IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 14, 2021, accepted November 29, 2021, date of publication December 3, 2021,
date of current version December 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3132453

A Reinforcement Learning Based Adaptive ROI Generation for Video Object Segmentation

USMAN AHMAD USMANI[1], JUNZO WATADA[2], JAFREEZAL JAAFAR[1], (Senior Member, IEEE), IZZATDIN ABDUL AZIZ[1], AND ARUNAVA ROY[3]

[1]Department of Computer and Information Science, Faculty of Science and IT, Universiti Teknologi PETRONAS (UTP), Seri Iskandar, Perak 32610, Malaysia
[2]Production and Systems, Graduate School of Information, Waseda University, Kitakyushu 808-0135, Japan
[3]Department of Computer Science, School of Information Technology, Monash University Malaysia, Subang Jaya, Selangor 47500, Malaysia

Corresponding author: Usman Ahmad Usmani (usman_19001067@utp.edu.my)

This work was supported by the Yayasan UTP Prestigious Scholarship (YUTP) under Universiti Teknologi PETRONAS (UTP) with Cost Center under Grant 015LC0-281.

ABSTRACT Video object segmentation's primary goal is to automatically extract the principal object(s) in the foreground from the background in videos. The primary focus of the current deep learning-based models is to learn the discriminative representations in the foreground over motion and appearance in small-term temporal segments. In the video segmentation process, it is difficult to handle various challenges such as deformation, scale variation, motion blur, and occlusion. Furthermore, relocating the segmentation target in the next frame is difficult if it is lost in the current frame during the segmentation process. This work aims at solving the zero-shot video object segmentation issue in a holistic fashion. We take advantage of the inherent correlations between the video frames by incorporating a global co-attention mechanism to overcome the limitations. We propose a novel reinforcement learning framework that provides competent and fast stages for gathering scene context and global correlations. The agent concurrently calculates and adds the responses of co-attention in the joint feature space. To capture the different aspects of the common feature space, the agent can generate multiple co-attention versions. Our framework is trained using pairs (or groups) of video frames, which adds to the training content, thus increasing the learning capacity. Our approach encodes the important information during the segmentation phase by a simultaneous process of various reference frames that are subsequently utilized to predict the persistent and conspicuous objects in the foreground. The proposed method has been validated using four commonly used video entity segmentation datasets: SegTrack V2, DAVIS 2016, CdNet 2014, and the Youtube-Object dataset. On the DAVIS 2016, the results reveal that the proposed results boost the state-of-the-art techniques on the F1 Measure by 4%, SegTrack V2 by a Jaccard Index of 12.03%, and Youtube Object by a Jaccard Index of 13.11%. Meanwhile, our algorithm improves the accuracy by 8%, F1 Measure by 12.25 %, and precision by 14% on the CdNet 2014, thus ranking higher than the current state-of-the-art methods.

INDEX TERMS Model adaptation, object detection, object tracking, reinforcement learning, video object segmentation.

LIST OF ABBREVIATIONS

Ea	Embedding Module Features for Frame A.
Eb	Embedding Module Features for Frame B.
Xa	Optimal Features.
Af	Affinity Matrix.
Za	Co-Attention Summary.
fg	Gating Function.
π	Segmentation Policy.
r_t	Reward Function.

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

I. INTRODUCTION

Video object segmentation (VOS) is a technique for automatically distinguishing the objects(s) in the foreground from the background in videos. Zero Shot Video Object Segmentation (ZVOS) is very helpful for both application and research since it does not need to interact manually during the assumption phase. Inspite of the common challenges in video processing (e.g., occlusion, object deformation, and backdrop clutter, etc.), ZVOS is confronted with a new challenge: how the primary objects can be correctly distinguished from the complex background when there is no prior object present. Two qualities are required for primary video object

recognition. The objects in ZVOS should be recognized in a single frame (locally prominent) and must appear throughout the video sequence (globally consistent). Although the primary objects at the macro-level are highly correlated (the whole video), the camera movements, articulated body motions, out-of-view movements, occlusions, and ambient changes often create discontinuities at the micro-level (the individual frames) (shorter video snippets). Consequently, when dealing with problems caused by micro-level changes, it's better to depend on data from other frames (such as the global consistency feature).

When we look from a global perspective at ZVOS, we reduce the ambiguity locally and identify the primary objects. Even though it was the inspiration for most conventional heuristic models of video segmentation [1], [2] it is not preferred by the current Deep Learning (DL) based techniques. The best performing deep models of ZVOS currently focus mainly on the distinguishing feature of intra-frame. The important objects in motion or appearance ignore the global occurrence consistency across multiple frames. The optical fluxes are typically calculated across a few frames consecutively [3]–[5], [66] and are limited to the narrow temporal receptive window. Even though recurrent neural networks (RNNs) [8], [9] were created to retain data from the previous frames, the processing of this sequential method is not successful in exploring the intricate relationships between distant frames effectively.

Most works in the Machine Learning (ML) domain solve the VOS problem at the pixel stage, using Fully Convolution Networks (FCN) to conduct dense pixel wise classification for each image. Some researches focused on the coherent classification of objects based on object proposals. Despite extensive research in both the domains of image and videos, several new techniques have a lot of drawbacks. The supervised learning guidance serves as a kind of teacher in models to handle individual structures or group decisions sequentially. Supervised learning is a kind of ML algorithm that trains computers to predict output using labeled training data. As demonstrated by the labeled data, a portion of the input data is already labeled with the desired result.

The training data given as an input to the machines serve as a supervisor in supervised learning, training the machines to predict the output correctly. Abdulhussain *et al.* [10] proposed a temporal video segmentation (TVS) approach for reliably recognizing various types of video transitions with low computation cost and high recall values. The orthogonal moments are used as features in detection of transitions in the proposed technique. To improve the accuracy and speed of the TVS technique, embedded orthogonal polynomial algorithms and fast block processing are used to extract features. Li *et al.* [11] proposed an Attention-Guided Network (AGNet) for adaptively strengthening the inter- frame and intra-frame features for more accurate segmentation predictions. They added a spatial attention module (SAM) to an adjacent attention module (AAM) to a dilated Fully Convolutional Network (FCN) to imitate the feature correlations in spatial and temporal dimensions.

Nakamura *et al.* [12] proposed a semi-supervised strategy where they assumed a set of poorly labeled videos with sparsely marked frames. The frames are supplied as input, with the annotated frames being utilized to train a feature extractor. The proposed method works by dividing the input videos into small chunks known as primitive segments of set length, which are then grouped using the visual characteristics collected by the aforementioned feature extractor. Wang *et al.* [13] introduced Noisy-LSTM, a novel model for capturing the temporal coherence in video frames that can be trained using ConvLSTMs from start to finish, as well as a simple but successful training method that substitutes a frame in a video sequence with noises.

Chakrobarty and Thounaojam [14] proposed a novel shot boundary recognition method based on color and gradient information. Luminance distortion and gradient similarity are used to calculate the structural and contrast changes of each frame. By using an adaptive technique to detect the probable transitions between videos using an adaptive threshold, the proposed approach accounts for the impacts of changes in brightness and contrast-structure. Also, Chakrobarty *et al.* [15] proposed a video segmentation method based on the mean luminance patterns and CIEDE2000 color-difference and mean luminance patterns. CIEDE2000 color-difference uses the lab color space, which is efficient and trustworthy. The main benefit of the lab color space model is that it can accurately recreate all of the available colors as viewed by the human eye. The highlights and limitations of other TVS algorithms are mentioned in Table 1.

VOS is formalized as a conditional decision-making process to tackle this problem. Two RL agent are employed to calculate the attention summaries between the two feature embeddings. If there are groups of frames, the agent calculates the enhanced feature based on pair-wise co-attention between the original frame and the correlation information from the other N frames. The VOS model uses this information for the update process and is further fed into the segmentation network to find accurate segmentation masks. Figure 1 shows the segmentation of three frames in the DAVIS 2016 video dataset by our Reinforcement Learning (RL) algorithm. As seen in the results in the 2^{nd} row and 4^{th} row, the boundary of our segmentation mask results is clearly visible.

To select the optimal segmentation mask for the frames, the features are fed into an RL model. Then the RL model determines the best action and chooses the most suitable mask for the current frame. Thus, an accurate segmentation result is obtained by the segmentation model. The RL agent learns to capture the complex correlations between a group or pair of frames from the same video. This is achieved via the use of a gated co-attention differentiable approach that enables the network to pay more attention to informative correlated regions while generating more discriminative foreground features. Our RL model can give more accurate

TABLE 1. Temporary video segmentation algorithm highlights and limitations.

S.No.	Algorithm Name	Highlights	Limitations
1.	Discrete Orthogonal Moments [10]	It utilizes fast block processing and embedded orthogonal polynomial algorithms to extract the features.	Soft transitions needs to be detected to guarantee the use of algorithm for video object detection.
2.	Attention-Guided Network [11]	Appends a spatial (SAM) and adjacent attention module (AAM) on the top of dilated FCN, which models the feature correlations in spatial and temporal dimensions.	Selection of more powerful backbones needs to be considered for replacing the upsample operation with a complex decoder for optimizing the final boundary results.
3.	Noisy-LSTM [13]	Temporal coherence in video frames are leveraged by using convolutional LSTMs which replaces a given video frame sequence with noises.	Technique should further explore the way to inject noises in model training and identify the types of noises.
4.	Hierarchical tree [12]	Proposed method assumed a set of weakly-labeled videos and hierarchical tree of the category labels performed recursively at each tree branch.	Unknown category labels cannot be handled.
5.	Motion Guided Attention [20]	Technique transfers information inherent in image-based instance embedding networks.	Incorrect foreground seeds are often discovered on static objects when errors occur in "objectness only" mode.
6.	SBD-Duo [14]	Proposes a novel shot boundary detection technique using colour and gradient information.	Should extend the proposed system to address non-uniform illumination effect and eliminate the effect of camera obstruction.
7.	Visual Colour Information [15]	Technique utilizes mean luminace pattern and CIEDE2000 colour-difference.	Should extend the proposed method for detecting wipes transition using the mean luminance pattern.
8.	Hybrid dual tree complex wavelet transform [21]	Technique uses Walsh Hadamard transform (DTCWT-WHT) with Hybrid Dual-Tree Complex Wavelet Transform.	The structure suffers from a basic limitation when the authors examine the overall frequency response of each channel.
9.	Color histogram [22]	Local descriptors and the image color feature are combined in a kind of motion area extraction algorithm.	SURF matching is not performed for all adjacent frames of each candidate segment frame by frame.
10.	Deep CNN [23]	Three stages are respectively used for abrupt detection, candidate boundary detection and gradual transition detection.	Variety of constraints present in the method for filtering non-boundaries and the model processing speed needs to be considered.

results for a testing frame when several reference frames and correlations between the testing frame are utilized. When the data is used only from a single testing frame, the results are poor. Another advantage of our RL model is that it can be used to supplement the training data. It enables a high number of random frame pairings to be utilized inside a single video. The proposed model also removes the necessity for time-consuming and computationally expensive optical flow calculations, because of the specified connections between video frames. Finally, our RL model offers a single framework for collecting rich contextual data from video sequences from start to finish. To summarize, this paper provides four significant contributions:

- We propose a single, end-to-end RL framework where two RL agents are employed to calculate the rich features between the video frames using the differentiable co-attention mechanism. This helps in recognizing the primary video foreground objects.
- The correlations are learned by the pair-wise co-attention mechanism between the frame pairs, which is further fed into the segmentation network to obtain the optimal segmentation mask. We adopt the Deep Deterministic Policy Gradient (DDPG) algorithm to primarily train the agent in producing the correct object segmentation masks.
- The RL agent calculates the correlations among the video frames using the group co-attention mechanism, resulting in a significant motion object pattern modeling framework.
- According to the final quantitative results, the suggested methods outperform state-of-the-art methods in the *F1* measure (F_m) on the DAVIS 2016 dataset [16] by 2%, SegTrack *V2* [17] by 12.03 %, and the Youtube-Object dataset [18] by 13.11% for the *F1* measure. Meanwhile, our algorithm outperforms the current state-of-the-art methods evaluated on the CdNET night segmentation dataset [19], with an Accuracy (*Acc*) of 87.99%, Precision (*Pre*) of 94.01%, and F_m of 92.51%.

The remainder of the paper is organized as follows. Section II describes the current state-of-the-art methods in video

segmentation. Section III gives our proposed RL technique for segmenting the video objects. It involves modeling RL actions, states, and rewards to enhance the performance of the VOS by using a pair-wise co-attention mechanism. Section IV describes in detail the Results and Discussion. We finally conclude in Section V and give our future directions.

II. RELATED WORK

The VOS problem is addressed in a zero-shot unsupervised or one-shot (semi-supervised) setting, depending on the degree of supervision given during the test time. In this research, the focus is given to the problem of ZVOS, which performs the extraction of the primary object(s) and does not require any intervention of humans in test time.

ZVOS rose from the long-studied issue of automated VOS in the field of computer vision. Automated video segmentation algorithms usually emphasize spatiotemporally

FIGURE 1. Segmentation of 3 frames in the DAVIS 2016 video dataset by our RL algorithm. As it can be seen in the results in the 2^{nd} row and 4^{th} row, our segmentation mask results are very clear.

connected groupings of video pixels and compact (consistent motion and appearance). Motion analysis is one of the early solutions [21], [22], and is based on assessing background-induced motion patterns and geometry constraints. A wide class of models based on trajectory are used for exploiting the long-term motion information. The popular techniques include super-voxels [24], temporal superpixel [25], and hierarchical segmentation [26]. The researchers shifted their focus towards video object pattern modeling after low-level video over-segmentation. The signals related to objects like object proposals [20], [23], [27] are used for the saliency information [1], [28]–[30], and inference of primary video objects are utilized. The example works above described made substantial improvements in VOS. Still, hand-crafted features' limited representation capacity failed when the heuristic assumptions were not applied.

Many methods [31]–[33] using DL features have recently started to tackle the ZVOS problem, and are inspired by the success of DL. The improvement in performance of these models are large weight fully connected topology networks [31], [33], but are limited by the lack of learning capabilities end-to-end. The research later focused on ZVOS models that were constructed entirely on convolutional neural networks. To differentiate between independent object and camera motion, Tokmakov *et al.* [9] proposed using a learnable motion pattern network. For recognizing the background [7], Li *et al.* [6] used static images to train an instance embedding network and then identified the background [7] by incorporating motion-based bilateral networks.

FCN are another popular method for combining appearance data and motion for inference of objects [3]–[5], [20], [34], [35]. Other research looked at ZVOS via exploring robust network topologies [9], [20], [36] teacher-student learning paradigm [37]. Wang *et al.* [38] created an attention-guided ZVOS model after viewing a dynamic task, and demonstrating the substantial relationship between moving object patterns and human attention. These deep ZVOS

models often provide excellent results, showing the usage of neural networks advantages for this task. On the other hand, they focus exclusively on the short-term temporal information and ZVOS sequential nature, using the beneficial, cross-frame correlation within videos and failing to take a global view. During the ZVOS testing phase, target object(s) mask(s) are often provided in a few frames or the first frame and are sent to the future frames [39]–[42] automatically. Several prior approaches have been proposed, including super-trajectories [43], object recommendations [44], graphical models [32], and so on. The results of DL algorithms are promising and have dominated the field.

Test-time supervision is done for the models online, including methods based on learning, which performs the models fine-tuning online. Others include frame-by-frame mask propagation [41], [45], [46] and propagation-based, which rely on previous frame segments and function on a frame-by-frame basis. Matching-based is another kind of common stream, in which each frame is split based on its matching connection/correspondence to the preceding frame [47]. Although many matching-based OVOS models use a Siamese network architecture, our changes are substantial. Aside from task parameters, our RL method captures the global and rich correspondence by training the Siamese network between groups or video frame pairs. The main aim is to utilize the cross-frame correlations to assist automatic segmentation and primary object recognition. RL is a machine learning training method based on the reward of the desired behaviors and punishing the undesired ones. At the same time, the connections are captured between the matching-based OVOS models between the first and subsequent frames.

Deep neural networks [48], [49] have been widely investigated for differentiable attention [50], [51], inspired by human vision. The networks can use neuronal attention with end-to-end training and focus on a subset of informative inputs. Beginning with the neural machine translation [52] and moving to a broad variety of NLP-related activities [48], a continual development of attention mechanisms has been witnessed in the field of natural language processing (NLP). Later, a wide range of computer vision applications utilized neural attentions, including object identification [53], image captioning [54], video processing [55], visual recognition [56], and visual dialogue [57], to name a few. It has been shown that differentiable attention can capture correlations/dependencies between the input components.

In particular, Chen *et al.* [58] proposed the use of channel and spatial- wise attention for choosing an image area dynamically while reducing the redundancy of the feature channel. Self- attention beats LSTM and traditional RNN in the sequence- to-sequence challenge, according to Vaswani *et al.* [48] computed the solution at a place by accessing all locations. A non-local operation was proposed by Wang *et al.* [59] that may be thought of as a broader kind of self-attention in a self-supervised environment. Sun *et al.* [60] trained a mixed visual, linguistic model using self-attention-based BERT. Co-attention mechanisms, a kind of differentiable attention, have recently been successful in language and vision tasks [61], [62]. In this research, co-attention techniques are used for efficiently mining the underlying relationships and projecting different modalities into a single feature space. In this research, the RL agent captures coherence across different frames by utilizing a co-attention module, resulting in an elegant and unified VOS network framework that focuses on the significance of identifying video object information globally. The RL algorithm learns from experience. Actor- Critic applies to a well-known RL model that inherits several previous RL constructs focused on values and policies, such as policy gradient and Deep Q-learning. In computer vision applications, RL techniques have been used to detect objects at the bounding box level in various computer vision applications. Yun *et al.* [63] used RL techniques to move the bounding box from the object's original position in the previous frame to the precise location. To put it another way, the predicted action allows the sensor to move away from its current site, and the next event is measured using the new location. Zhang *et al.* [64] proposed a new RL-driven model that can choose deep convolutional layers based on the complexity of the current image, reducing run time while maintaining precision.

The basic principle is that the less convolutional layers' features are used to process the simpler frames. In contrast, the fully convolutional layers' costly and invariant deep features are used to process the more complicated frames. To find the best relation between the filter hyperparameters, Dong and Yang [65] used the RL technique. Since conventional continuous deep Q learning algorithms are challenging to implement, they may help speed up the convergence phase. Just one attempt to integrate RL into the role of VOS has been made to our knowledge.

Chen *et al.* [66] created a new RL architecture that chooses the bounding object box and the background box. Context and object boxes are distinguished in the exploration, resulting in distinct segmentation masks for an analogous segmentation model. As a result, using the RL technique to choose the best object context box pair for the best segmentation result is usually appropriate. Unlike Sun *et al.* [60], who used RL to determine the size of the quest area fed into the segmentation network, our agent is capable of generating multiple co-attention versions for capturing the different aspects of the learned joint feature space

III. METHODOLOGY

Our RL method can recognize the primary video objects which appear throughout the video sequence and are distinguishable in each frame. Our RL method formulates the ZVOS problem as a co-attention approach, and a novel co- attention Siamese Network is constructed to represent it from a global perspective. Our method learns to capture the complex correlations between a pair (or group) of frames from the same video during training. This is achieved by using a gated, differentiable co-attention mechanism, enabling the network to focus on informative, correlated regions while also

producing more discriminative foreground features. From a global perspective, our RL method provides more accurate results for a testing frame, i.e., it considers the correlations between the testing frame and many reference frames. To be more explicit, we first explore the co-attention between the paired frames. Consequently, during the testing phase, the pair-wise co-attention features of many inference frames are concatenated to form a global representation. We describe a group co-attention module directly constructed across several frames. Based on the pair-wise co-attention module, we can capture global information more naturally and elegantly. Additionally, our RL model uses a large number of arbitrary frames which helps it augment the training data.

The proposed model avoids the need for computationally expensive and time-consuming optical flow calculations since the video frame interactions are fully specified. Finally, our model offers an end-to-end trainable framework for collecting rich contextual information from video sequences from start to finish. We demonstrate that our co-attention strategy helps improve performance and focuses on the importance of global information and its usability for ZVOS. Our model can catch the rich relations between the video frames due to the differentiable co-attention mechanism, which is essential for distinguishing the key video foreground objects. Our method also employs group co-attention for extracting high and rich order correlations between video frames, resulting in a more powerful moving object pattern modeling framework.

The experimental results in the section IV infer that our method can suppress similar target distractions and capture the common objects even when no annotation is supplied during the segmentation task. Our approach can handle the sequential learning of data and can readily be applied to various video analysis applications, such as optical flow predictions and video saliency detection. We use a co-attention strategy in our RL framework. The RL agent encodes the correlations between video frames directly. This enables our model to focus its attention on regions that are often coherent, helping in the identification of foreground objects and providing acceptable segmentation results. Our method is able to detect the irregular objects in the standard datasets. The coherent region are extracted finely and helps in extracting good quality segmentation masks. We give an overview of our RL method in Section A. In Section B, the agent action is provided to calculate the pair-wise co-attention summaries for the features obtained from the feature embedding module. In Section C, the State and Reward is given for the RL agent, while the training in Actor-Critic Framework is given in Section D. In Section IV, we give the analysis of Results and Discussion while Section V concludes our work.

A. OVERVIEW

The main objective of the proposed work is to utilize RL for zero-shot VOS. Unlike the current DL-based methods, which primarily focus on learning the foreground representations, our approach uses RL to find the correlation between the video frames. This helps our model attend the frequently coherent regions and discover the foreground object(s), thus producing good segmentation results. The RL agent evaluates the correlation learning between any frame pairs from the same video. It also helps assess the groupwise co-attention mechanism, which addresses the high order relationship amongst a group of video frames. The framework of our RL model is shown in Figure 2.

$\mathbf{E}_a$ and $\mathbf{E}_b$'s features are fed into the RL model to obtain the optimal $\mathbf{X}_a$ and $\mathbf{X}_b$ for each frame to address the problem. Two RL models are built to choose the most suitable $\mathbf{X}_a$ and $\mathbf{X}_b$ for each frame and choose the appropriate group co-attention. A state $s \in S$, a co-attention computation action $a_x \in A_x$ that helps in determining the value of $\mathbf{X}_a$, and a computation action $a + y \in A_y$ that determines the value of $\mathbf{X}_b$, a state transition function that is denoted as $s' = T(s, A_x, A_y)$, and a reward function g denoted by $g(s, A_x, A_y)$.

The provided frames Fa and Fb are fed into a feature embedding module to get $\mathbf{E}_a$ and $\mathbf{E}_b$'s features. The two RL agents then calculate the pair-wise co-attention module and attention summaries that encapsulate the correlations between $\mathbf{E}_a$ and $\mathbf{E}_b$. $\mathbf{E}$ and $\mathbf{Z}$ are then concatenated and passed to a segmentation module, which produces the final segmentation predictions, i.e., the group co-attention enhanced embedding is fed into the segmentation network, the optimal segmentation masks $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are achieved. To mine the correlations between Fa and Fb in their respective feature embedding space, the co-attention mechanism [67], [68] is employed.

To begin with, the affinity matrix $\mathbf{A}_f$ between $\mathbf{E}_a$ and $\mathbf{E}_b$, is calculated in Equation 1 as follows:

$$\mathbf{A}_f = \mathbf{E}_b^T \mathbf{W} \mathbf{E} a \in \mathrm{R}^{(WH)\times(WH)} \tag{1}$$

where the weight matrix is $\mathbf{W} \in \mathrm{R}^{\mathrm{C}\times\mathrm{C}}$. Here $\mathbf{E_b} \in \mathrm{R}^{\mathrm{C}\times(\mathrm{WH})}$ and $\mathbf{E_a} \in \mathrm{R}^{\mathrm{C}\times(\mathrm{WH})}$ are represented by flattening of matrix representations. Every column $\mathbf{E_a^{(i)}}$ at position $i \in \{1,\ldots,\mathrm{WH}\}$ in $\mathbf{E}_a$ represents the C dimensions feature vector. Therefore, every entry of $\mathbf{A_f}$ provides each row $\mathbf{E_b^t}$ and each column $\mathbf{E}_a$ similarity. Because, the W weight matrix is a square matrix, it can be represented as a diagonalization of W is calculated in Equation 2 as follows:

$$\mathbf{W} = \mathbf{P^{-1}DP} \tag{2}$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is an invertible matrix. Then, it can be rewritten as follows in Equation 3:

$$\mathbf{A_f} = \mathbf{E_b^T P^{-1} D P E_a} \tag{3}$$

Before computing the distance between any of their locations, the characteristic of each frame is linearly modified in Equation 3. The projection matrix $\mathbf{P}$ becomes an orthogonal matrix when the weight matrix also symmetric: $\mathbf{P} > \mathbf{P} = \mathrm{I}$, where $\mathbf{I}$ represents the identity matrix $\mathbf{C}\times\mathbf{C}$. To compute symmetric co-attention we use the Equation below as follows in Equation 4:

$$\begin{aligned} \mathbf{A_f} &= \mathbf{E_b^T P^T D P E_a} \\ \mathbf{A_f} &= (\mathbf{P E_b})^{\mathbf{T}} \mathbf{D P E_a} \end{aligned} \tag{4}$$

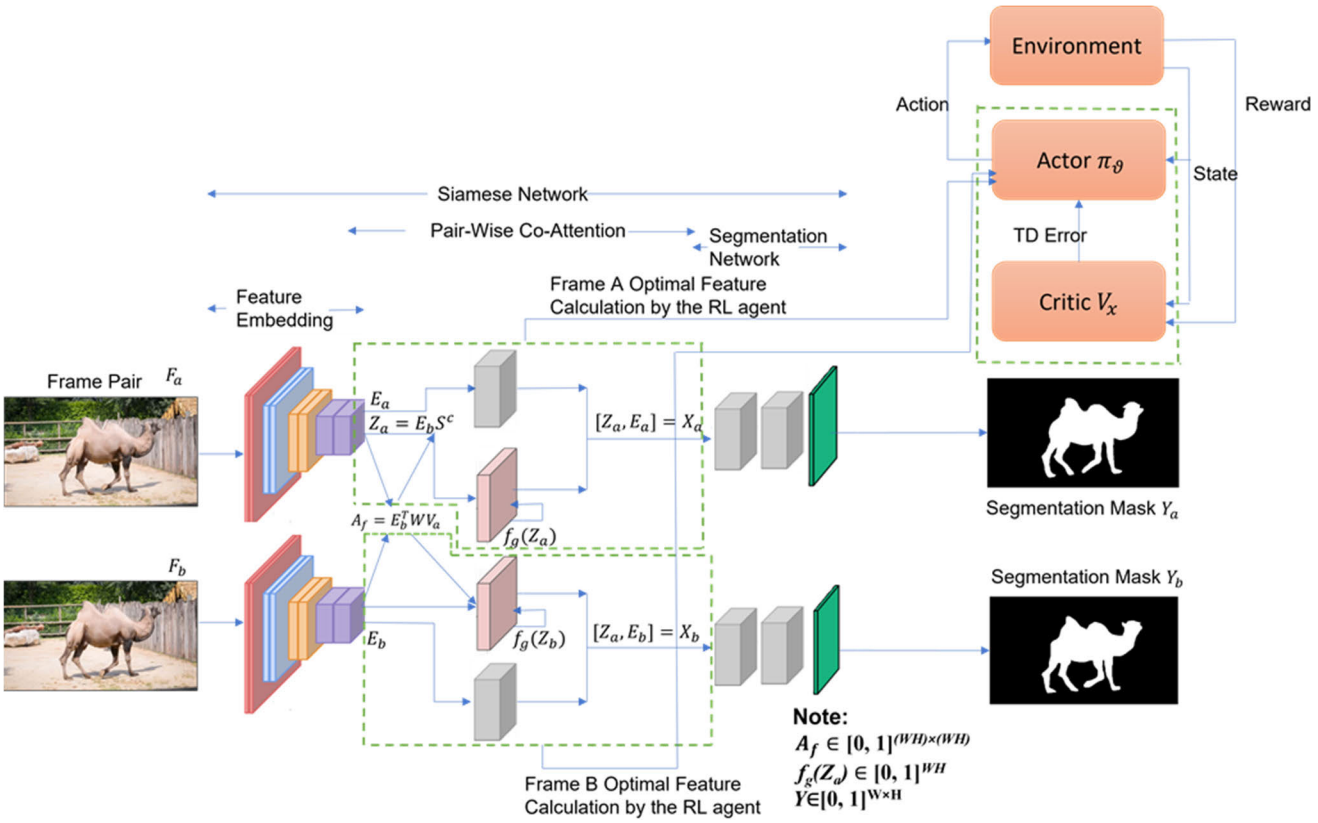

FIGURE 2. Overview of our RL framework based pair-wise co-attention in the phase of training. The frame pairs F_a and F_b is fed as an input to the feature embedding module for obtaining the two features E_a and E_b. Then, our RL agent calculates the co-attention summaries and finds the correlations between the two frame embedded features E_a and E_b. We adopt two actor-critic type model pairs to calculate the attention summaries Z_a and Z_b. Two roles are performed by the "actor-critic" framework which includes an "actor" role for generating an action and a "critic" role for measuring how good the action is. Then a gating function is used for allocation of co-attention confidence to each attention summary. Finally, the concatenation of the attention summary Z and the features E obtained from the embedded feature module is done and then handed over to the segmentation module for producing accurate segmentation masks.

We project the features $\mathbf{E}_a$ and $\mathbf{E}_b$ into a common orthogonal space while preserving their norm, according to Equation 4. This characteristic has been shown to assist in the removing correlation across many channels (i.e., the C dimension) [69] as well as the enhancement of the network's generalization ability [70]. There is a degree of co-attention on each channel. Furthermore, the projection matrix $\mathbf{P}$ can be reduced to an $\mathbf{I}$ identity matrix and the weight matrix W can be reduced to the diagonal matrix. $\mathbf{W}$ (i.e. $\mathbf{D}$) can be diagonalized into two diagonal matrices, $\mathbf{D}_a$ and $\mathbf{D}_b$. As a result, Equation 4 can be re-written as channel-wise co-attention in Equation 5 as follows:

$$\mathbf{A_f} = \mathbf{E_b^T I^{-1} D I E_a} = \mathbf{E_b^T D_a^T T D_b E_a} = \mathbf{(D_a E_b)^T} > \mathbf{D_b E_a} \tag{5}$$

This method can be compared to applying a channel-wise weight to $\mathbf{E}_a$ and $\mathbf{E}_b$ before computing the similarity. This lowers channel redundancy in a similar way to the Squeeze and-Excite (SE) [71] method.

B. AGENT ACTION

The architecture of the proposed method comprises two RL models that choose the best group co-attention enhanced embedding. To select the pair-wise co-attention $\mathbf{X}_a$, the action set A_x is utilized and for selecting $\mathbf{X}_b$ the action set A_y is utilized. $\mathbf{A}_f$ i$^{\text{th}}$ column contains the WH-length vector. This vector depicts the relationship between the i^{th} feature in $\mathbf{E}_b$ and each feature (1,…., WH) in $\mathbf{E}_a$. Using a softmax function, the column-and row-wise normalized symmetric co-attention is as follows in Equation 6:

$$\begin{aligned} \mathbf{A_c^f} &= softmax(\mathbf{A_f}) \in [0, 1]^{(WH)\times(WH)} \\ \mathbf{A_r^f} &= softmax(\mathbf{A_c^f}) \in [0, 1]^{(WH)\times(WH)} \end{aligned} \tag{6}$$

Next, the attention summaries for the feature embedding $\mathbf{E}_a$ w.r.t. $\mathbf{E}_b$ can be computed as follows in Equation 7:

$$\begin{aligned} \mathbf{Z_a} &= \mathbf{E_b S^c} = [\mathbf{Z_a^{(1)} Z_a^{(2)}} \ldots\ldots \mathbf{Z_a^{(i)}} \ldots\ldots \mathbf{Z_a^{(WH)}}] \\ \mathbf{Z_a^{(i)}} &= \mathbf{E_b} \circledast \mathbf{A_f^{c(i)}} = \sum\nolimits_{j=1}^{WH} \mathbf{E_b^{(j)}} \mathbf{A_{fc}^{ij}} \in R^C, \end{aligned} \tag{7}$$

where $\mathbf{Z_a^{(i)}}$ denotes the $\mathbf{Z}_a i^{th}$ column, $\circledast$ indicates the matrix times vector, $\mathbf{A_f^{c(i)}}$ is the i^{th} column of $\mathbf{A}_c^f$, $\mathbf{E_j^b}$ indicates the j^{th} column of $\mathbf{E}^{(j)}$ and $\mathbf{A_{fc}^{ij}}$ is the j^{th} element in $\mathbf{A}_f^{c(i)}$.The corresponding co-attention enhanced feature is calculated as: $\mathbf{Z}_b = E_a \mathbf{A}_r^f$.

Given the subtle variations in appearance across input pairs, occlusions, and background noise, it's better to calculate the information from different input frames rather than treating all co-attention data identically. To accomplish this, the network uses a self-gate technique for assigning every attention summary with a co-attention confidence. The gate inscription reads as follows in Equation 8:

$$f_g(\mathbf{Z_a}) = \sigma(w_f \mathbf{Z_a} + b_f) \in [0, 1]^{WH},$$
$$f_g(\mathbf{Z_b}) = \sigma(w_f Z_b + b_f) \in [0, 1]^{WH}, \quad (8)$$

where σ is the logistic sigmoid activation function, b_f and w_f is the bias and convolution kernel, respectively. The f_g gate regulates the amount of data from the frame of reference and can be learned automatically. The updation of attention summaries is done once the gate confidences have been computed and are described in the Equation 9 as follows:

$$\mathbf{Z_a} = \mathbf{Z_a} \star f_g(\mathbf{Z_a}), \quad \mathbf{Z_b} = \mathbf{Z_b} \star f_g(\mathbf{Z_b}) \quad (9)$$

where the channel-wise Hadamard product is represented by '$\star$'. As a consequence of these actions, a co-attention structure with gates is formed. The resulting co-attention representation $\mathbf{Z}$ is then concatenated with the original feature $\mathbf{E}$ by the two RL agents to calculate the values of $\mathbf{X}_a$ and $\mathbf{X}_b$ respectively in Equation 10 as follows:

$$\mathbf{X_a} = [\mathbf{Z_a}, \mathbf{E_a}] \in \mathrm{R}^{W \times H \times 2C}, \mathbf{X_b} = [\mathbf{Z_b}, \mathbf{E_b}] \in R^{W \times H \times 2C} \quad (10)$$

where '[ ]' denotes the concatenation operation. The pairwise co-attention enhanced feature X calculated by the RL agent is fed into a segmentation network to produce the final output $\mathbf{Y} \in [0, 1]^{W \times H}$.

With respect to the reference group embedding [$\mathbf{E}_1$,.... $\mathbf{E}_{n-1}$, $\mathbf{E}_{n+1}$,..., $\mathbf{E}_N$+**1**], the Equation below is the inference feature embedding $\mathbf{E}_n$ based on group co-attention summary in Equation 11 as follows:

$$\mathbf{Z_n} = [\mathbf{E_1}, ..., \mathbf{E_{n-1}}, \mathbf{E_{n+1}}, ... \mathbf{E_{N+1}}]\mathbf{A^c_{fn}} \in R^{C \times WH} \quad (11)$$

Each column in the Equation 11 above represents a linear combination of reference frame embeddings, resulting in $\mathbf{Z}_n$. $\mathbf{Z}_n$ now contains the whole information of the reference group. By using Equation 8 and Equation 9, the gated co-attention are calculated, and the concatenation of the co-attention summary with the original features obtained from the feature embedding is done by the RL agent as follows in Equation 12:

$$\mathbf{X_n} = [\mathbf{Z_n}, \mathbf{E_n}] \in \mathbf{R}^{W \times H \times 2C} \quad (12)$$

where E_n is the enhanced group co-attention embedding that is calculated by considering the feature of the original frame and the correlation information of the preceding N frames' correlation information. By focusing on the group as a whole, we can improve the features $\{\mathbf{X}_n\}_{n=1}^{N+1}$ for all the frames $\{\boldsymbol{F}_n\}_{n=1}^{N+1}$.

C. STATE AND REWARD

Since this method consists of two RL-based models, each model employs a separate collection of states. The RL model's input is actually in states. The feature map is produced by states obtained after extracting the features from the feature embedding module. The attention summaries are computed using the group-wise co-attention modules that encode the correlations between$\mathbf{E}_a$ and $\mathbf{E}_b$. Finally, the concatenation of $\mathbf{E}$ and $\mathbf{Z}$ is done and given over to the segmentation module to produce the final segmentation process predictions.

The features $\mathbf{E}_a$ and $\mathbf{E}_b$ are obtained from the feature embedding module in the provided input frames $\mathbf{F}_a$ and $\mathbf{F}_b$. Following that, the attention summaries for the feature embedding $\mathbf{E}_a$ with respect to $\mathbf{E}_b$ are computed. Co-attention confidence is assigned to each attention summary by using a self-gate technique. Concatenating the initial features $\mathbf{E}$ and $\mathbf{Z}$ yields the final co-attention representation. For Frame F_a, the state f_a is defined as follows:

$$state_{f_a} = feature\{\mathbf{E_a}\} + feature\{\mathbf{E_b}, \mathbf{A^f_c}\} + feature\{f_g(\mathbf{Z_a}), \mathbf{E_a}\} \quad (13)$$

In the case for the second frame $\mathrm{F_b}$, $\mathbf{E_a}$ is used for calculating the affinity matrix and the state is given by the following Equation below as:

$$state_{\mathbf{f_b}} = feature\{\mathbf{E_b}\} + feature\{\mathbf{E_a}, \mathbf{A^f_r}\} + feature\{f_g(\mathbf{Z_b}), \mathbf{E_b}\} \quad (14)$$

Finally, states $state_{f_a}$ and $state_{f_b}$ will be fed into the corresponding RL model and result in the actions to choose the optimum values of group co-attention enhanced embedding. The reward function is defined as $r_t = g(s_t, a_x, a_y)$, which helps in reflecting the final segmentation result performance of each frame in the video sequence and is given by:

$$g(s_t, a_x, a_y) = \begin{cases} \boldsymbol{\alpha.1} & if\ \Delta > 0.1 \\ -\alpha.1 & if\ \Delta < 0.1 \end{cases} \quad (15)$$

where $\Delta = IoU(m_t, _{k+1}, y_t) - IoU(m_t, _k, y_t)$

This Equation 15 reflects each of the final segmentation results for every frame in the video sequence: where IOU represents the Intersection-over-Union (IOU) between the ground truth and the predicted ROI, thus indicating the accuracy of the predicted segmentation's mask.

D. TRAINING IN ACTOR-CRITIC FRAMEWORK

DDPG is an off-policy, model-free technique for continuous action learning [90]. DQN (Deterministic Quality Network) and DPG (Deterministic Policy Gradient) are combined in this model (Deep Q-Network). It is based on DPG, which is capable of operating over continuous action spaces and employs DQN's Experience Replay and slow-learning target networks. The architecture of our actor-critic framework is shown in Figure 3. For RL training, this analysis uses the "actor-critic" framework. The "actor-critic" framework

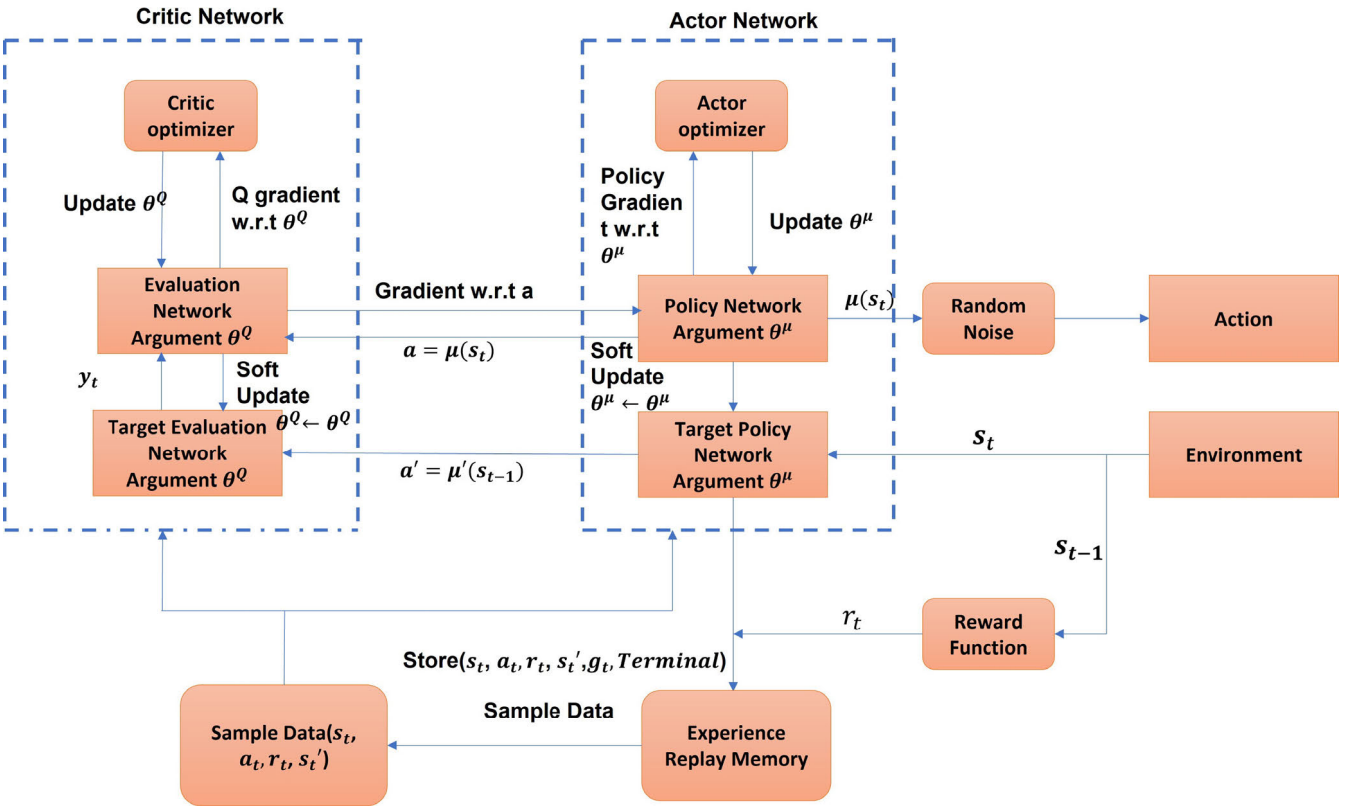

FIGURE 3. Actor-Critic Network. When the current frame *Fa* is provided, the first step is to feed the state into the "actor" network, selecting the optimal features for producing the segmentation masks. The corresponding reward *rt* will be rewarded once this phase is completed. The *rt* is calculated by the IOU that depends on the final results of the segmentation process.

consists of two parts: an "actor" who performs an action and a "critic" who evaluates the actor's performance. The proposed RL algorithm selects the optimum group coattention enhanced embedding for each frame. We selected two "actor-critic" type pairs to choose the segmentation masks for each of the frames. We take four individual RL models. When the current frame F_a is provided, the first step is to feed the state into the "actor" network, selecting the optimal improved feature to produce the segmentation mask. The corresponding reward *rt* will be rewarded once this phase is completed. The *rt* is calculated by the IOU that depends on the final results of the segmentation process. During the phase of training, the "critic" network is generally updated first, in the form of a value-based fashion which is shown as follows: The critic updates the network, which is quite similar to the average cost temporal-difference method of [72]:

$$\lambda_{k+l} = \lambda_k + jY_k(g(X_k, U_k)\lambda_k) \tag{16}$$

where

$$r_{k+1} = r_k + \gamma k(g(X_k, U_k)\lambda_k + Q\theta_k r_k(X_{k+1}, U_k + \theta_k) \tag{17}$$

In Equation 16 and Equation 17, λ_{k+1} and λ_k represent the weight of the critic model before and after the update, and γ is the learning rate of the critic model. $g(X_k, U_k)$ means the accumulated award of the state s_t the critic predicts that before the update of the model. r_k, X_k, and U_k are the reviewer's criteria, and θ_k is the actor's vector r at time k. (X_k, U_k) is the current state-action pair. X_{k+l} is the new state obtained after performing action U_k.

$$\theta_{k+1} = \theta_k \beta_k \tau(r_k)\theta_k Q(X_{k+1}, U_{k+1})\theta k \psi(X_{k+1}, U_{k+1}) \tag{18}$$

where $\psi(s, a)$ denotes the advantage function, and θ_k and θ_{k+1} in Equation 18 denotes the weight of the "actor" model after and before the transition, respectively. The policymaking function $Q(X_{k+1}, U_{k+1})$ is a network with the state s and unique action as inputs and the likelihood of the chosen action happening in the state s as output. As a result, the "actor-critic" framework avoids the shortage of policy- based and value-based approaches while training the RL models. The RL models are modified during each pass instead of waiting for the end of the process, thus significantly minimizing the training time while preserving the stability of RL training.

E. IMPLEMENTATION DETAILS

The backbone network for our RL framework is DeepLabv3 [73]. In this network, the ASPP atrous spatial pyramid pooling module and starting blocks of the five convolutional layers from [74] ResNet form the backbone network for our RL framework. The weight matrix W of the co-attention modules is built by utilizing

256 × 256 parameters of a fully connected layer. The SE-like module is used to describe the channel-wise co-attention. The 256 nodes fully connected layer is generated by the channel weights by having 256 nodes in one branch, which utilize a fully connected layer. It is then given as an input to the other embedded branch feature. Equation 8 is achieved using an 1 × 1 layer convolutional layer having a sigmoid activation function. The following parameters apply to the different co-attention variations as well as group co-attention. The 3×3 convolutional layers form the segmentation module (with the batch norm and 256 filters) and 1 × 1 layers of convolutional for final segmentation prediction (with one filter and sigmoid). Our RL model training process is split into two parts. We fine-tune the feature embedding module based on DeepLabV3 [73] is using the YouTube-Object VOS [18] and SegTrack V2 [17] image saliency datasets by using the static data.

The model is then fine-tuned using the DAVIS16 [10] training videos. At this point, two frames are chosen randomly from the same sequence and given to our RL model to use as pairs in the training process. We choose six frames at random for the training process of group co- attention from the same sequence. To conserve GPU RAM, we compress the frames in the input to $384 \times 384 \times 3$, yielding a (56; 56; 256)-d tensor as the frame's initial feature embedding V. The whole network is constructed using TensorForce, and the training is done with the learning rate schedule and SGD optimizer shown below: init $lr = 1.5 \times 10-4$. The total epochs and batch size are both set to 4 and 30. Data augmentation is beneficial to static images and video data (e.g., flipping, resizing, and cropping). All testing and analyses are performed on an Intel (R) Xeon 2vCPU@2.2GHz and an NVIDIA GeForce 1080 Ti GPUs. It takes around 3 days to complete the training. The network design has provided outstanding results for the classification of ImageNet dataset and the segmentation of the PASCAL tasks [29]. A batch of four frames, consisting of three reference frames and one inference frame is used by our RL model. A three frames batch, i.e., two reference frames with just one inference frame, is used by our RL model and is sufficient for generating promising results.

Algorithm 1 RL Based Video Object Segmentation

Input: Ground Truth of the First Frame extracted img(1)
The Length of the sequence M
The threshold T
Pretrained ResNet network
RL model to choose the pair-wise co-attention X_a
RL model to choose the pair-wise co-attention X_b
Output: Segmentation result Y_t
1: Fine-tune Segmentation Network on the frame extracted F_a and F_b.
2: Extract the features E_a and E_b from the feature embedding space
3: **for** $t = 2$ to L do
4: Obtain the two RL machine learning states using (3) and (4), respectively.
5: Feed the states into the RL Model and calculate the optimal X_a and X_b values.
6: Obtain the Segmentation mask Y_t for the two frames F_a and F_b.
7: Update the segmentation networks on frames F_t using E_a and Z_a.
8: K_t = Update(Seg_Algo, F_t)
9: $r(t) = m(t)$
10: **end for**

IV. ANALYSIS OF RESULTS AND DISCUSSION

To calculate the performance of the proposed model, the basic statistical parameters used in other literature works have been studied. Sensitivity (*Sen*) is calculated in Equation 19, as follows:

$$Sen = TP/TP + FN \tag{19}$$

It means that the number of object pixels in the image is distributed uniformly. Similarly, the parameter Specificity (Spe) determines if the pixels proportions have been correctly assigned to the image and is given in Equation 20 as follows:

$$Spe = TN/TN + FP \tag{20}$$

The rate of pixel classification referred to as Acc is determined in Equation 21, as follows:

$$Acc = TP/TN + TP + FN + FP \tag{21}$$

The spatial overlap that is present between the assigned binary mask and the segmented image is defined as the

Dice coefficient (Dice), and is measured in Equation 22 as follows:

$$Dice = 2TP/2TP + FP + FN \tag{22}$$

The Jaccard Index (J_m) is the relationship between the binary labels and the pixel values analyzed for the input image. The J_m is determined in Equation 23 as follows:

$$Jaccard\ Index = TP/TP + FN + FP \tag{23}$$

It is generally used to measure the change in the center of transformation present in the image axis. While true positive (*TP*) correctly depicts the object pixels, false positive (*FP*) incorrectly depicts non-object pixels as objects, true negative (*TN*) depicts all incorrectly labeled non-object pixels, and false negative (*FN*) represents the incorrectly identified object pixels.

The F_m is a measure of a test's accuracy in binary classification statistical analysis. It's determined by dividing the number of true positive results by the total number of positive results (including those that were erroneously recognized), and the *Rec* is the number of genuine positive findings divided by the total number of samples that could have been detected as positive. In diagnostic binary

classification, Pre is also known as positive predictive value, while *Rec* is also known as *Sen*. The F_m score is calculated using the harmonic mean of *Acc* and *Rec*. The additional weights generally score favor accuracy or memory over the other. The conventional F_m, often known as the balanced F_m, is the harmonic mean of Pre and Rec in Equation 24 as follows:

$$F_m = \frac{2}{recall^{-1} + precision^{-1}} = \frac{tp}{tp + \frac{1}{2}(fp + fn)} \quad (24)$$

The Segtrack V2 dataset [17], the DAVIS 2016 dataset [16], and the Youtube-Object dataset [18] are all used to test the proposed RL segmentation method. The DAVIS 2016 dataset contains 50 high-quality video sequences and 3455 frames that address a wide range of VOS issues, including transitions, occlusions, and motion blur. 30 DAVIS 2016 video sequences are used for training, and 20 video sequences are used for analysis. Each video series in DAVIS 2016 has only one object instance that is annotated. In contrast, the DAVIS 2017 dataset expands on the DAVIS 2016 dataset by annotating several elements in a single frame. The proposed approach is limited to the DAVIS 2016 dataset and only considers single instance segmentation. This dataset contains 155 video sequences, with a total of 570,000 frames in the Youtube-Object.

The YouTube-Objects dataset contains videos retrieved from YouTube after looking for ten distinct object types' names. There are around 9 to 24 videos for each class. Each video is approximately 30 minutes and 3 seconds long. While the videos are annotated indirectly, we guarantee each includes at least one object from the relevant class. The video sequences are then separated into ten categories. The Youtube-Object is commonly used solely for analysis since the training and test sets are not isolated. By comparing the Youtube-Object dataset to the SegTrack V2 dataset, the occlusions outnumber the appearance changes in around 14 video frames. The area similarity J_m and contour accuracy F_m are the calculation parameters. Other state-of-the-art semi-supervised and unsupervised VOS strategies are now being used to test the proposed study, including Lucid- Tracker [75], STV [76], MSK [77], ObjectFlow [17], TRC [30], CVOS [85], KEY [25], MSG [86] and NLC [1]. Our results are compared with these methods and our quantitative results outperform these methods in terms of J_m and F_m. On the CdNET 2014 night video dataset, the comparison is made between the ground truth segmentation and the segmentation generated by the proposed approach. It is then calculated for each frame in the night videos.

This helps us to evaluate the output of our method with the current state-of-the-art approaches. The results reflect the error flow over time and highlighting the points of failure [78]. In each method, deep neural networks evaluate the necessary parameter values and add the best features from the training data. These algorithms are trained and tested on two partitions of the same sequence. The deep learning network has poor generalization over unseen scenes due to a lack of annotated datasets. Our system was trained using CDNet videos, and it demonstrated strong generalization performance. In the fact that the current methods are mostly supervised, our method outperforms many of these methods. Our RL learning system is inherently more robust to relevant shifts in the frames than various methods such as COLBMOG [79], EFIC [80], and C-EFIC [81] based on *Rec* and *Pre* values. The *Rec* and *Acc* values for our method are 0.8799 and 0.9401, respectively. This means that our method has correctly obtained the segmentation masks and performs well compared to the other methods. Our RL agent is capable of making autonomous decisions with strong accuracies and *Pre*.

Figure 4, Figure 5, Figure 6, Figure 7 depicts the comparison of the foreground masks which are generated by our proposed RL method, EFIC [80], C-EFIC [81], and COLBMOG [79]. In Figure 4, from top to bottom, these images display the initial frame (Input), ground truth (GT), and foreground segmentation mask effects of our proposed RL method, EFIC [80], C-EFIC [81], and COLBMOG [79] for the frames numbered 1056 and 1220 of the winterStreet video. The misclassified pixels are highlighted in red. The quality of the segmentation masks provided by our approach and all the other three proposed methods are somewhat similar for frame 1056, which is a ''simple'' frame. In contrast, our RL method performs significantly better than the other approaches for frame 1220, which is a ''hard'' frame to detect. These qualitative results demonstrate our method's superior performance in strong reflections and the camouflaged objects originating from the street's headlights.

From left to right in Figure 4, it contains the bridgeEntry 1662nd frame, busyBoulvard frame number 820, fluidHighway frame 443, frame 2665 of streetCornerAtNight, frame 1636 of frame tramStation video, and frame 1278 of winterStreet video. The methods evaluated on these frames determine the shape of the cars in these videos. Our method can accurately determine the car's shape and flashlight (which is the front lightbox that emits light). This demonstrates the better performance of our RL method to extract the vehicle's shape. The red color highlights pixels that have been incorrectly labeled. Because the CDnet Night Videos dataset contains a wide range of obstacles, it's also essential to test our RL method's success in other difficult scenarios, such as shadows and complex backgrounds.

Figure 5 shows the comparison of our method's results with COLBMOG [79] for the input image. As it can be seen, our method segmentation mask is more accurate than the COLBMOG [79] method. In the second row the Figure 5 shows our RL algorithm performing significantly well than the COLBMOG [79] method. In the segmentation mask extracted, even the hands are visible clearly with our method. Figure 6 depicts the qualitative segmentation masks proposed by various VOS approaches. Our RL agent can accurately retrieve the primary objects by calculating the co-attention summaries that takes into account the global

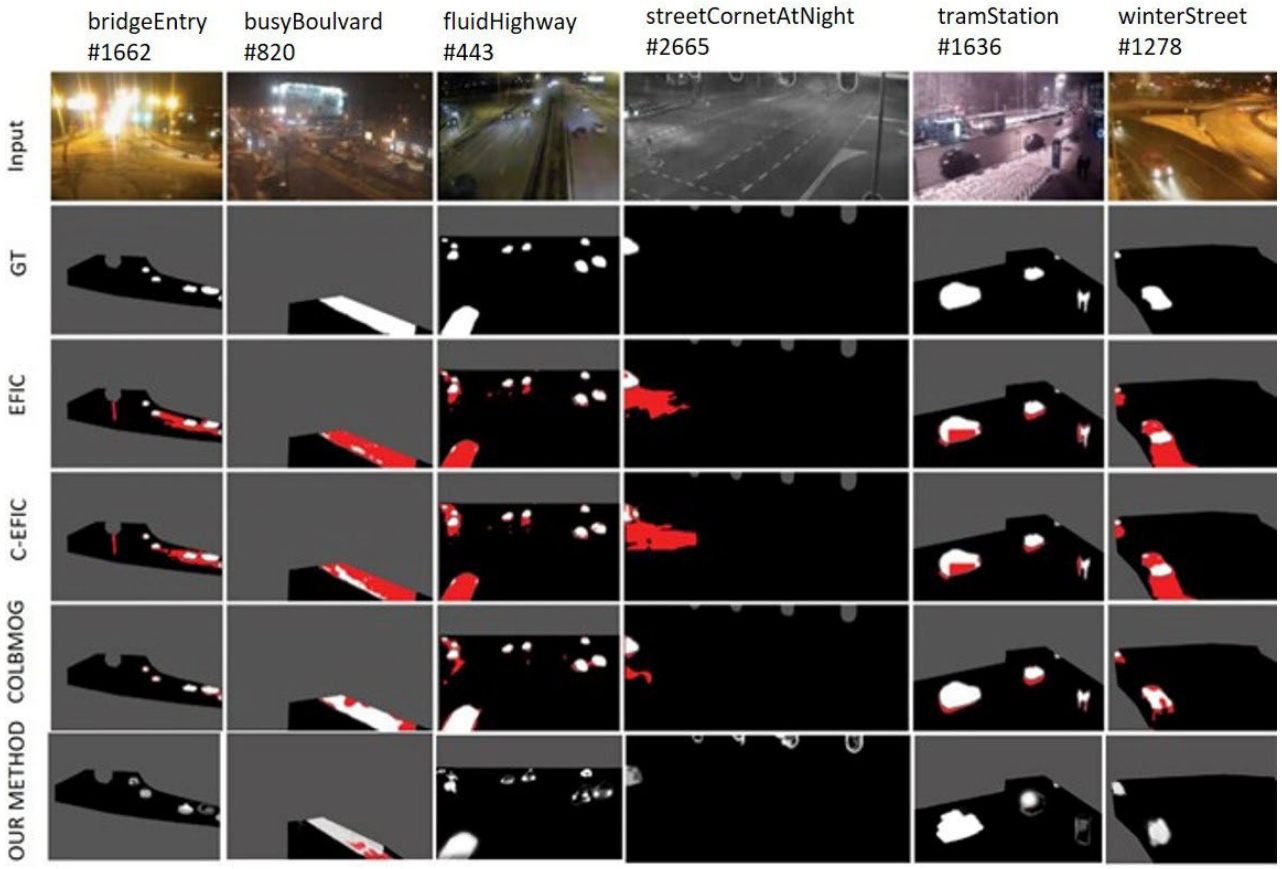

FIGURE 4. Example of segmentation masks for COLBMOG [79], C-EFIC [81], and EFIC [80], and the corresponding ground truth (GT), for the night videos from the CDnet 2014 dataset. Our method performs well than the current existing methods. As seen from the results it can be seen that the segmentation masks from our results are quite clear and our method is able to segment the objects more clearly. Even the headlights of the car are visible with our RL method, and the shape detected by our algorithm is quite similar to the Ground Truth.

temporal information (s). The agent can deal with fast motion scenarios (e.g., packour) and cluttered backgrounds (e.g., dancingtwirl).

Our RL approach emphasizes the primary subject while minimizing the comparable object distractions using both videos and static saliency images. This technique fits well in video clips with a lot of variation in the presence of the target entity, such as the camel and breakdance videos. Our method successfully separates the target object from many other similar items, mainly when they are near together, as in the camel series. Since the SegTrack V2 and Youtube-Object datasets do not distinguish between the training and test sets, both of these video sequences can be included for the evaluation of VOS approaches. Figure 7 shows the segmentation masks for the various methods such as COLBMOG [79], EFIC [80] and C-EFIC [81]. As it can be seen from the figure our method performs much better than the other method, giving clear Region of Interest (ROI).

Table 2 gives the average metrics for each of the videos and across the overall set of videos for our RL learning method. As seen in Table 2, the average F_m in the streetCornerAtNight dataset for the first half of the daytime videos is 0.9251, which is higher than COLBMOG [79] (0.7853), C-EFIC (0.7223) [81], and EFIC (0.6704) [80]. The values of other statistical measures are significantly higher than the other methods. The Night Videos genre is devoid of dynamic surroundings, which are significantly challenging to work with. With values of 0.9251 (bridgeEntry), 0.9401 (busyBoulvard), and 0.8378 (fluidHighway), we were able to achieve higher F_m averages for all video categories of Dynamic Background (DB) type, putting us well ahead of the methods evaluated on these video categories. This is an error measure instead of the F_m, so a lower value means better Acc. As it can be shown in the most challenging conditions, all algorithms fail in the same frames. In these frames, our RL method performs significantly well than the other state of the art methods evaluated on these dataset.

As seen in Table 3, our approach outperforms the others on the DAVIS 2016 dataset, SegTrack V2 and Youtube Object datasets. Our method is compared with eight deep learning models [3]–[5], [8], [9], [9], [36], [89], and eight conventional methods [1], [20], [25], [85]–[88], all of which are based on the DAVIS16 benchmark [17]. This shows that our method is more effective in terms of integrating the

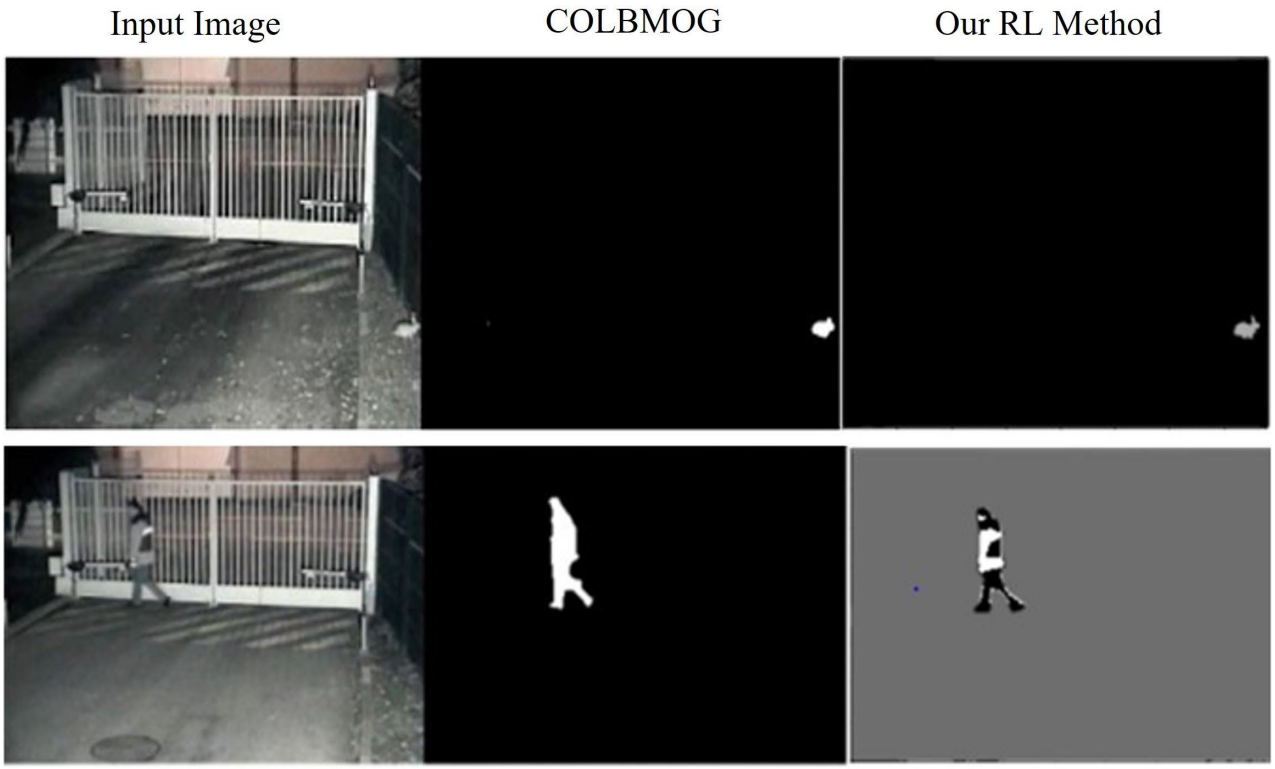

FIGURE 5. Our methods results compared with the COLBMOG [79] method for the Input Image. As it can be seen, our method segmentation mask is more accurate than the COLBMOG method in both the first and the second row.

TABLE 2. Average metrics for each of the CdNet 2014 videos dataset and across the overall set of videos for our RL method.

Video	*Spe*	*FPR*	*Sen*	*FNR*	MCC	*Pre*	F_m
streetCornerAt-Night	0.9231	0.0769	0.8750	0.1250	0.7943	0.9333	0.9251
tramStation	0.9133	0.0757	0.8690	0.1310	0.7899	0.9267	0.9401
winterStreet	0.9114	0.0786	0.8742	0.1258	0.7906	0.9112	0.8378
fluidHighway	0.9266	0.0734	0.9016	0.0984	0.8229	0.9451	0.9010
busyBoulvard	0.9345	0.0655	0.9059	0.0941	0.8393	0.9895	0.9519
bridgeEntry	0.9041	0.0959	0.8635	0.1365	0.7654	0.9171	0.9107
Average	0.9188	0.0776	0.8815	0.1184	0.8004	0.9371	0.9111
St. Dev.	0.01022	0.00920	0.01620	0.1172	0.0264	0.0258	0.0039

information common for the inference of mask. We infer that our RL method considers more information related to cross-frame relation. Compared to previous state-of-the-art methods, this technique raises the mean area similarity J_m on the Seg-Track *V2* dataset by 12.03%, demonstrating the usefulness of RL models when selecting online adaptation ROIs. The proposed method also improves the mean-field correlation J_m by 13.11% on the Youtube-Object dataset. The grid objects (such as trains and aircraft) and non-grid objects are the two types of categories in Youtube-Things (e.g., Cat, Bird). Despite the fact that the objects in the latter class often undergo fast appearance change and shape deformation our RL method maintains and captures long-term dependency better than any other method evaluated on this dataset.

The proposed approach outperforms the other methods on the DAVIS 2016 dataset with an F_m of 91.85%. The values of J_m on the three datasets are 90.68%, 89.63% and 92.61% for DAVIS 2016, SegTrack V2 and Youtube Object dataset respectively. Our approach can handle significant appearance variations caused by interacting objects, size differences, appearance change and background clutter. Because all of the techniques investigated, including optical flow fusion in FSEG [4], multi-tasks estimation in SFL [5] and ConvLSTM in PDB [8], we use the temporal information to estimate the segmentation mask. Our RL agent has the benefit of leveraging temporal correlations through the co-attention mechanism which becomes apparent from a global perspective. Our method can capture temporal coherence and differentiate between foreground and background objects. The better performance of our RL model is due to the group co-attention calculated by our agent which learns the fusing and capturing of the correlation information from multiple reference frames.

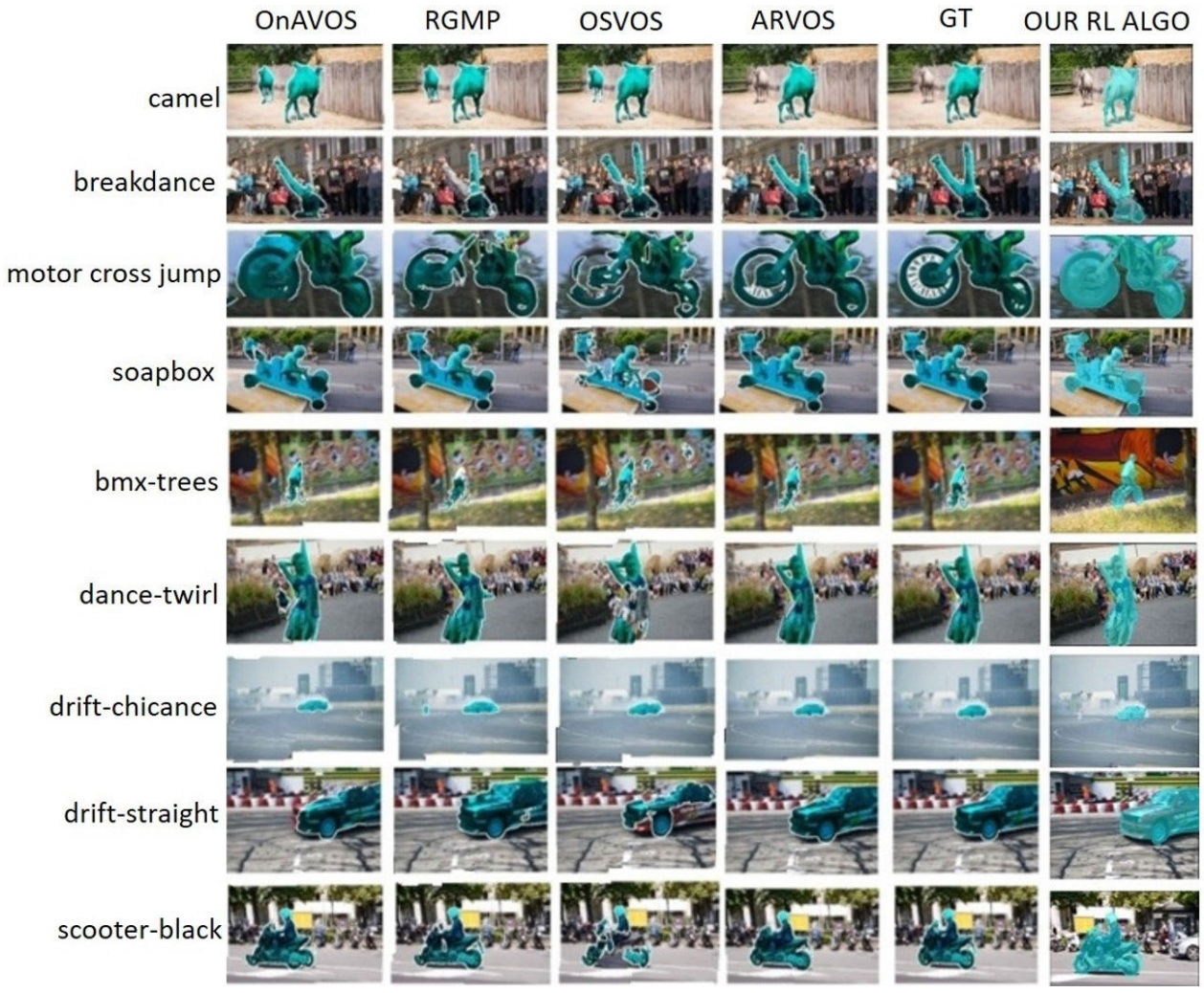

FIGURE 6. The segmentation masks for the various methods. As it can be seen from the figure our method performs much better than the other method, giving very clear ROIs. Our RL method helps in easy detection of all the segmentation masks for the ROIs taken into consideration.

In Table 4, the validity of the suggested approach is shown by the fact that our algorithm runs better than the other approaches. The quantitative results for the average value metrics to evaluate the night video dataset segmentation results are compared with the other state-of-the-art methods. In this, we assess the foreground segmentation results by using around seven metrics that are well-known and are based on the number of correct and incorrect classified pixels. The seven measures used are *Spe*, *Rec*, FalsePositiveRate (FPR), FalseNegativeRate (FNR), *Sen*, and F_m. We calculate and evaluate the overall success of our method using these various parameters. The metrics for each video group are then measured to ensure that our RL approach is performing well. We also calculate the average metrics in addition to the calculation of statistical metrics for each video. The average F_m for each video and the entire set of other videos are shown in Table 4. The standard deviation of a group of videos is often measured. When we do a significance test such as the FriedMan test on each of the experimental outcome, we infer that the total ranks values of the respective column sum up to 34,25, 50 and 52. After calculating the p value, we find that the value is 0.00042, which is less than 0.05. This suggests that our results are significant.

Our system consistently outperforms the COLBMOG [79], C-EFIC [81], and EFIC [80] approaches in terms of F_m ranking, with a 16.8 % relative overall F-Measure improvement over the previously suggested strategies, COLBMOG [79], a 25.74 % relative overall F-Measure improvement over C-EFIC [81], and a 27.03 % relative overall F-Measure improvement over the EFIC [80]. Our RL method outperforms the other approaches in all video types, including bridgeEntry, busyBoulvard, and fluidHighway. Just a few red-marked regions in Figure 4 infer that our algorithm can extract the shapes with great precision. COLBMOG is focused on the low complexity color-based classification algorithm BMOG. Due to the low-quality video with visible compression noise, it produces inaccurate textures and has a significant effect on texture representation. The statistical measures infer that

TABLE 3. On the DAVIS 2016, SegTrack *V2*, and Youtube-Object datasets, the quantitative results of our RL method are compared with the other state-of-the-art methods. Our method performs relatively well in comparison with the other methods.

Method	DAVIS-16 J_m	DAVIS-16 F_m	SegTrack V2 J_m	Youtube Object J_m
PreMVOS [82]	84.9	88.6	-	-
OnAVOS [83]	85.7	84.8	66.7	77.4
CINM [42]	83.4	85.0	77.1	78.4
Lucid [75]	83.7	-	76.8	76.2
MSK [77]	79.7	75.4	72.1	75.6
OSVOS [40]	79.8	80.6	65.4	78.3
STV [76]	73.6	-	78.1	-
ObjFLow [16]	68.0	-	74.1	77.6
ARVOS [84]	87.1	86.1	77.6	79.5
SAG [27]	42.6	38.3	38.6	70.9
TRC [30]	47.3	44.1	49.3	69.3
CVOS [85]	48.2	44.7	54.0	73.6
KEY [19]	49.8	42.7	59.1	86.2
MSG [86]	53.3	50.8	61.6	78.0
NLC [1]	55.1	52.3	55.8	81.7
CUT [87]	55.2	55.2	57.5	65.6
FST [88]	55.8	51.1	64.9	60.5
SFL [5]	67.4	66.7	81.4	77.2
LMP [9]	70.0	65.9	85.0	70.6
FSEG [4]	70.7	65.3	83.0	76.0
LVO [3]	75.9	72.1	77.1	56.1
ARP [89]	76.2	70.6	79.2	81.0
PDB [8]	77.2	74.5	73.8	80.0
LSMO [9]	78.2	75.9	83.4	63.8
MOT [36]	77.2	77.4	83.5	65.4
Our RL Method	90.68	91.85	89.63	92.61

TABLE 4. Average metrics across the overall set of videos for our RL Method, COLBMOG [79], C-EFIC [81], EFIC [80].

Metric	EFIC [80]	C-EFIC [81]	COLBMOG [79]	Our Method
Re	0.6704	0.7223	0.8047	0.8799
Sp	0.9893	0.9866	0.9889	0.9618
FPR	0.0107	0.0134	0.0111	0.0382
FNR	0.3296	0.2777	0.1953	0.1184
Pr	0.6869	0.6636	0.7287	0.9401
F-measure	0.6548	0.6677	0.7564	**0.9251**

our system outperforms other approaches in these uneven datasets. The proposed system's *Acc* and *Pre* scores also raise the F_m value, which is considerably better than the other methods. Compared to different algorithms, our RL algorithm has a lower standard deviation of the F_m across the entire range of images, suggesting better efficiency. The F_m is found to be 0.0552. In the COLBMOG [79] method, the BMOG model is used. The contribution of the value-added applied by COLBMOG to the device's overall performance is apparent when contrasting the F_m obtained by BMOG for the Night Videos segment is 0.4982, with the one received by COLBMOG [73] (0.7564).

For video datasets, our method outperforms other methods in both qualitative and quantitative terms, with some promising results in pedestrian segmentation. It detects the dark region at the bottom right of the images accurately. Our results for the fluidHighway video, a low-quality video with visible compression noise that produces false textures and directly affects texture representation, are excellent. Our system outperforms the other approaches by a broad margin, in all the challenging cases. Our method has a significantly lower standard deviation of the F_m over the entire continuum of videos, implying more accurate *Pre* across various obstacles. These statistical measures serve as a key metric for inferring the better performance of our method. We evaluate our system's *Acc* in some more complex cases, such as shadows and complicated backgrounds. According to ablation studies, the proposed method outperforms current methods since it uses the co-attention mechanism. We propose a co-attention mechanism-based RL framework using Siamese Networks. The Siamese neural network (sometimes referred to as a twin neural network) is a kind of artificial neural network that uses the same weights to produce identical output vectors while processing two different input vectors. Table 5 gives the average performance metrics for the different categories of DAVIS 2016 video dataset. Our method performs relatively well for all categories of videos.

Our VOS approach can achieve long-term, difficult to achieve outcomes. The model fix mistakes made during the training phase. Once the model has resolved an error, the likelihood of the same error happening again is very low. Our RL method strikes a balance between exploration and exploitation. Exploration is the practice of searching for new samples and exploiting the promising areas explored

TABLE 5. Performance Metrics for the different categories of videos. The performance metrics are *Acc*, Dice, *Fm*, *Jm*, Matthews Correlation Coefficient (MCC), *Pre*, *Sen* and *Spe*.

Video	*Acc*	*Dice*	F_m	J_m	MCC	*Pre*	*Sen*	*Spe*
camel	0.8893	0.9185	0.9185	0.8493	0.74599	0.9191	0.9179	0.8284
breakdance	0.8491	0.8052	0.8052	0.6739	0.6829	0.7838	0.8278	0.8260
motor-crossjump	0.9180	0.8305	0.8305	0.7101	0.7777	0.8677	0.7963	0.9590
soapbox	0.8597	0.8694	0.8694	0.7690	0.7180	0.8612	0.8778	0.8392
bmx-trees	0.8295	0.7601	0.7601	0.6131	0.6308	0.8045	0.7204	0.8950
dance-twirl	0.8704	0.8969	0.8969	0.8130	0.7230	0.8863	0.9077	0.8094
drift chicance	0.9563	0.9543	0.9543	0.9126	0.9126	0.9615	0.9472	0.9648
drift-straight	0.8299	0.8554	0.8554	0.7474	0.6626	0.7959	0.9246	0.7169
scooter black	0.8002	0.8214	0.8214	0.6969	0.6051	0.7670	0.8841	0.7096

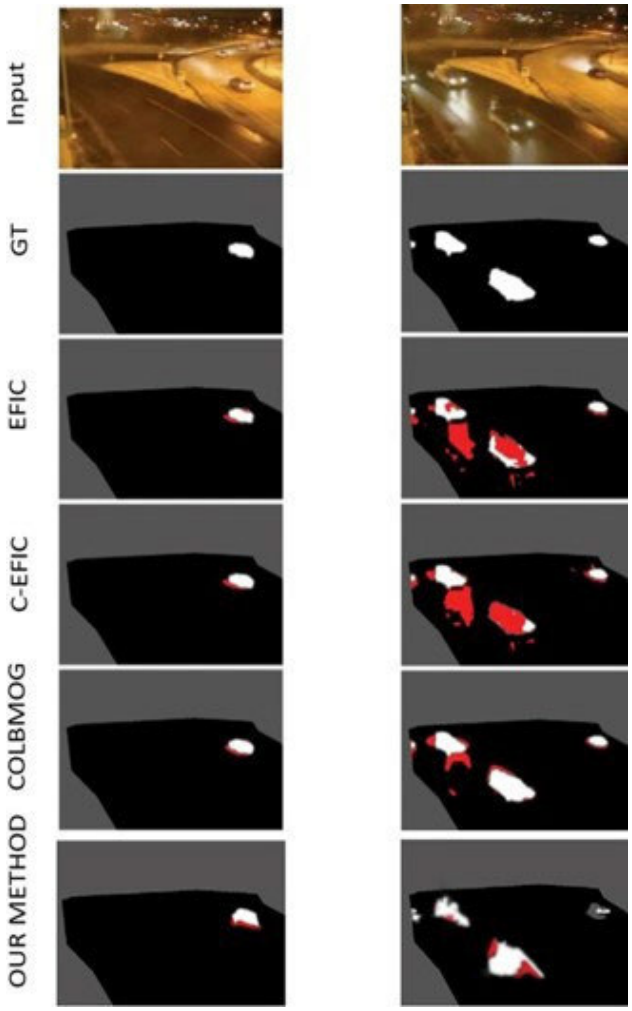

FIGURE 7. Comparison of our segmentation masks with COLBMOG, C-EFIC and EFIC. The segmentation masks calculated by our method is better as compared to the other state of the art methods.

during exploitation. Most machine learning algorithms do not maintain this balance. Furthermore, the mentioned issue is a general problem in various video-related tasks, and our proposed RL approach can be applied to other video-related tasks. The stronger the match between our predicted objects and the ground truth, the higher the value of *Pre*. The main advantage of an encoder-based Siamese network over a regular encoder network is the ability to quickly detect similar target objects and foreground information propagation. With ZVOS, the Siamese network function performs well. As a consequence, it can entirely replace online fine-tuning while also substantially speeding up the segmentation process. The segmentation accuracy is higher than the online fine-tuning, and the Siamese network can achieve the speed-accuracy trade-off. As a result, the amount of error produced during the segmentation process is reduced. The error minimization function has been extended to various other video-related functions, enabling the current frame's output to monitor the output of subsequent frames. In several other VOS methods, the current frame's segmentation results are paired with the information from the next frame. Our technique benefits from the use of the error minimization process.

RL algorithms have their state space, function space, transfer process, and reward defined. From this vantage point, our state-space comprises the assignment network's fixed inputs, including normal (image, flow, etc.) and unique to the current proposal inputs (current mask, appearance, etc.). It can be seen from the results, our method needs much less training data (especially video data) than other methods like LVO [3], FSEG [4], OBN [7], LSMO [9], MOT [36] while still obtaining better results. DDPG is composed of two models: actor and critic [90]. Rather than the probability distribution over the actions, the state is given as an input to the actor (policy network) and outputs the exact action (continuous). The state and action is given as an input to the critic and it produces a Q-value as an output. The term "deterministic" in DDPG refers to the fact that the actions are computed directly by the actor rather than utilizing a probability distribution across actions. We also compare our computational time and hardware resources required for various deep architectures with our RL method.

The methods taken for comparison are SegNet [91], VNet [92], UNet [93] and Autoencoders [94]. Our RL algorithm performs significantly better than the other methods in terms of GPU training memory, GPU inference memory, forward pass, and the backward pass time with a value of 4069MB, 2700MB, 102.22ms and 144.49ms. From the results we infer that our RL algorithm can solve the complex problem of VOS in the night video dataset. Our method does not need a large number of scene flow images (such as those used in LVO [3] and LSMO [9] to train an optical flow module since it takes the video frames as input. Our method benefits from the natural data augmentation property of our Siamese network-based learning method. Our method

also outperforms FSEG [4], LSMO [9], MOT [36] which all need extra video datasets. Our method also outperforms the PDB [8] with the same training data, showing the importance of global knowledge for ZVOS tasks.

V. CONCLUSION

Our RL model automatically recognizes and isolates the major object regions in each frame of a video. Unlike the conventional methods, which focused on sequential and local data, this research emphasizes the significance of the global co-attention mechanism. We propose our RL model to capture temporal coherence by gathering correlation information between frames group (or pairs) via a differentiable co-attention mechanism. The proposed method identifies the significant background objects for each frame, capturing the temporal correlation across frames. Our model can capture similar objects and minimize comparable target distraction even when no annotation is given during segmentation. We can extend our RL model to other video analysis applications such as video saliency detection and optical flow estimates. In the future, more powerful co-attention mechanisms can be exploited and the idea of meta-learning can be incorporated into the design. The algorithm can be tested for the detection of the primary objects in more complex scenarios. Our RL method ranks first in the CDnet ''Night Videos'' with an F_m score of 0.9251. This makes it the best performing method for the segmentation of irregular objects in night video datasets. The results reveal that the proposed results boost the state-of-the-art techniques in the F1 measure on the DAVIS 2016 dataset by 2%, SegTrack V2 by a J_m of 12.03%, and on the Youtube Object dataset by a J_m of 13.11 %. Meanwhile, our algorithm achieves an accuracy of 87.99%, precision of 94.01%, and F_m of 92.51% on the DAVIS 2016 dataset, thus ranking higher than the current state-of-the-art methods on the video segmentation datasets.

VI. DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this article.

REFERENCES

[1] A. Faktor and M. Irani, ''Video segmentation by non-local consensus voting,'' in *Proc. BMVC*, vol. 2, 2014, p. 8.

[2] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, and B. Chen, ''JumpCut: Non-successive mask transfer and interpolation for video cutout,'' *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–195, 2015.

[3] P. Tokmakov, K. Alahari, and C. Schmid, ''Learning video object segmentation with visual memory,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4481–4490.

[4] S. D. Jain, B. Xiong, and K. Grauman, ''FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2126.

[5] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, ''SegFlow: Joint learning for video object segmentation and optical flow,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 686–695.

[6] S. Li, B. Seybold, A. Vorobyov, A. Fathi, Q. Huang, and C.-C.-J. Kuo, ''Instance embedding transfer to unsupervised video object segmentation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6526–6535.

[7] S. Li, B. Seybold, A. Vorobyov, X. Lei, and C.-C. J. Kuo, ''Unsupervised video object segmentation with motion-based bilateral networks,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 207–223.

[8] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, ''Pyramid dilated deeper ConvLSTM for video salient object detection,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 715–731.

[9] P. Tokmakov, C. Schmid, and K. Alahari, ''Learning to segment moving objects,'' *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 282–301, 2019.

[10] S. H. Abdulhussain, S. A. R. Al-Haddad, M. I. Saripan, B. M. Mahmmod, and A. Hussien, ''Fast temporal video segmentation based on Krawtchouk-Tchebichef moments,'' *IEEE Access*, vol. 8, pp. 72347–72359, 2020.

[11] J. Li, Y. Zhao, J. Fu, J. Wu, and J. Liu, ''Attention-guided network for semantic video segmentation,'' *IEEE Access*, vol. 7, pp. 140680–140689, 2019.

[12] K. Nakamura, N. Nitta, N. Babaguchi, K. Fujii, S. Matsumura, and E. Nabata, ''Semi-supervised temporal segmentation of manufacturing work video by automatically building a hierarchical tree of category labels,'' *IEEE Access*, vol. 9, pp. 68017–68027, 2021.

[13] B. Wang, L. Li, Y. Nakashima, R. Kawasaki, H. Nagahara, and Y. Yagi, ''Noisy-LSTM: Improving temporal awareness for video semantic segmentation,'' *IEEE Access*, vol. 9, pp. 46810–46820, 2021.

[14] S. Chakraborty and D. M. Thounaojam, ''SBD-duo: A dual stage shot boundary detection technique robust to motion and illumination effect,'' *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 3071–3087, Jan. 2021.

[15] S. Chakraborty, D. M. Thounaojam, and N. Sinha, ''A shot boundary detection technique based on visual colour information,'' *Multimedia Tools Appl.*, vol. 80, no. 3, pp. 4007–4022, Jan. 2021.

[16] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, ''A benchmark dataset and evaluation methodology for video object segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 724–732.

[17] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, ''Motion coherent tracking using multi-label MRF optimization,'' *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 190–202, 2012.

[18] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, ''Learning object class detectors from weakly annotated video,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3282–3289.

[19] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, ''CDnet 2014: An expanded change detection benchmark dataset,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 387–394.

[20] H. Li, G. Chen, G. Li, and Y. Yu, ''Motion guided attention for video salient object detection,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7274–7283.

[21] R. Mishra, ''Video shot boundary detection using hybrid dual tree complex wavelet transform with Walsh Hadamard transform,'' *Multimedia Tools Appl.*, vol. 80, pp. 1–27, May 2021.

[22] S. Zhou, X. Wu, Y. Qi, S. Luo, and X. Xie, ''Video shot boundary detection based on multi-level features collaboration,'' *Signal, Image Video Process.*, vol. 15, no. 3, pp. 627–635, Apr. 2021.

[23] T. Wang, N. Feng, J. Yu, Y. He, Y. Hu, and Y.-P. P. Chen, ''Shot boundary detection through multi-stage deep convolution neural network,'' in *Proc. Int. Conf. Multimedia Modeling*, 2021, pp. 456–468.

[24] C. Xu and J. J. Corso, ''Evaluation of super-voxel methods for early video processing,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1202–1209.

[25] J. Chang, D. Wei, and J. W. Fisher, III, ''A video representation using temporal superpixels,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2051–2058.

[26] M. Grundmann, V. Kwatra, M. Han, and I. Essa, ''Efficient hierarchical graph-based video segmentation,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2141–2148.

[27] D. Zhang, O. Javed, and M. Shah, ''Video object segmentation through spatially accurate and temporally dense extraction of primary object regions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 628–635.

[28] W. Wang, J. Shen, R. Yang, and F. Porikli, ''Saliency-aware video object segmentation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[29] Y.-H. Tsai, G. Zhong, and M.-H. Yang, ''Semantic co-segmentation in videos,'' in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 760–775.

[30] Y.-T. Hu, J.-B. Huang, and A. G. Schwing, "Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 786–802.
[31] K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik, "Learning to segment moving objects in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4083–4090.
[32] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3899–3908.
[33] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan, "Learning features by watching objects move," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2701–2710.
[34] G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin, "Flow guided recurrent neural encoder for video salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3243–3252.
[35] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3119–3127.
[36] M. Siam, C. Jiang, S. Lu, L. Petrich, M. Gamal, M. Elhoseiny, and M. Jagersand, "Video object segmentation using teacher-student adaptation in a human robot interaction (HRI) setting," in *Proc. Int. Conf. Robot. Automat. (ICRA)*, May 2019, pp. 50–56.
[37] X. Lu, W. Wang, J. Shen, Y.-W. Tai, D. J. Crandall, and S. C. H. Hoi, "Learning video object segmentation from unlabeled videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8960–8970.
[38] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S. C. H. Hoi, and H. Ling, "Learning unsupervised video object segmentation through visual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3064–3074.
[39] J. Cheng, Y.-H. Tsai, W.-C. Hung, S. Wang, and M.-H. Yang, "Fast and accurate online video object segmentation via tracking parts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7415–7424.
[40] K. K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. V. Gool, "Video object segmentation without temporal information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1515–1530, May 2018.
[41] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos, "Efficient video object segmentation via network modulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6499–6507.
[42] L. Bao, B. Wu, and W. Liu, "CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5977–5986.
[43] W. Wang, J. Shen, F. Porikli, and R. Yang, "Semi-supervised video object segmentation with super-trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 985–998, Apr. 2019.
[44] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung, "Fully connected object proposals for video segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3227–3234.
[45] V. Jampani, R. Gadde, and P. V. Gehler, "Video propagation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 451–461.
[46] S. W. Oh, J.-Y. Lee, K. Sunkavalli, and S. J. Kim, "Fast video object segmentation by reference-guided mask propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7376–7385.
[47] Y.-T. Hu, J.-B. Huang, and A. G. Schwing, "VideoMatch: Matching based video object segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 54–70.
[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[49] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6087–6096.
[50] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas, "Learning where to attend with deep architectures for image tracking," *Neural Comput.*, vol. 24, no. 8, pp. 2151–2184, Aug. 2012.
[51] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.
[52] S. Jiang, A. Armaly, and C. Mcmillan, "Automatically generating commit messages from diffs using neural machine translation," in *Proc. 32nd IEEE/ACM Int. Conf. Automated Softw. Eng. (ASE)*, Oct. 2017, pp. 135–146.
[53] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1913–1927, Aug. 2020.
[54] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[55] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 220–237, Jan. 2021.
[56] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5012–5021.
[57] Z. Zheng, W. Wang, S. Qi, and S.-C. Zhu, "Reasoning visual dialogs with structural and partial observations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6669–6678.
[58] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.
[59] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[60] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7464–7473.
[61] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 289–297.
[62] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan, "Beyond RNNs: Positional self-attention with co-attention for video question answering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8658–8665.
[63] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi, "Action-decision networks for visual tracking with deep reinforcement learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2711–2720.
[64] T. Zhang, M. Huang, and L. Zhao, "Learning structured representation for text classification via reinforcement learning," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.
[65] X. Dong and Y. Yang, "Searching for a robust neural architecture in four GPU hours," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1761–1770.
[66] T. Chen, J. Liu, Y. Xiang, W. Niu, E. Tong, and Z. Han, "Adversarial attack and defense in reinforcement learning-from AI security view," *Cybersecurity*, vol. 2, no. 1, pp. 1–22, Dec. 2019.
[67] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.
[68] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.
[69] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 365–381.
[70] X. Wu, K. Xu, and P. Hall, "A survey of image synthesis and editing with generative adversarial networks," *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 660–674, 2017.
[71] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[72] M. Paniri, M. B. Dowlatshahi, and H. Nezamabadi-Pour, "Ant-TD: Ant colony optimization plus temporal difference reinforcement learning for multi-label feature selection," *Swarm Evol. Comput.*, vol. 64, Jul. 2021, Art. no. 100892.
[73] S. C. Yurtkulu, Y. H. Sahin, and G. Unal, "Semantic segmentation with extended DeepLabv3 architecture," in *Proc. 27th Signal Process. Commun. Appl. Conf. (SIU)*, Apr. 2019, pp. 1–4.
[74] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo, and T. Liu, "Residual networks of residual networks: Multilevel residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1303–1314, Jun. 2018.

[75] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele, "Lucid data dreaming for object tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, 2017, pp. 1680–1688.
[76] W. Wang, J. Shen, J. Xie, and F. Porikli, "Super-trajectory for video segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1671–1679.
[77] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2663–2672.
[78] L. A. Lim and H. A. Keles, "Learning multi-scale features for foreground segmentation," *Pattern Anal. Appl.*, vol. 23, no. 3, pp. 1369–1380, 2020.
[79] I. Martins, P. Carvalho, L. Corte-Real, and J. L. Alba-Castro, "Texture collinearity foreground segmentation for night videos," *Comput. Vis. Image Understand.*, vol. 200, Nov. 2020, Art. no. 103032.
[80] G. Allebosch, F. Deboeverie, P. Veelaert, and W. Philips, "EFIC: Edge based foreground background segmentation and interior classification for dynamic camera viewpoints," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.*, 2015, pp. 130–141.
[81] G. Allebosch, D. V. Hamme, F. Deboeverie, P. Veelaert, and W. Philips, "C-EFIC: Color and edge based foreground background segmentation with interior classification," in *Proc. Int. Joint Conf. Comput. Vis., Imag. Comput. Graph.*, 2015, pp. 433–454.
[82] J. Luiten, P. Voigtlaender, and B. Leibe, "PReMVOS: Proposal-generation, refinement and merging for video object segmentation," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 565–580.
[83] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.
[84] M. Sun, J. Xiao, E. G. Lim, Y. Xie, and J. Feng, "Adaptive ROI generation for video object segmentation using reinforcement learning," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107465.
[85] B. Taylor, V. Karasev, and S. Soattoc, "Causal video object segmentation from persistence of occlusions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4268–4276.
[86] P. Ochs and T. Brox, "Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1583–1590.
[87] M. Keuper, B. Andres, and T. Brox, "Motion trajectory segmentation via minimum cost multicuts," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3271–3279.
[88] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1777–1784.
[89] Y. J. Koh and C.-S. Kim, "Primary object segmentation in videos based on region augmentation and reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7417–7425.
[90] J. K. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, 2017, pp. 66–83.
[91] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[92] A. Abdollahi, B. Pradhan, and A. Alamri, "VNet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data," *IEEE Access*, vol. 8, pp. 179424–179436, 2020.
[93] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.
[94] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. ICML Workshop Unsupervised Transf. Learn.*, 2012, pp. 37–49.

USMAN AHMAD USMANI was born in Aligarh, India, in April 1993. He is currently pursuing the Ph.D. degree in computer science with the Universiti Teknologi PETRONAS, Malaysia. He has worked as a Research Assistant at IIT Kanpur and as a Researcher with Massey University, New Zealand. He has built up a social network named Zamber that has been published in around 14 national newspapers. His area of research interests include artificial intelligence, computer vision, computer security, wearable sensors, and cloud computing.

JUNZO WATADA received the B.Sci. and M.Sci. degrees in electrical engineering from Osaka City University, Japan, and the Ph.D. degree from Osaka Prefecture University, Japan. After retiring from Waseda University, he contributed as a Full Professor with the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia, and a Professor Emeritus with Waseda University, Japan. He is a Life Fellow of Bio Medical Fuzzy System Association (BMFSA) and the Japan Society of Fuzzy Theory and Intelligent Informatics. His research interests include big data analytics, soft computing, image processing systems to track human behaviors and understand pictures and videos, knowledge engineering, and management engineering.

JAFREEZAL JAAFAR (Senior Member, IEEE) received the Ph.D. degree from The University of Edinburgh, U.K., in 2009. He is currently an Associate Professor and the former Head of the Computer and Information Sciences Department at Universiti Teknologi PETRONAS, Malaysia. His main research areas include big data analytics, soft computing, and software engineering. He has secured a number of research projects from the industry and government agencies. Based on his publication track records, he had been appointed as the chief editor and a reviewer for several journals, and also the chair, technical chair, and committee member for several international conferences. He is also active in IEEE Computer Society, Malaysia Chapter, and he has been appointed as the Executive Committee Member in 2016 and 2017.

IZZATDIN ABDUL AZIZ received the Ph.D. degree in information technology from Deakin University, Australia, working in the domain of hydrocarbon exploration and cloud computing. He is currently a Researcher with the High Performance Cloud Computing Centre (HPC3), Universiti Teknologi PETRONAS (UTP), where he focuses in solving complex upstream oil and gas (O&G) industry problems from the view point of computer sciences. He currently serves as the Deputy Head of the Computer and Information Sciences Department, UTP. He is working closely with O&G companies in delivering solutions for complex problems, such as offshore O&G pipeline corrosion rate prediction, O&G pipeline corrosion detection, securing data on clouds and designing and implementing Metocean prediction systems, and bridging upstream and downstream oil and gas businesses through data analytics. In addition, he is also working on big data transmission, security, and optimization problems on high performance clouds.

ARUNAVA ROY received the Ph.D. degree from the Department of Applied Mathematics, Indian School of Mines, Dhanbad. He currently works as a Researcher with the Department of Industrial and Systems Engineering, National University of Singapore, Singapore. Previously, he was a Postdoctoral Fellow with the Department of Computer Science, The University of Memphis, TN, USA. His areas of interests include web software reliability, software reliability, cyber security, algorithm design and analysis, data structure, and statistical and mathematical modeling.