RefMOS: A Robust Referred Moving Object Segmentation framework based on text query

Prafulla Saxena[1], Susim Mukul Roy[2], Dinesh Kumar Tyagi[1], Santosh Kumar Vipparthi[3], Subrahmanyam Murala[3,5], R. Balasubramanian[4]

[1]MNIT Jaipur, [2]IIT Jodhpur, [3]IIT Ropar, [4]IIT Roorkee [5]SCSS Trinity College Dublin

Abstract

Referred Moving object segmentation is a very challenging task in automated video surveillance applications as it requires additional information to learn about object representation referred by natural language expression. In segmenting specific moving objects targeted by a text, suppressing other moving as well as stationary objects is a crucial task. A better context needs to be learned where linguistic, spatial, and temporal features need to be taken into account. In this work, we have proposed a robust referred moving object segmentation (RefMOS) framework to capture moving objects referred by text query. Most of the earlier state-of-the-art methods exploit a different type of supervision by treating video frames as images but lack temporal information during processing. In this work, we have proposed an inter-frame movement detector (IFCD) module, which extracts the movement information between the consecutive frames and helps integrate temporal information with spatial visual features. Language embedding is utilized to capture the information of referred moving objects in the text by extracting linguistic features from a pre-trained language model, i.e., BERT. Furthermore, the cross-entropy loss and SGD optimizer are used to train the network. Our RefMOS framework competes with the state-of-the-art approaches and achieves 48.6 mean IOU on the ref-DAVIS 17 dataset.

1. Introduction

Referred video object segmentation (RVOS) is a potentially significant and emerging field at the intersection of computer vision and natural language processing. RVOS aims to segment an object with the reference of a language expression. On the other hand, object segmentation focuses on the spatial domain at the image level and learns the context within the single frame to understand the object representation better. In capturing the movement of an object in a video, these approaches do not consider temporal information. In comparison, standard moving object segmentation (MOS) focuses on extracting temporal and generates a mask of moving objects but does not consider the semantic understanding and context information of objects[12].

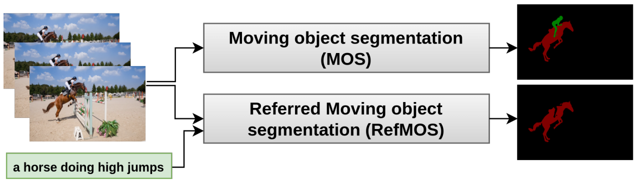
Moving object segmentation (MOS)
Referred Moving object segmentation (RefMOS)
a horse doing high jumps

Figure 1. In RefMOS, the segmentation of moving objects depends on a text query. Whether in MOS task, all moving objects are segmented irrespective of the text query.

Fig. 1 shows the significance of Referred moving object segmentation (RefMOS) over the standard moving object segmentation (MOS). Consecutive frames are analyzed to find moving regions by capturing changes in pixels, but specific characteristics must be targeted when referring to objects through text. Contextual information is essential for understanding the object representation in videos rather than solely identifying moving regions. Thus, spatiotemporal features become more important for the referred moving object segmentation (RefMOS). Even though prevailing challenges of MOS, like illumination changes, low visibility, dynamic backgrounds, hidden objects, shadow, noise, and camera motion, similarly affect RefMOS, text embedding adds another challenge to tackle. Moreover, the segmentation task involves a significantly smaller number of relevant pixels, and much of the information seems irrelevant and redundant. This irrelevant information can have a negative impact on the objective, leading to a degradation in the overall performance of moving object segmentation. Most of the approaches in the literature have proposed various learning-based approaches to alleviate these problems by designing different types of deep learning techniques [1, 6, 8, 12, 14, 15, 20] and given a significant performance improvement in MOS task. However, the segmentation of moving objects referred by language expressions has received relatively little attention in the literature. compared to the conventional semi-supervised video object segmentation [17, 24], Referred moving object segmentation is more difficult to proceed with and typically needs more ground-truth mask annotations in advance. This needs a thorough grasp of the cross-modal sources, such as vision and language. As a result, the model should be highly capable of determining which moving object a reference is made to. In literature, existing methods utilize two strategies to handle the task's difficulty. 1) Bottom-up approaches [2]. These techniques combine language and visual information in an early-fusion manner. (2) Top-down approaches [11]. These approaches employ a two-stage pipeline where instance segmentation is done at the early stage, and then, based on a query, a grounding model segments the desired object in the second stage. Both approaches have pros and cons. Bottom-up approaches fail to capture instance-level information, and object association does not follow across multiple frames. Furthermore, top-down approaches proceed with a complected, multistage pipeline and suffer from a heavy workload. Both methods may need more crucial information if temporal information is not utilized for the referred moving object segmentation task. So, in this work, we have incorporated the temporal information to capture the change between consecutive frames. An object's temporal and visual features will guide the process of segmenting the real moving object. Our method considers the multiple frames as input to capture the moving objects. The linguistic features are extracted from the Language Feature Extractor (LFE), which helps segment the targeted referred moving object. We evaluated our state-of-the-art RefMOS method and achieved 48.6 mean IOU on the Ref-DAVIS-17 [18] dataset. We propose the following contribution to this paper.

- An end-to-end referred moving object segmentation (RefMOS) framework for segmenting objects which aims to target a moving object referred by a query.
- The Inter-Frame Change Detector (IFCD) module extracts substantial changes that capture movement information and spatio-temporal structural dependencies between consecutive frames.
- RefMOS framework does not store history for the background model, which is required in standard MOS methods and processes with three consecutive frames during training.
- Utilizes the BERT language model for linguistic features extraction.
- Evaluate on Ref-DAVIS17 and DAVIS-16 benchmark dataset and verify the proposed module's effectiveness with quantitative and qualitative analysis.

2. Related work

Referred video object segmentation is a crucial task in computer vision and video analysis. It involves the segmentation of a specific object or region throughout the video sequence, guided by a provided language description. In the literature, extensive research has been carried out, and various segmentation approaches have been introduced that are tailored to specific requirements for images or videos. These approaches target specific objects, whether moving or stationary, depending on the problem domain.

Image Segmentation (IS) Image segmentation is one of the pioneer tasks in computer vision applications. Image segmentation is the process of partitioning an image into multiple segments or regions. This is a highly researched area in computer vision with various approaches like U-net[19], Deeplabv3[4] in various domains. U-net[19] is CNN architecture designed for biomedical image segmentation. Deeplabv3[4] is a neural network architecture for semantic image segmentation.

Video object Segmentation (VOS) - VOS aims to separate a specific object of interest and segment it throughout the video. In this task, the target object is known in advance with a set of features. The goal is to segment it from the background and keep track of its position, shape, and appearance. The standard VOS method seeks to extend the frame's ground truth object masks throughout the video. Recent efforts [5, 13, 22, 25] fall within the matching-based techniques and track the target object using feature matching.

Moving object Segmentation (MOS) - Standard MOS techniques involve identifying and separating moving objects in a video sequence from the background. This can be used in a variety of applications such as video surveillance, self-driving cars, and border security [12, 15, 20]. The process usually involves background subtraction, optical flow estimation, and blob detection techniques to extract moving objects.

Referred video object segmentation (RVOS) RVOS aims to segment specific objects referred by text in the video. In this particular task, learning object representation is affected by the natural language query. Various approaches [2, 3, 23] work in segmenting referred objects frame by frame as the video progresses with time. In the case of referring moving objects, a better understanding of movement is desired. To segment a moving object referred by a text query requires both temporal and visual features information [10]. Hence, Spatio-temporal and linguistic features collectively become essential for the RVOS task. Furthermore, for the RVOS task, the incorporation of language embedding into the database is required, leading to a scarcity of available databases in the literature. RVOS[24] introduces a benchmark dataset and evaluation metrics where a natural language query guides segmentation.

3. Proposed method

This paper introduces a novel framework for Referred Moving Object Segmentation (RefMOS) that leverages an encoder-decoder architecture. Our approach is motivated by the need for robust and efficient segmentation of moving objects in videos guided by textual descriptions. The proposed method consists of three fundamental building blocks to address specific challenges in RefMOS.

3.1. Visual feature extractor (VFE)

The VFE module is responsible for extracting visual features from input frames. This module leverages the DeepLabv3[4] model, which is a semantic segmentation architecture for image segmentation based on the fully convolutional network (FCN) architecture. It uses atrous convolutions to increase the resolution of the output feature maps without increasing the number of parameters or computational cost. DeepLabv3 utilizes the atrous spatial pyramid pooling (ASPP) module to capture both global context and local detail information.

3.2. Inter-Frame Change Detector (IFCD)

The IFCD module makes the encoder network more robust toward generating effective foreground-background probability maps to detect moving objects. Our proposed approach IFCD module is responsible for capturing inter-frame change features. Each input frame is processed through the VFE module to get the visual features. Subtraction is performed on all three permutations of obtained result maps to capture the inter-frame changes. All these subtracted feature maps get added to expose the rich edge information as temporal features. Furthermore, temporal features are concatenated with a visual representation. The working of the IFCD module is shown in the e.q. 1

$$IFCD = \bigoplus \{F0, F1, F2\} \tag{1}$$

$\bigoplus$ indicates the addition of three obtained change feature maps. F is obtained from e.q. 2

$$F_i = VFE_{(i+1)mod2} \, \Theta \, VFE_i ; \quad i \in (0, 1, 2) \tag{2}$$

Θ indicates the element-wise subtraction where paired frames are subtracted with one another.

Obtained IFCD change feature map is concatenated with single frame visual representation as shown in e.q. 3

$$VIFCD = \bowtie \{VFE_i, IFCD\} \tag{3}$$

3.3. Language Feature Extractor (LFE)

To learn the language embedding, LFE leverages the Bidirectional encoder representations from the transformers (BERT[7]) model, which is trained on a large amount of text corpus. BERT is a transformer-based model that understands the meaning of ambiguous language in text by using surrounding text to establish context. Text embedding is extracted and merged with the spatiotemporal (VIFCD) feature map. While merging the linguistic features, first, they are upsampled to the same resolution to match the visual features by concatenating the same 256-size vectors in parallel. Once both domain features' sizes are matched, they are combined with multiplication. The merging of Linguistic feature encoding assists the model in targeting the desired object referred by the text query. A language-guided Spatial-temporal (LST) feature map can be obtained using e.q. 4.

$$LST = \bigotimes \{VICFD, BERT(TEXT)\} \tag{4}$$

Here $\bigotimes$ denotes the multiplication of language embedding with Saptio-temporal features. These LST features are processed through a decoder network to help the model in segmenting a target referred object in the video.

3.4. Overview of proposed architecture

The proposed RefMOS framework assumes three consecutive video frames as input. All three frames are processed through the encoder network to extract visual features. IFCD module works towards capturing inter-frame changes where inter-frame change features are concatenated with the visual representation of a single frame. Furthermore, the spatiotemporal feature map merges with natural language embedding extracted from the language feature extractor and helps the model to segment the targeted moving object. The encoded features are processed through a decoder network to generate a corresponding segmentation map of the referred object. Fig. 2 shows the general overview of the proposed architecture.

3.5. Network losses

We have used cross-entropy loss and SGD approach in the training process for model optimization. The objective of the loss function is defined in e.q. 5.

$$\mathcal{L}_{CE} = \sum_{i=1}^{C=2} t_i \log(s_i) = -t_1 \log(s_1) - (1 - t_1) \log(1 - s_1) \tag{5}$$

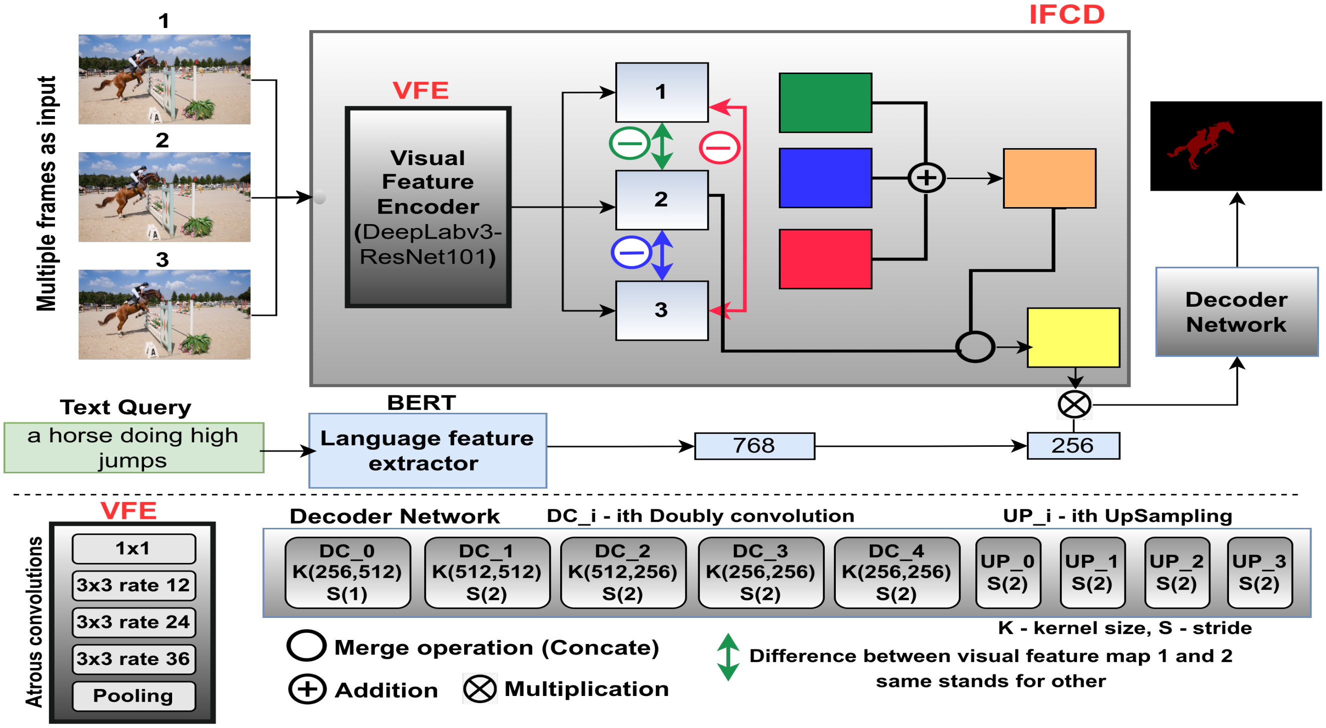

Figure 2. Overview of the proposed framework for RefMOS. Substantial change features between frames are captured using the proposed inter-frame change Detector (IFCD) module. These movement features are concatenated with visual features to learn the visual appearance of moving objects. Linguistic features are extracted from the BERT model and merged with spatio-temporal information to target the referred object by the text query.

where $\mathcal{L}_{CE}$ is cross-entropy loss. In this particular objective, binary cross entropy is utilized. C denotes classes, whereas 1 and 2 denote foreground and background classes, respectively. $t_1 \in \{0, 1\}$, and s_1 are ground truth and the predicted probability for class C_1. t_2 = 1-t_1, and s_2 = 1-s_1 stand for ground truth and predicted probability for class C_2, respectively. Loss is calculated in an iterative manner for each pixel during training. The goal of training is to minimize the loss by using backpropagation and SGD optimization algorithms. Minimizing the loss encourages the model to produce an accurate segmentation mask closely related to the ground truth. For the RefMOS objective, we utilize cross-entropy loss as this effectively measures the difference between probability distribution and is suitable for predicting pixel-wise class labels.

4. Experimental setup

We have evaluated RefMOS on DAVIS-2017 [18] and DAVIS-2016 [16] datasets which are widely used benchmarks in video object segmentation tasks. To assess the performance of the proposed method, quantitative results of the Jaccard and the F-measure similarity index have been compared with state-of-the-art methods. The Jaccard index (J), also known as Intersection over union (IoU), measures the similarity between predicted output and ground truth mask by computing the ratio of Intersection over union among two classes. Jaccard Index can be represented in eq. 6

$$J = \frac{Area\ of\ Overlap}{Area\ of\ Union} = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

where A and B are the sets that represent the predicted output and group truth, and the F-measure index is evaluated as a harmonic mean of precision and recall. It provides a balanced evaluation of the model performance. F-measure can represented as in eq. 7

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{7}$$

where precision is the ratio of predicted true positives (TP) to the total predicted positives (TP + FP), and recall is the ratio of true positives predicted to the total actual positives.

$$precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}$$

4.1. Dataset:

DAVIS-2017 dataset contains 90 videos that serve as crucial benchmarks in VOS. Among those 90 videos, 60 are

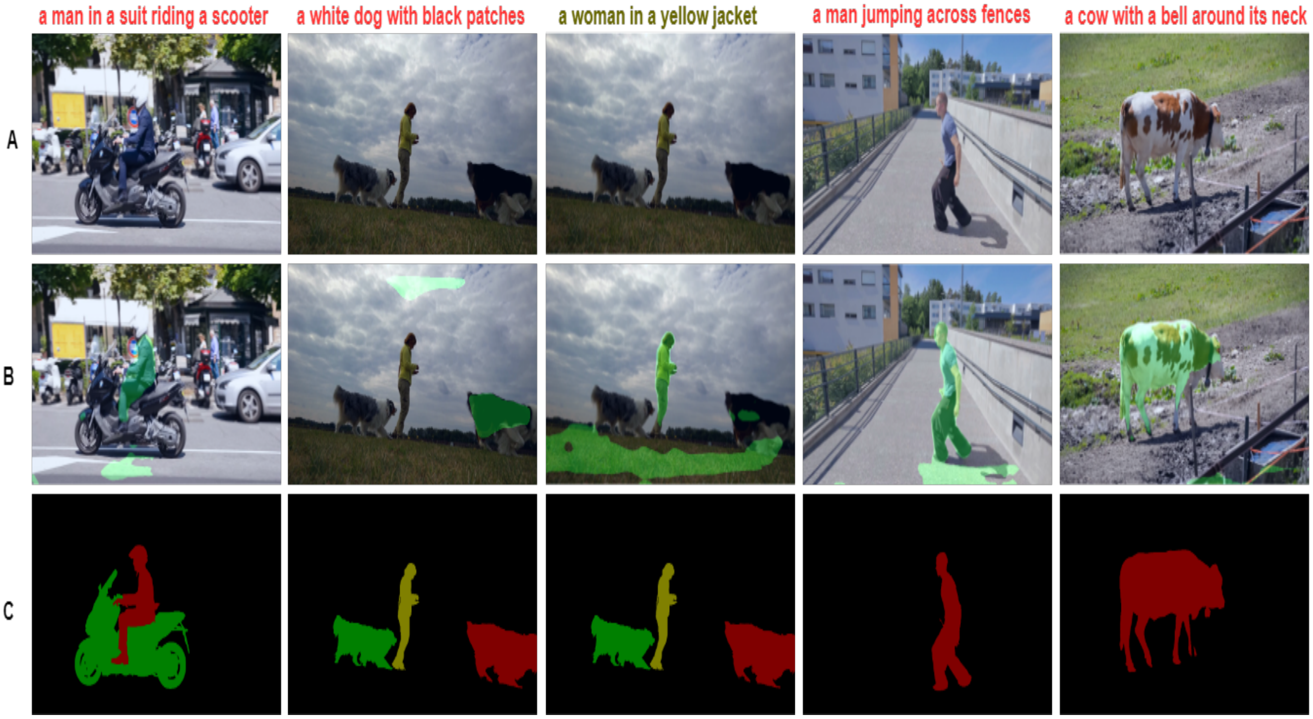
a man in a suit riding a scooter
a white dog with black patches
a woman in a yellow jacket
a man jumping across fences
a cow with a bell around its neck
A
B
C

Figure 3. **Text queries are written on the top of the inputs. Row A: Inputs, Row B: Query-based segmentation of moving objects in green mask superimposed on inputs, C: Ground Truth of moving objects irrespective of text query**.

allocated for the training set, and the remaining 30 videos are categorized for testing purposes. Similarly, in DAVIS-2016, 30 and 20 videos are categorized as training and testing. Moreover, to enhance the utility and make it work for the RefMOS task, both datasets are augmented with Regular expressions (REs). These REs work as text queries to target specific objects; hence, ground-truth maps are dynamically generated for input frames corresponding to the input text query. If a video contains more than one object, then corresponding objects get enabled for masking based on a text query.

4.2. Training details:

We employed the PyTorch-1.12 framework to train our network, leveraging its capabilities for efficient model development. The objective of the RefMOS task is to generate a binary mask that highlights relevant objects in video frames, aiding in tasks like object tracking and segmentation. To enhance the robustness of our model, we applied various data augmentations during training, such as random flips, rotations, random crops, etc. Our training process spanned 50 epochs, during which we saved all checkpoints to ensure the retention of the best weights for future use. We monitored the training process closely, adjusting hyperparameters and model architecture to improve performance.

5. Results and Discussion

We have reported the mean Intersection over Union (J) and average F-measure (F) on the DAVIS-17 and DAVIS-16 datasets to evaluate the segmentation performance of our model. Additionally, we applied post-processing techniques, such as morphological operations, on the feature maps to improve segmentation accuracy and obtain better object boundaries. These techniques help refine the segmentation results by smoothing the boundaries and filling in small gaps or holes in the masks.

5.1. Quantitative Analysis

We evaluate the proposed method on the dataset in Table 1 and Table 2. The quantitative results of RefMOS demonstrate satisfying performance compared to the state-of-the-art methods when evaluated using the Jaccard Index and F-score. These results highlight the robustness and efficacy of our approach in handling complex video object segmentation guided by text query.

5.2. Qualitative Analysis

We perform a qualitative analysis of our proposed method. Fig. 3 shows the output segmented maps in green color superimposed on input frames. As demonstrated in the visual results, it is evident that the RefMOS framework efficiently segments the referred object. In Fig. 3, five examples are shown in five columns. On the top of each column, a text query is written that defines a specific target to segment. The first row, 'A,' contains the input frame. Second row 'B' contains the output segmented map in green superimposed on input to demonstrate the effectiveness of the proposed model. The third row, 'C,' contains the ground truths that show all moving objects irrespective of the targeted object. Visual results show that our model effectively segments the moving object referred to by the text. Row 'C' shows ground truths of all moving objects, but only referred objects need to be segmented based on a text query. Column-2 and column-3 have the same input frame, but output map segments dog and lady, respectively, based on the provided text query, which targets the desired object only.

Methods	Backbone	J	F	J&F
Khoreva et al.[9]	-	-	-	39.3
CMSA [26]	ResNet-50	32.2	37.2	34.7
CMSA+RNN[26]	ResNet-50	36.9	43.5	40.2
URVOS[21]	ResNet-50	47.3	56.0	51.5
RefMOS (Ours)	**ResNet-101**	**48.6**	**53.4**	**51.0**

Table 1. Quantitative results comparison of RefMOS on DAVIS-2017. **J**: Jaccard, **F**: F-measure, **J&F** is average of J and F

Methods	Backbone	J	F	J&F
RefMOS (Ours)	ResNet-101	56.7	59.3	55.1

Table 2. Quantitative results of RefMOS on DAVIS-2016. **J**: Jaccard, **F**: F-measure, **J&F** is average of J and F

5.3. Discussion

Merging text and visual features for referred video object segmentation is a promising yet challenging task. The primary limitations include the complexity of fusing these features due to their different representations and dimensionalities, leading to potential information loss. We fuse the linguistic features by multiplying them with Visual features after upsampling BERT features, which are subject to experimentation. The natural language introduces ambiguity, synonymy, and polysemy, complicating the accurate interpretation of text about visual data. Hence, a better language model and simultaneous training of language models, which is the future scope of the current approach, can lead to better performance. Furthermore, the computational load is significant, requiring resource-intensive neural network architectures and posing challenges for real-time processing. Data annotation is limited, with high-quality annotated datasets being rare and domain-specific. Evaluating the performance of models that combine text and visual features is also challenging, requiring novel metrics that consider both segmentation accuracy and text-visual alignment correctness. Addressing these limitations requires advancements in multi-modal learning techniques, dataset creation, and efficient algorithms. Despite these challenges, the integration of text and visual features holds great potential for enhancing the accuracy and usability of referred video object segmentation systems.

6. conclusion

In this work, we designed a robust referred moving object segmentation (RefMOS) framework to segment the moving object guided by a text query. RefMOS comprises three basic building blocks. The First VFE module utilizes Deeplabv3 architecture, which is responsible for extracting visual features from input frames. Second, Interframe change detector (IFCD) module captures relevant changes in consecutive frames. Obtained IFCD features are concatenated with visual object representation from the input frame to learn spatiotemporal structural dependencies. Third, Language feature extractor (LFE) utilizes the BERT language model to extract linguistic features provided by the text query. The spatiotemporal features from the IFCD module are merged with linguistic features, which assist the model in segmenting referred objects. Furthermore, the cross-entropy loss and SGD optimizer are utilized to train the designed network. We evaluate our method on the DAVIS-2017 and DAVIS-2016 datasets, where our proposed method competes with state-of-the-arts ones.

Acknowledgments

This project was supported by project SERB/2022/000332.

References

[1] Thangarajah Akilan, Qingming Jonathan Wu, Amin Safaei, Jie Huo, and Yimin Yang. A 3D CNN-LSTM-based image-to-image foreground segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 21:959–971, 2019. 2

[2] Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres, and Xavier Giro-i Nieto. Refvos: A closer look at referring expressions for video object segmentation. *arXiv preprint arXiv:2010.00263*, 2020. 2

[3] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF CVPR*, pages 4985–4995, 2022. 2

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the ECCV*, pages 801–818, 2018. 2, 3

[5] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2

[6] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *Proceedings of the IEEE CVPR*, pages 7415–7424, 2018. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[8] Ping Hu, Jun Liu, Gang Wang, Vitaly Ablavsky, Kate Saenko, and Stan Sclaroff. Dipnet: Dynamic identity propagation network for video object segmentation. In *Proceedings of the IEEE/CVF WACV*, pages 1904–1913, 2020. 2

[9] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV 2018*, pages 123–141. Springer, 2019. 6

[10] Dezhuang Li, Ruoqi Li, Lijun Wang, Yifan Wang, Jinqing Qi, Lu Zhang, Ting Liu, Qingquan Xu, and Huchuan Lu. You only infer once: Cross-modal meta-transfer for referring video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1297–1305, 2022. 2

[11] Chen Liang, Yu Wu, Tianfei Zhou, Wenguan Wang, Zongxin Yang, Yunchao Wei, and Yi Yang. Rethinking cross-modal interaction from a top-down perspective for referring video object segmentation. *arXiv preprint arXiv:2106.01061*, 2021. 2

[12] Murari Mandal, Prafulla Saxena, Santosh Kumar Vipparthi, and Subrahmanyam Murala. Candid: Robust change dynamics and deterministic update policy for dynamic background subtraction. In *IEEE/ICPR*, pages 2468–2473, 2018. doi: 10.1109/ICPR.2018.8545504. 1, 2

[13] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF ICCV*, pages 9226–9235, 2019. 2

[14] Prashant W Patil and Subrahmanyam Murala. MSFgNet: A novel compact end-to-end deep network for moving object detection. *IEEE Transactions on Intelligent Transportation Systems*, 20:4066–4077, 2018. 2

[15] Prashant W Patil, Kuldeep M Biradar, Akshay Dudhane, and Subrahmanyam Murala. An end-to-end edge aggregation network for moving object segmentation. In *proceedings of the IEEE/CVF CVPR*, pages 8149–8158, 2020. 2

[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 *proceedings of the CVPR*, 2016. 4

[17] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE CVF/CVPR*, pages 724–732, 2016. 2

[18] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2, 4

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: Munich, Germany, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[20] Prafulla Saxena, Kuldeep Biradar, Dinesh Kumar Tyagi, and Santosh Kumar Vipparthi. Richex: A robust inter-frame change exposure for segmenting moving objects. In *IEEE CVF/ICIP*, pages 2172–2176. IEEE, 2022. 2

[21] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV 2020*, pages 208–223. Springer, 2020. 6

[22] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF CVPR*, pages 9481–9490, 2019. 2

[23] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF CVPR*, pages 4974–4984, 2022. 2

[24] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2, 3

[25] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV 2020*, pages 332–348. Springer, 2020. 2

[26] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF CVPR*, June 2019. 6