hollywood-2, and UCF-Sports.

	DHF1K					Hollywood-2				UCF-Sports			
model	AUC-J	SIM	sAUC	CC	NSS	AUC-J	SIM	CC	NSS	AUC-J	SIM	CC	NSS
DeepVS [32]	0.856	0.256	0.583	0.344	1.911	0.887	0.356	0.446	2.313	0.870	0.321	0.405	2.089
ACLNet [37]	0.890	0.315	0.601	0.434	2.354	0.886	0.542	0.623	3.086	0.897	0.406	0.510	2.567
STRANet [34]	0.895	0.355	0.663	0.458	2.558	0.923	0.536	0.662	3.478	0.910	0.479	0.593	3.018
SalEMA [33]	0.890	0.466	0.667	0.449	2.574	0.919	0.487	0.613	3.186	0.906	0.431	0.544	2.638
TASED-Net [41]	0.895	0.361	0.712	0.470	2.667	0.918	0.507	0.646	3.302	0.899	0.469	0.582	2.920
Chen et al. [31]	0.900	0.353	0.680	0.476	2.685	0.928	0.537	0.661	3.804	0.917	0.494	0.599	3.406
SalSAC [35]	0.896	0.357	0.697	0.479	2.673	0.931	0.529	0.670	3.356	0.926	0.534	0.671	3.523
UNISAL [36]	0.901	0.390	0.691	0.490	2.776	0.934	0.542	0.673	3.901	0.918	0.523	0.644	3.381
ViNet [42]	0.908	0.381	0.729	0.511	2.872	0.930	0.550	0.693	3.730	0.924	0.522	0.673	3.620
HD2S [43]	0.908	0.406	0.700	0.503	2.812	0.936	0.551	0.670	3.352	0.904	0.507	0.604	3.114
STA3D [39]	0.908	0.390	0.721	0.515	2.877	0.927	0.534	0.659	3.329	0.900	0.465	0.560	2.754
ECANet [46]	0.903	0.385	0.717	0.500	2.814	0.929	0.526	0.673	3.380	0.917	0.498	0.636	3.189
TSFP-Net [44]	0.912	0.392	0.723	0.517	2.966	0.936	0.571	0.711	3.910	0.923	0.561	0.685	3.698
STSANet [45]	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
VSFT [50]	0.911	0.411	0.720	0.519	2.977	0.936	0.577	0.703	3.916	-	-	-	-
GFNet [40]	0.913	0.379	0.724	0.526	2.995	0.938	0.585	0.719	3.952	0.933	0.544	0.694	3.723
TinyHD [47]	0.909	0.396	0.714	0.505	2.921	0.935	0.561	0.690	3.815	0.918	0.510	0.624	3.280
HFTR-Net [48]	0.914	0.391	0.731	0.536	3.086	0.940	0.572	0.724	3.930	0.939	0.563	0.702	3.910
TMFI-Net [51]	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
UniST [52]	-	-	-	-	-	0.951	0.632	0.777	4.397	0.932	0.576	0.706	3.718
MSFF-Net [49]	0.913	0.392	0.728	0.534	3.066	0.940	0.574	0.723	3.930	0.939	0.563	0.710	3.913
OFF-ViNet	0.914	0.419	0.726	0.538	3.089	0.942	0.594	0.737	4.051	0.936	0.589	0.730	4.180

The best and second-best scores are marked by red and blue respectively.

TABLE 2. Quantitative comparison of various models of video saliency prediction results on the validation dataset of DHF1K.

	DHF1K				
model	AUC-J	SIM	sAUC	CC	NSS
SalEMA [33]	0.886	0.360	0.690	0.450	2.495
TASED-Net [41]	0.894	0.362	0.718	0.481	2.706
Chen et al. [31]	0.905	0.358	0.689	0.467	2.651
SalSAC [35]	0.898	0.364	0.729	0.480	2.624
UNISAL [36]	0.907	0.381	0.691	0.487	2.755
ViNet [42]	-	0.388	-	0.521	2.957
HD2S [43]	0.904	0.403	0.705	0.489	2.806
STA3D [39]	0.911	0.385	0.624	0.516	2.877
ECANet [46]	0.910	0.394	0.725	0.515	2.877
TSFP-Net [44]	0.919	0.398	-	0.529	3.009
STSANet [45]	0.920	0.411	-	0.539	3.082
VSFT [50]	0.909	0.409	0.721	0.522	2.992
GFNet [40]	0.920	0.402	-	0.542	3.088
TinyHD [47]	0.908	0.389	-	0.495	2.874
HFTR-Net [48]	-	0.425	-	0.559	2.901
TMFI-Net [51]	0.924	0.428	-	0.554	3.201
UniST [52]	0.920	0.423	-	0.541	3.113
MSFF-Net [49]	-	0.421	-	0.557	2.885
OFF-ViNet	0.922	0.418	0.734	0.548	3.143

The best and second-best scores are marked by red and blue respectively.

and output of the Warping module. Therefore, the number of encoder and decoder parameters is the same for the w/o Warping module and OFF-ViNet.

Table 3 presents a quantitative comparison of OFF-ViNet and the w/o Warping module. OFF-ViNet showed the highest accuracy for all evaluation metrics. This experiment shows that the features obtained from the Warping module are more effective for video saliency prediction than those obtained using implicit learning from the data.

F. DISCUSSION

This section presents failure cases of OFF-ViNet and analyzes the limitations of the proposed model. Fig. 7 shows the failure cases of OFF-ViNet. Fig. 7 (a) shows no specific object of interest and where the camera zooms to the center of the frame. The w/o Warping module performs relatively well, whereas OFF-ViNet predicts saliency not only in the center of the image but also over a wide area on the bottom and right sides of the image. When the camera zooms, the optical flow is directed outward from the image. Therefore, the warped features in the Warping module using optical flow are spread out in concentric circles. This study suggests that regions of high saliency appear in a wide range of regions. This indicates that the proposed Warping module is ineffective for videos with intense camera motion and no specific salient objects. Although a feature representation that predicts the future using optical flow was explicitly obtained, this study expects that the performance of video saliency prediction can be improved by using a feature representation that predicts the future and is more suitable for video saliency prediction without relying on optical flow.

Fig. 7 (b) shows a scene with multiple objects. Both OFF-ViNet and the w/o Warping module predict saliency over a wide range but not properly for the ground truth. In the ground truth, saliency is scattered over various regions within the frame. Fig. 8 shows the relationship between the sum of the pixel values of the normalized ground truth saliency maps averaged for each video and the average OFF-ViNet evaluation value over each video in the DHF1K validation dataset. The sum of the pixel values in the normalized ground truth saliency map represents the size of the salient regions. Moreover, the larger the value, the more salient the regions that appear in various areas of the frame. As shown in Fig. 8,

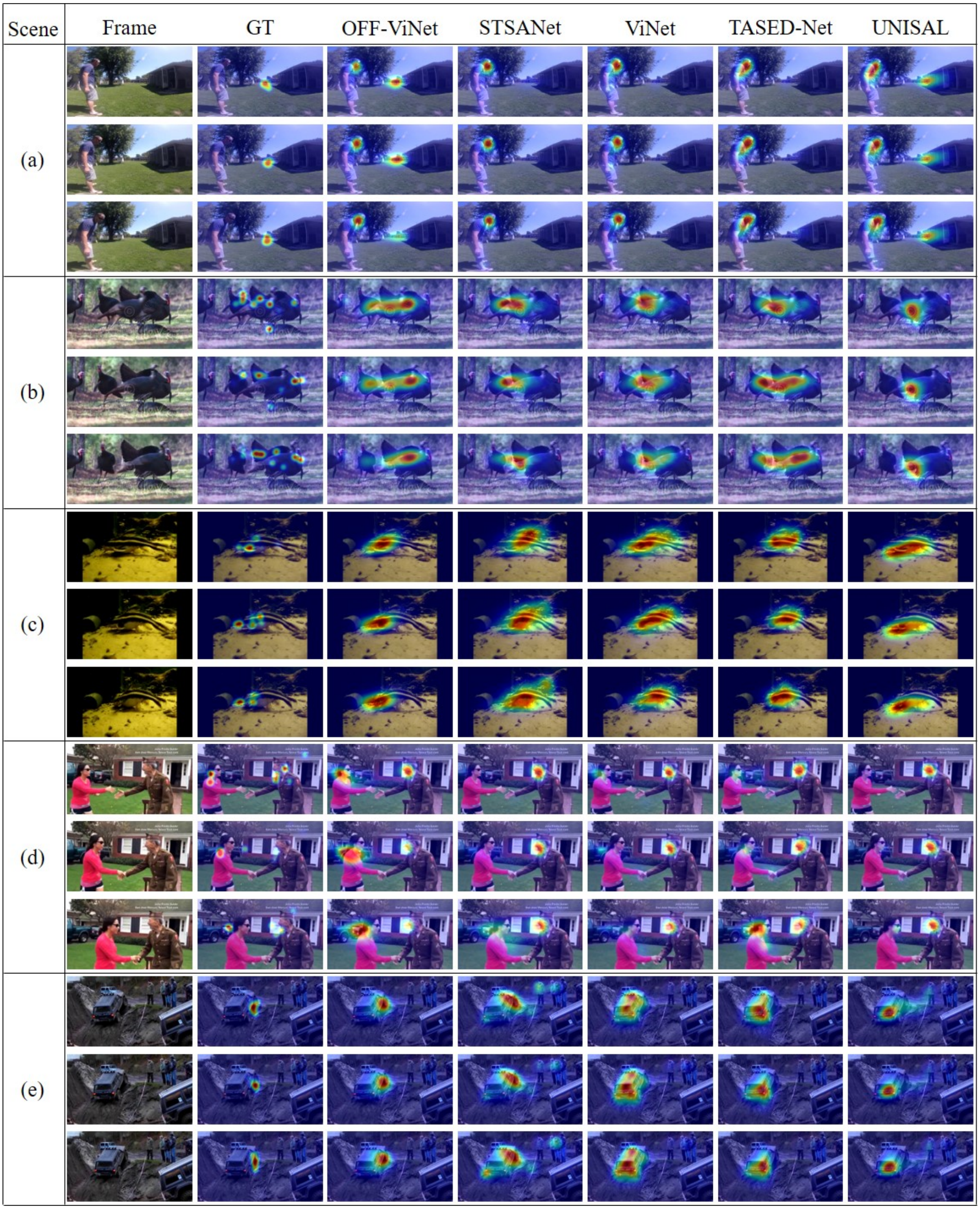

FIGURE 5. Qualitative comparison of various video saliency prediction models. A comparison with ground truth (GT), OFF-ViNet, STSANet, ViNet, TASED-Net, and UNISAL is shown for each scene, sampling three frames per scene. We used the source code published by the authors of TMFI-Net for training and prediction, and did not obtain valid results due to insufficient convergence of the loss during training. Therefore, the results are not shown here.

a negative correlation exists between the CC of OFF-ViNet and the size of the salient regions. Therefore, OFF-ViNet has difficulty predicting video saliency for videos in which salient objects are in a large area within the frame. However, it is

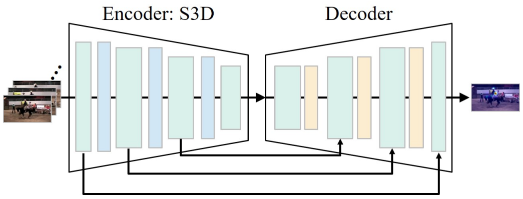

FIGURE 6. Structure of the w/o Warping module, which consists of an encoder and a decoder with the same parameters as OFF-ViNet.

TABLE 3. Ablation study on Warping module

model	AUC-J	SIM	sAUC	CC	NSS
w/o Warping module	0.920	0.414	0.730	0.542	3.111
OFF-ViNet	0.922	0.418	0.734	0.548	3.143

Ablation study on the validation set of DHF1K.

possible that video saliency prediction is difficult for these videos. When there are many salient objects in a frame, it is difficult for the observer and model to allocate attention appropriately, making video saliency prediction difficult. These videos are also likely to be insufficiently annotated in the dataset. Fig. 7 (b) shows an image from a video with the widest area of saliency in the DHF1K validation data. When saliency is sparsely distributed, as in this video, the number of observers may not be sufficient to determine the size of the salient regions in the frame. In this case, there is a risk that the ground truth of the dataset may diverge from the true ground truth.

V. CONCLUSION

This study proposes OFF-ViNet, which is based on ViNet, a state-of-the-art video saliency prediction model, with the addition of a Warping module, a mechanism to explicitly predict future features. The results of quantitative and qualitative experiments on a representative video saliency prediction dataset show that OFF-ViNet is competitive with existing state-of-the-art models. In particular, OFF-ViNet outperforms the existing models in the evaluation metrics SIM, CC, and NSS on the UCF-Sports dataset, which contains several videos with moving objects. The ablation study shows the usefulness of the Warping module for video saliency prediction and that feature representation predicting the future is effective for video saliency prediction.

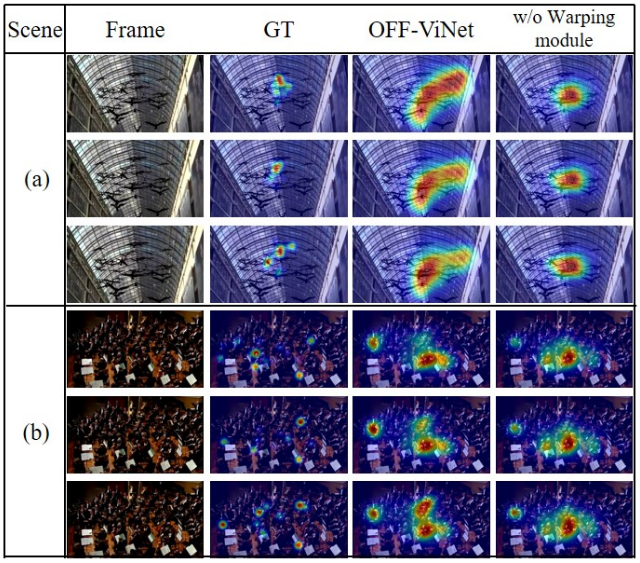

FIGURE 7. Failure cases. Comparison with ground truth (GT), OFF-ViNet, and w/o Warping module.

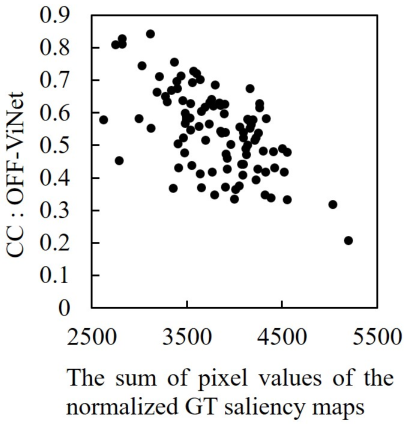

FIGURE 8. Relationship between the sum of pixel values of the normalized ground truth saliency maps averaged over each video and the average OFF-ViNet evaluation value over each video. Pearson's Correlation Coefficient (CC) is used as the evaluation value for OFF-ViNet.

REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Mach. Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Adv. in Neural Inf. Process. Syst.*, pp545–552, 2006.

[3] H. J. Seo, and P. Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," *2009 IEEE Comput. Society Conf. on Comput. Vis. and Pattern Recognit. Workshops*, Miami, FL, USA, pp. 45–52, 2009.

[4] B. Schauerte, and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," *Europian Conf. on Comput. Vis.*, pp. 116–129, 2012.

[5] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," *2009 IEEE 12th Int. Conf.on Comput. Vis., Kyoto, Japan*, pp. 2106–2113, 2009.

[6] X. Hou, J. Harel, and C. Koch , "Image signature: highlighting sparse salient regions," *IEEE Trans. on Pattern Analysis and Mach. Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[7] N. Riche, M. Mancas, M. Duvinage, B. Gosselin, and T. Dutoit, "RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Processing: Image Communication*, vol. 28, Issue 6, pp. 642–658, 2013.

[8] N. Bruce, and J. Tsotsos, "Saliency based on information maximization," *Neural inf. Process. Syst.*, pp. 155–162, 2005.

[9] J. Zhang, and S. Sclaroff , "Saliency detection: A boolean map approach," *IEEE Int. Conf. on Comput. Vis.*, pp. 153–160, 2013.

[10] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell , "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vis.*, vol. 8, no. 7, pp. 32–32, 2008.

[11] X. Hou, and L. Zhang , "Dynamic visual attention: searching for coding length increments," *Int. Conf. on Neural Inf. Process. Syst.*, pp. 681-688, 2008.

[12] K. Ishikura, N. Kurita, D. M. Chandler, and G. Ohashi, "Saliency detection based on multiscale extrema of local perceptual color differences," *IEEE Trans. on Image Process.*, vol. 27, no. 2, pp. 703–717, 2018.

[13] B. Nicholas, Z. Lingyun, C. Garrison, and M. Javier, "Visual saliency model for robot cameras," *IEEE Int. Conf. on Robotics and Automation*, Pasadena, CA, USA, pp. 2398–2403, 2008.

[14] N. Tam, X. Mengdi, G. Guangyu, K. Mohan, and T. Qi, Y. Shuicheng, "Static saliency vs. Dynamic saliency: A comparative study," *ACM Multimedia Conf.*, pp.987–996, 2013.

[15] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image and Vis. Computing*, vol. 29, no. 1, pp. 1–14, 2011.

[16] H. Hadi, and B. Ivan , "Saliency-aware video compression," *IEEE trans. on image process. : a publication of the IEEE Signal Process. Society,* 2013.

[17] L. Gongyang, L. Zhi, S. Ran, and W. Weijie, "Constrained fixation point based segmentation via deep neural network," *Neurocomputing*, vol. 368, no. 27, pp. 180–187, 2019.

[18] W. Wang, J. Shen, R. Yang, and F. Porikli , "Saliency-aware video object segmentation," *IEEE Trans. on Pattern Analysis and Mach. Intelligence*, vol. 40, no. 1, pp. 20–33, 2018.

[19] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," *IEEE Conf. on Comput. Vis. and Pattern Recognit.*, pp. 2798–2805, 2014.

[20] A. Linardos, M. Kümmerer, O. Press, and M. Bethge, "DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling,"*IEEE/CVF Int. Conf. on Comput. Vis.*, pp. 12919–12928, 2021.

[21] S. Jia, N. D. B. Bruce, "EML-NET: An expandable multi-layer network for saliency prediction," *Image and Vis. Computing*, vol. 95, 2020.

[22] G. Ding, N. İmamoğlu, A. Caglayan, M. Murakawa, and R. Nakamura, "SalFBNet: Learning pseudo-saliency distribution via feedback convolutional networks," *Image and Vis. Computing*, 2022.

[23] J. Low, H. Lin, D. Marshall, D. Saupe, and H. Liu, "TranSalNet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, vol. 494, no. 14, pp. 455–467, 2022.

[24] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Trans. on Multimedia*, vol. 20, no. 7, pp. 1688–1698, 2018.

[25] K. Zhang, and Z. Chen, "Video saliency prediction based on spatial-temporal two-stream network," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3544–3557, Dec. 2019.

[26] Z. Wu, L. Su, and Q. Huang, "Learning coupled convolutional networks fusion for video saliency prediction," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 2960–2971, Oct. 2019.

[27] A. Kocak, E. Erdem, and A. Erdem, "A gated fusion network for dynamic saliency prediction," *IEEE Trans. on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 995–1008, Sept. 2022.

[28] K. Fu, P. Shi, Y. Song, S. Ge, X. Lu, and J. Li, "Ultrafast video attention prediction with coupled knowledge distillation," in *Proc. AAAI Conf. on Artificial Intelligence*, vol. 34, no. 7, pp. 10802–10809, 2020.

[29] L. Bazzani, H. Larochelle, and L. Torresani, "Recurrent mixture density network for spatiotemporal visual attention," in *Proc. Int. Conf. Learning Represenation*, 2017.

[30] S. Gorji, and J. J. Clark, "Going from Image to Video Saliency: Augmenting Image Salience with Dynamic Attentional Push," *IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, pp. 7501–7511, 2018.

[31] J. Chen, H. Song, K. Zhang, B. Liu, and Q. Liu, "Video saliency prediction using enhanced spatiotemporal alignment network," *Pattern Recognit.*, vol. 109, 2021.

[32] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "DeepVS: A deep learning based video saliency prediction approach," in *Proc. Eur. Conf. Comput. Vis.*, pp. 602–617, 2018.

[33] P. Linardos, E. Mohedano, J. J. Nieto, N. E. O'Connor, X. Giro-i-Nieto, and K. McGuinness, "Simple vs complex temporal recurrences for video saliency prediction," *30th British Machine Vis. Conf.*, Sep. 2019.

[34] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *IEEE Trans. on Image Process.*, vol. 29, pp. 1113–1126, 2020.

[35] X. Wu, Z. Wu, J. Zhang, L. Ju, and S. Wang, "SalSAC: A Video Saliency Prediction Model with Shuffled Attentions and Correlation-Based ConvLSTM," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 34, no. 7, 2020.

[36] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," *European Conf. on Comput. Vis.*, Springer, Cham, pp. 419–435, 2020.

[37] W. Wang, J. Shen, J. Xie, M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Analysis Mach. Intelligence*, vol. 43, no. 1, pp. 220–237, Jan. 2021.

[38] Z. Sun, X. Wang, Q. Zhang, and J. Jiang, "Real-time video saliency prediction via 3D residual convolutional neural network," in *IEEE Access*, vol. 7, pp. 147743–147754, 2019

[39] W. Zou, S. Zhuo, Y. Tang, S. Tian, X. Li, and C. Xu, "STA3D: Spatiotemporally attentive 3D network for video saliency prediction," *Pattern Recognit. Letters*, vol. 147, pp. 78–84, 2021.

[40] S. Wu, X. Zhou, Y. Sun, Y. Gao, Z. Zhu, J. Zhanga, and C. Yan, "GFNet: gated fusion network for video saliency prediction," *Appl. Intelligence*, vol. 53, pp. 27865–27875, 2023.

[41] K. Min, and J. Corso, "TASED-Net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 2394–2403, Oct. 2019.

[42] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, "ViNet: Pushing the limits of visual modality for audio-visual saliency prediction," *IEEE/RSJ Int. Conf. on Intell. Robots and Systems*, Prague, Czech Republic, pp. 3520–3527, 2021.

[43] G. Bellitto, F. P. Salanitri, S. Palazzo, F. Rundo, D. Giordano, and C. Spampinato, "Hierarchical domain-adapted feature learning for video saliency prediction," *Int. Journal of Comput. Vis.*, vol. 129, pp. 3216–3232, 2021.

[44] Q. Chang, and S. Zhu, "Human vision attention mechanism-inspired temporal-spatial feature pyramid for video saliency detection," *Cognitive Computation*, vol. 15, pp. 856–868, 2023.

[45] Z. Wang, Z. Liu, G. Li, Y. Wang, T. Zhang, L. Xu, and J. Wang, "Spatio-temporal self-attention network for video saliency prediction," *IEEE Trans. Multimedia*, vol. 25, pp. 1161–1174, 2023.

[46] H. Xue, M. Sun, and Y. Liang, "ECANet: Explicit cyclic attention-based network for video saliency prediction," *Neurocomputing*, vol. 468, pp. 233–244, Jan. 2022.

[47] F. Hu, S. Palazzo, F. P. Salanitri, G. Bellitto, M. Moradi, C. Spampinato, and K. McGuinness, "TinyHD: Efficient video saliency prediction with heterogeneous decoders using hierarchical maps distillation," *IEEE/CVF Winter Conf. on Applications of Comput. Vis.*, Waikoloa, HI, USA, pp. 2050–2059, 2023.

[48] Y. Zhang, T. Zhang, C. Wu, and Y. Zheng, "Accurate video saliency prediction via hierarchical fusion and temporal recurrence," *Imag. and Vis. Comput.*, vol. 136, 2023.

[49] Y. Zhang, T. Zhang, C. Wu, and R. Tao, "Multi-scale spatiotemporal feature fusion network for video saliency prediction," *IEEE Trans. on Multimedia*, vol. 26, pp. 4183–4193, 2023.

[50] C. Ma, H. Sun, Y. Rao, J. Zhou, and J. Lu, "Video saliency forecasting transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6850–6862, Oct. 2022.

[51] X. Zhou, S. Wu, R. Shi, B. Zheng, S. Wang, H. Yin, J. Zhang, and C. Yan, "Transformer-based multi-scale feature integration network for video saliency prediction," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7696–7707, Dec. 2023

[52] J. Xiong, P. Zhang, C. Li, W. Huang, Y. Zha, and T. You, "UniST: Towards unifying saliency transformer for video saliency prediction and detection," *arXiv.2309.08220*, 2023.

[53] A. Borji, "Saliency prediction in the deep learning era: successes and limitations," in *IEEE Trans. on Pattern Analysis and Mach. Intelligence*', vol. 43, no. 2, pp. 679–700, 1 Feb. 2021

[54] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Vis. Cognition*, vol. 12, no. 6, pp. 1093–1123, 2005.

[55] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cognitive Computation*, vol. 3, no. 1, pp.5–24, 2011.

[56] R. Ikenoya, and G. Ohashi, "Fixation Analysis for Video Saliency Prediction," in *IEEJ Trans. on Electronics, Information and Systems*, vol. 143, no. 9, pp. 885–894, 2023, (in Japanese).

[57] Y. -F. Ma, and H. -J. Zhang, "A new perceived motion based shot content representation," *Proc. of the Int. Conf. on Image Process.*, vol. 3, pp. 426–429, 2001.

[58] Y. -F. Ma, and H. -J. Zhang, "A model of motion attention for video skimming,"*Proc. of the Int. Conf. on Image Process.*, 2002.

[59] G. Agarwal, A. Anbu, and A. Sinha, "A fast algorithm to find the region-of-interest in the compressed MPEG domain," *Proc. 2003 Int. Conf. on Multimedia and Expo. ICME '03.*, pp. II–133.

[60] A. Sinha, G. Agarwal and A. Anbu, "Region-of-interest based compressed domain video transcoding scheme," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Montreal*, pp. iii–161, 2004.

[61] X. Hou, and L. Zhang, "Dynamic visual attention: searching for coding length increments," *Adv. in Neural Inf. Process. Syst.*, 2008.

[62] H. j. Seo, and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vis.*, vol. 9, no. 12, 2009.

[63] Z. Liu, H. Yan, L. Shen, Y. Wang, and Z. Zhang, "A motion attention model based rate control algorithm for H.264/AVC," *2009 Eighth IEEE/ACIS Int. Conf. on Comput. and Information Science*, 2009, pp. 568–573, 2009.

[64] C. Guo, and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," in *IEEE Trans. on Image Process.*, vol. 19, no. 1, pp. 185–198, 2010.

[65] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," *IEEE Conf. on Comput. Vis. and Pattern Recognit.*, pp. 1147–1154, 2013.

[66] K. Muthuswamy, and D. Rajan, "Salient motion detection in compressed domain," in *IEEE Signal Process. Letters*, vol. 20, no. 10, pp. 996–999, 2013.

[67] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," in *IEEE Trans. on Image Process.*, vol. 23, no. 9, pp. 3910–3921, 2014.

[68] S. H. Khatoonabadi, N. Vasconcelos, I. V. Bajić, and Yufeng Shan, "How many bits does it take for a stimulus to be salient?," *IEEE Conf. on Comput. Vis. and Pattern Recognit.*, pp. 5501–5510, 2015.

[69] V. Leboran, A. Garcia-Diaz, X. R. Fdez-Vidal, and X. M. Pardo, "Dynamic whitening saliency," *IEEE Trans. on Pattern Analysis and Mach. Intelligence*, vol. 39, no. 5, pp. 893–907, 2017.

[70] M. Xu, L. Jiang, X. Sun, Z. Ye, and Z. Wang, "Learning to detect video saliency with HEVC features," in *IEEE Trans. on Image Process.*, vol. 26, no. 1, pp. 369–385, Jan. 2017.

[71] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *IEEE Conf. on Comput. Vis. and Pattern Recognit.*, pp. 4510–4520, 2018.

[72] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," *European Conf. on Comput. Vis.*, pp. 305–321, 2018.

[73] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, and A. Zisserman, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[74] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical Image segmentation," *Medical Image Computing and Comput.-Assisted Intervention*, 2015.

[75] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Int. Conf. on Learning Representations*, 2020.

[76] Z. Liu, Y. Lin, Y. Cao, H. hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *IEEE/CVF Int. Conf. on Comput. Vis.*, Montreal, QC, Canada, pp. 9992–10002, 2021.

[77] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision Transformer," *IEEE/CVF Int. Conf. on Comput. Vis.*, Montreal, QC, Canada, pp. 6816–6826, 2021.

[78] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, pp. 3202–3211, 2022.

[79] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "UniFormer: Unifying Convolution and Self-Attention for Visual Recognition," in *IEEE Trans. on Pattern Analysis and Mach. Intelligence*, vol. 45, no. 10, pp. 12581–12600, Oct. 2023,

[80] Z. Teed, and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," *Proc. Eur. Conf. Comput. Vis.*, pp. 402–419, 2020.

[81] S. Baker, S. Roth, D. Scharstein, M. J. Black, J. P. Lewis, and R. Szeliski, "A Database and Evaluation Methodology for Optical Flow," *IEEE 11th Int. Conf. on Comput. Vis.*, Rio de Janeiro, Brazil, pp. 1–8, 2007.

[82] S. Mathe, and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Trans. on Pattern Analysis and Mach. Intelligence*, vol. 37, no. 7, pp.1408–1424, 2015.

[83] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. "Pytorch: An imperative style, high-performance deep learning library." *Adv. in Neural Inf. Process. Syst.*, no.721, pp. 8026–8037, 2019.

[84] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[85] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," in *IEEE Trans. on Pattern Analysis and Mach. Intelligence*, vol. 41, no. 3, pp. 740–757, 1 March 2019.

REITA IKENOYA received his B.E. from the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shizuoka University, Hamamatsu, Japan in 2022. He is currently enrolled in the Department of Electrical and Electronic Engineering, Graduate School of Integrated Science and Technology, Shizuoka university. His research interests include deep learning, artificial intelligent, and computer vision.

TOMONORI TASHIRO received his B.E., M.E. and D.E. degrees from Utsunomiya University in Tochigi, Japan in 2009, 2011 and 2014, and obtained his Ph.D. degree from University of Eastern Finland in Joensuu, Finland in 2016, respectively. He was a post-doctoral researcher at Utsunomiya University from 2014 to 2016, and that at Yamagata University from 2016 to 2021, and had been a senior researcher in Industrial Research Institute of Shizuoka Prefecture from 2022 to 2023. He is currently an assistant professor in Department of Electrical and Electronic Engineering at Shizuoka University since 2024. His research interests include cognitive, vision and color science.

GOSUKE OHASHI (Member, IEEE) received his B.E., M.E. and D.E. degrees from Keio University in Yokohama, Japan in 1992, 1994 and 1997, respectively. He had been an assistant Professor since 1997 and he is currently a Professor in Department of Electrical and Electronic Engineering at Shizuoka University. He was a visiting researcher at University of California, Santa Barbara from 2003 to 2004. His research interests include image processing, computational vision, and visual perception.