A Comparative Analysis of Video Object Segmentation Models

Raahim Siddiqi*
Department of Computer Science
FAST-NUCES
Karachi, Pakistan
k200146@nu.edu.pk

Syed Ahmed Mahmood*
Department of Computer Science
FAST-NUCES
Karachi, Pakistan
k200153@nu.edu.pk

Mahad Hameed*
Department of Computer Science
FAST-NUCES
Karachi, Pakistan
k200338@nu.edu.pk

***Abstract*—This research paper provides a comparative analysis of popular Video Object Segmentation (VOS) models introduced till Dec. 2022 to identify the state-of-the-art techniques in the field. A total of 5 Models are compared, OSVOS, FEELVOS, BubbleNets, XMem, and BATMAN (listed in chronological order) on their performance on the DAVIS-2017 challenge dataset. This study finds that XMem and BATMAN models, both of which were introduced in 2022, achieved the best performance with J-F scores of 86.2, and became state-of-the-art models for VOS.**

***Index Terms*—Video Object Segmentation, OSVOS, FEELVOS, BubbleNets, XMem, BATMAN, DAVIS-2017 dataset, J-F score**

I. Introduction

Video object segmentation (VOS) is a challenging task in computer vision that involves detecting and segmenting objects in a video sequence. VOS has gained significant attention in recent years due to its numerous practical applications, ranging from video editing and surveillance to autonomous driving and robotics. With the increasing availability of high-resolution cameras and sophisticated imaging technologies, the demand for accurate and efficient VOS algorithms has also grown.

One of the main applications of VOS is in video editing, where it allows for the separation of foreground and background elements, enabling the creation of special effects and compositing. In the field of surveillance, VOS can be used to detect and track objects of interest, such as individuals or vehicles, and to extract relevant information from video feeds. Additionally, VOS is a critical component of autonomous driving systems, where it enables the detection and tracking of other vehicles, pedestrians, and obstacles in real-time.

Overall, the field of VOS is rapidly evolving, with new algorithms and approaches being developed and refined regularly. As such, this research paper aims to provide an overview of the current state-of-the-art techniques in VOS, their strengths and limitations, and potential directions for future research.

A. Related Work

Video Object Segmentation (VOS) has been a topic of research for many years. In recent years, the development of deep learning techniques has led to significant advancements in VOS models. This section provides a chronological-based review of related works on VOS models, and how they have solved various challenges in the field of VOS.

VOS models typically fall into two categories: (1) memory-based methods and (2) propagation-based methods. Memory-based methods use a feature memory to store information given in the first frame and to segment any new frames. On the other hand, propagation-based methods perform frame-to-frame propagation and are thus efficient at test-time. Both methods have their advantages and disadvantages, and researchers have proposed various models to address these limitations.

This paper will be using the One-Shot Video Object Segmentation (OSVOS) model (2017) as our baseline model. At the time it was introduced, it achieved state-of-the-art performance at the time [9]. OSVOS used a fully convolutional network (CNN) to perform per-pixel classification, which was then refined using a recurrent neural network [9]. However, there were many shortcomings in this model, which later models have addressed.

In 2018, the FEELVOS model was proposed, which used a feature enhancement module to improve the speed of the model significantly while maintaining accuracy without the need of heavy fine-tuning [21]. FEELVOS achieved real-time performance, making it possible to use VOS in real-time applications with acceptable accuracy.

In 2019, a novel VOS model called BubbleNets was proposed. Unlike previous VOS models which relied on the first-frame for segmentation, BubbleNets selected the best-annotated frame to guide the segmentation process [5]. BubbleNets also used a gating mechanism to suppress irrelevant information and reduce the computational cost [5].

XMem [1], which was unveiled in July 2022, employs several memory stores to capture various temporal contexts, while ensuring that GPU memory usage remains tightly bounded by means of long-term memory consolidation. Thanks to its innovative approach, XMem outperformed all other methods in terms of both accuracy and speed [1]. In August 2022, BATMAN was proposed, which uses a bi-directional attention module to capture both spatial and temporal information [17]. BATMAN also uses a feature alignment module to better match features across frames, improving accuracy [17].

In summary, the development of VOS models has come a long way, with researchers proposing various models to improve accuracy and speed. Memory-based and propagation-based methods both have their advantages and disadvantages, and recent models have attempted to address the limitations of both methods. XMem and BATMAN are the most recent state-of-the-art models, showing significant improvements in both accuracy and speed.

B. Problem Statement

Over the last few years, several VOS models have been proposed in the literature, each with its own strengths and weaknesses. However, there is a lack of comprehensive comparative studies that evaluate the performance of these models using a common set of metrics and datasets. Previous studies are outdated, and do not include the recent state-of-the-art models [26].

C. Objectives

In this review paper, we aim to compare and evaluate multiple VOS models to identify the state-of-the-art techniques in the field.

D. Research Question

What are the strengths and weaknesses of popular video object segmentation models, and how do they compare in terms of accuracy, performance, and speed to different types of video content, as evaluated on the benchmark dataset DAVIS 2017?

E. Significance Study

Our comparative analysis on Video Object Segmentation Models will help researchers and practitioners in the field gain insights into the performance of these models and make informed decisions when selecting the appropriate VOS model for their specific applications.

F. Limitations

This research paper has some limitations that must be considered. Firstly, the comparison is limited to only five popular VOS models, and therefore, it does not cover all the available models in the literature. The selection of these models was based on their popularity and availability. Consequently, some models that may have performed better than the selected models may have been excluded.

Secondly, the evaluation of the models was conducted on a particular dataset. While the chosen dataset (DAVIS 2017) is widely used in the scientific community as a benchmark for VOS models, it may not cover all the possible scenarios that can be encountered in real-world applications. Therefore, the results may not be generalizable to all possible scenarios.

Finally, the comparison is limited to the performance metrics used in this research, and other metrics, such as computational efficiency and memory usage, have not been considered. Due to lack of funds and hardware, this paper relies completely on metrics acquired from other research papers, and we were unable to perform additional tests or produce any of our own results.

G. Methodology

The data used in this study was obtained from the website PapersWithCode, which provided access to the relevant research papers and code implementations. The study spanned a week and involved gathering and analyzing data from various VOS models.

To begin the study, a list of relevant VOS models was compiled based on their popularity and recent advancements in the field. Papers and code implementations for each model were then obtained from PapersWithCode. Next, the evaluation metrics used in each paper were identified, and the results for each model were recorded. The evaluation metrics that we consider in this paper are the Jaccard Index ($\mathcal{J}$) and the F-measure ($\mathcal{F}$), which are standard metrics used to evaluate the performance of VOS models.

Our study will be using the DAVIS-2017 Challenge dataset, one of the widely used datasets for benchmarking VOS Model performance.

II. Video Object Segmentation Models

A. OSVOS Model

Most current video object segmentation methods enforce temporal consistency to propagate the initial mask into subsequent frames, which can be computationally expensive. These methods often use superpixels [11], patches [12], or object proposals [13] to reduce complexity. Some methods also require the computation of optical flow, which further slows down the process. Other methods incorporate deep learning techniques to refine masks frame by frame. However, OSVOS (One-Shot Video Object Segmentation) simplifies the pipeline by segmenting each frame independently, producing more accurate results in a significantly faster manner. OSVOS is a fully-convolutional neural network architecture for semi-supervised video object segmentation [9].

OSVOS starts with a pre-trained base CNN that is trained on ImageNet [14] for image labeling. The results of this pre-trained network are useful for segmenting some general image features, but not specific objects. Next, OSVOS trains a parent network on the training set of DAVIS(2016). This helps improve the segmentation results, but still doesn't focus on a specific target object. Finally, by fine-tuning the network on a single frame segmentation example for the specific target object, the network rapidly focuses on that object. This fine-tuning step helps the network learn more specific features and improve the accuracy of the segmentation results for that object.

The major contribution of OSVOS is the ability to work at various points in the trade-off between speed and accuracy. The user can choose the level of fine-tuning for a faster method or more accurate results. Furthermore, OSVOS is able to handle occlusions, various ranges of motion, and interlaced videos.

e-OSVOS(Efficient OSVOS) [10] is an improvement upon the previous OSVOS method, which is a semi-supervised approach for segmenting objects in each frame of a video. The main drawback of OSVOS is that it requires fine-tuning [15] the segmentation model separately for each given object mask at test time, which is time-consuming and computationally expensive.

e-OSVOS uses an encoder-decoder network with a memory module to segment objects in video frames. The memory module allows E-OSVOS to store relevant information from previous frames and reuse it in subsequent frames, reducing the need for the network to process all frames independently. This leads to a significant reduction in computational cost while maintaining high segmentation accuracy [16].

B. BubbleNets Model

The innovative BubbleNets model for video object segmentation developed by Gryphon and Corso utilizes a deep sorting frames (DSF) module to organize input frames according to their relevance to the foreground object. The DSF module employs self-attention to select the most informative frames, which are then ranked according to their relevance.

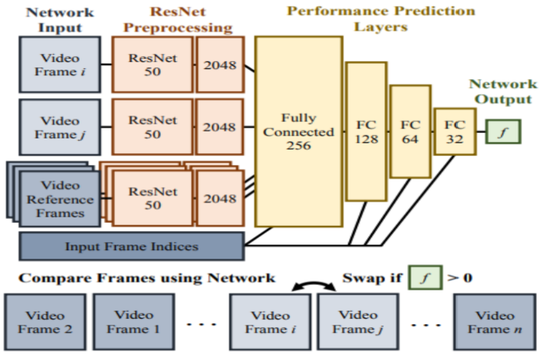


Fig. 1. BubbleNets Framework.

This procedure generates a set of frames that are suitable for training a convolutional neural network (CNN) to generate accurate object approximations. The ever-growing feature memory bank of earlier memory-based VOS models, which causes the model to substantially slow down when playing longer videos [5] is one issue that this method effectively resolves.

The DSF module enables the network to effectively employ the most instructive frames in a video sequence, thereby enhancing performance and generating more accurate object masks.

However, the method may be computationally intensive and may not perform as well with more complex or diverse video sequences [6]. The research papers ”FusionSeg: Learning to Combine Motion and Appearance for Fully Automatic Segmentation of Generic Objects in Videos” [7] and ”Siamese Instance Search for Tracking” [8] are cited by the authors and provide additional context for understanding the state of the art in this field.

C. FEELVOS Model

Various popular approaches for video object segmentation (VOS) in recent years are complex, slow, and ”heavily rely on fine-tuning” the model using the annotations from the first frame making them unsuitable for the majority of practical use cases [21]. Many of these methods also rely on ”extensive engineering”, which results in high system complexity with many components, like the winner of the 2018 DAVIS challenge, PReMVOS [22], [23], [24], which produced remarkable results but was of limited realistic usage.

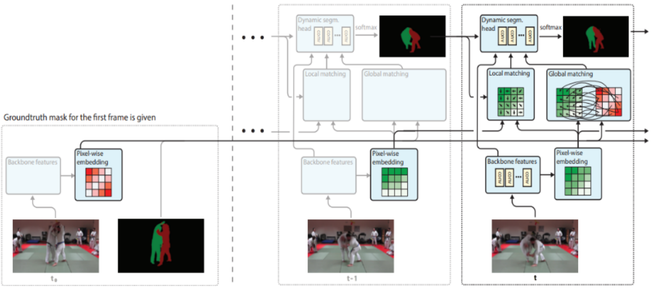


Fig. 2. Overview of the proposed FEELVOS method. In order to segment the image of the current frame, backbone features and pixelwise embedding vectors are extracted for it. Afterwards the embedding vectors are globally matched to the first frame and locally matched to the previous frame to produce a global and a local distance map. These distance maps are combined with the backbone features and the predictions of the previous frame and then fed to a dynamic segmentation head which produces the final segmentation.

FEELVOS is presented as a simple and fast method that does not rely on fine-tuning and instead it uses a ”semantic pixel-wise embedding together with a global and a local matching mechanism” for each frame to relay information from the initial frame and from the previous frame of the video to the current frame [21]. They also use a novel procedure called the ”dynamic segmentation head” to train the model, to segment various objects in a video [21].

FEELVOS is inspired by the Pixel-Wise Metric Learning (PML) [25] process and takes some ideas from it to allow FEELVOS to be simple and fast, but unlike PML it does not use it for the final segmentation decision, allowing it to ”learn the embedding in an end-to-end way and the network can recover from partially incorrect nearest neighbor assignments and still produce accurate segmentations” [21].

FEELVOS achieves fast inference speeds by leveraging an efficient online fine-tuning strategy that avoids the need to store the entire feature memory. Instead of storing all the features, FEELVOS only stores the most relevant features, which allows it to achieve comparable or even better results than previous methods while considerably reducing the computational and memory costs. Additionally, FEELVOS uses a novel feature-enhanced cross-entropy loss function that encourages the model to focus on the most important, and informative regions of the input frames, further improving its efficiency and accuracy [21].

D. BATMAN Model

Previous Video Object Segmentation (VOS) models faced challenges in matching features between frames in a video without any class annotations. VOS models need to match objects between frames without considering their specific classes, so that they can propagate these segmentation masks.. Previous models attempted to establish correspondence between frames using global attention mechanisms [18], which can fail to distinguish target objects from the background if numerous objects have a similar appearance. Some models also leveraged optical flow to capture the motion of the object [19], but this can be noisy and accumulate errors along the video sequence. As a result, previous VOS models struggled to accurately segment objects in complex video sequences.

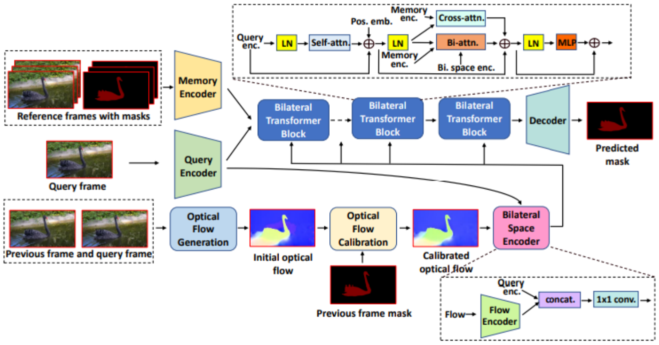


Fig. 3. Overview of the BATMAN architecture. Frame-level features of the reference frames and the query frame are extracted through the memory and query encoders, respectively. A pre-trained FlowNet is used to generate an initial optical flow estimation between the previous frame and the query frame, which is then improved by the optical flow calibration module. A bilateral space encoder is used to encode the query features and the calibrated optical flow into a bilateral space encoding, which is used by the bilateral attention. Multiple layers of bilateral transformer blocks are stacked for matching the correspondence between the reference and query features. Lastly, a decoder is used to predict the query frame segmentation mask.

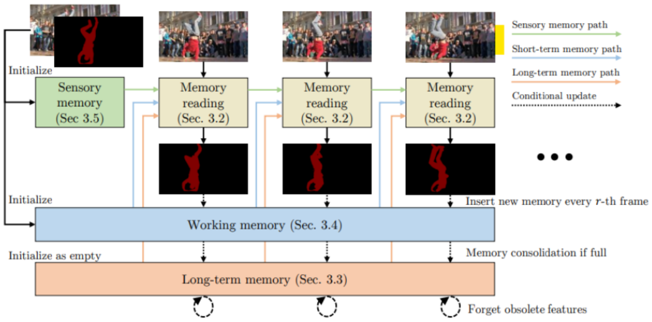
Bilateral Attention Transformer in Motion-Appearance Neighboring space (BATMAN) is a transformer based model [20] which applies bilateral attention for VOS. It addresses the limitations of previous VOS models by capturing the motion of the object in the video using a novel optical flow calibration module. This module merges the segmentation mask with optical flow estimation to improve the smoothness of optical flow within objects and reduce noise at object boundaries. Bilateral Attention Transformer in Motion-Appearance Neighboring space (BATMAN) is a transformer-based model [20] that uses bilateral attention for VOS. It overcomes the limitations of previous VOS models by capturing object motion in videos through a novel optical flow calibration module. This module merges segmentation masks with optical flow estimation to improve the smoothness of optical flow within objects and reduce noise at object boundaries. The calibrated optical flow is then used in BATMAN's novel bilateral attention mechanism, which matches the query and reference frames in the adjacent bilateral area., taking into account both the motion of the object and its appearance. This approach allows BATMAN to better handle challenging scenarios, such as occlusions and motion blur.

BATMAN is composed of two core modules.

1) A novel bilateral attention module that helps match object features better by focusing on important features and ignoring the background noise. It works by calculating attention between the query and memory features in the bilateral space of motion and appearance [17].
2) A novel optical flow calibration module to improve the accuracy of optical flow estimation. This technique uses a combination of the initial optical flow estimation and the object segmentation mask to improve the flow estimation and reduce noise at the boundary of the object [17].

E. XMem with Atkinson-Shiffrin Memory Model

Prior memory-based approaches in VOS models have suffered from a variety of issues. Tracking-based approaches lacked long-term context and often could not handle occlusions because of a shortage of long-term memory. [4]. Methods that "remembered" past frames as feature memory could not process larger videos [3], which would cause the model to slow down drastically(videos with several hundred frames or more).



Initialize
Sensory memory (Sec 3.5)
Memory reading (Sec. 3.2)
Memory reading (Sec. 3.2)
Memory reading (Sec. 3.2)
Sensory memory path
Short-term memory path
Long-term memory path
Conditional update
Initialize
Insert new memory every r-th frame
Working memory (Sec. 3.4)
Initialize as empty
Memory consolidation if full
Long-term memory (Sec. 3.3)
Forget obsolete features

Fig. 4. Overview of XMem. The memory reading operation extracts relevant features from all three memory stores and uses those features to produce a mask. To incorporate new memory, the sensory memory is updated every frame while the working memory is only updated every r-th frame. The working memory is consolidated into the longterm memory in a compact form when it is full, and the long-term memory will forget obsolete features over time.

XMem is presented as a semi-supervised Video Object Segmentation architecture for long videos using multiple feature memory stores inspired by the Atkinson-Shiffrin memory model [2]. Based on the human memory, this model incorporates three independent yet deeply-connected feature memory stores [2]:

1) A volatile short term memory.
2) A high end processing memory.
3) A Long-term sustainable memory.

The working memory model stores "high-resolution" features, but only for a few seconds [1]. This keeps the memory usage limited. The long term memory stores compressed features from the working memory, allowing it to keep these features in memory for up to 10,000 frames without memory issues. This solves both the problem of remembering long-term features, and the memory-explosion issue [1].

Despite the rapidly updating sensory memory, the XMem model fails to detect and track fast moving objects [1].

One point of interest is the fast performance of the XMem model, which under reasonably powerful hardware (RTX 2080

Ti) was able to perform at 22.6 FPS on the DAVIS set [1]. This performance is almost close enough for real-time applications.

III. Discussion

The field of Video Object Segmentation (VOS) is rapidly evolving, with new algorithms and approaches being developed and refined regularly. While previous studies have compared VOS models [26], with the recent introduction of newer models, such as XMem and BATMAN models in late 2022, there is a need for a new comparative analysis that evaluates the performance of the latest VOS models. This research paper aims to fill this gap by comparing popular VOS models, including XMem and BATMAN, using the DAVIS-2017 challenge dataset.

A. Evaluation Metrics

Jaccard Index ($\mathcal{J}$): Also known as Intersection over Union (IoU), Jaccard Index measures the similarity between two sets of data. In the case of video object segmentation (VOS) evaluation, it is used to measure the similarity between the ground-truth object mask and the predicted object mask generated by the VOS model. The Jaccard Index value ranges between 0 and 1, where 1 means a perfect overlap between the predicted and ground-truth masks. The Jaccard Index is important because it provides a quantitative measure of how well the model performs in accurately predicting the object mask.

F-measure ($\mathcal{F}$): The F-measure can be defined as the harmonic mean of precision and recall, two important metrics in classification problems. Precision measures the proportion of true positives (correctly identified object pixels) to the number of predicted object pixels. Recall measures the proportion of true positives to the total number of ground-truth object pixels. The F-measure balances precision and recall and is a widely used metric in VOS evaluation. It ranges between 0 and 1, where 1 indicates perfect performance. The F-measure is important because it provides a single metric that summarizes the overall model's performance in terms of both precision and recall.

J-F score ($\mathcal{J}$&$\mathcal{F}$): The J-F score is defined as a combination of Jaccard Index and F-measure metrics. It is calculated as the geometric mean of these two metrics and ranges between 0 and 1, where 1 indicates perfect performance. The J-F score is important because it provides a balanced evaluation of the VOS model performance in terms of both spatial and temporal accuracy.

B. Results and Analysis

The below results represent the performance of these 5 models on the DAVIS-2017 validation dataset.

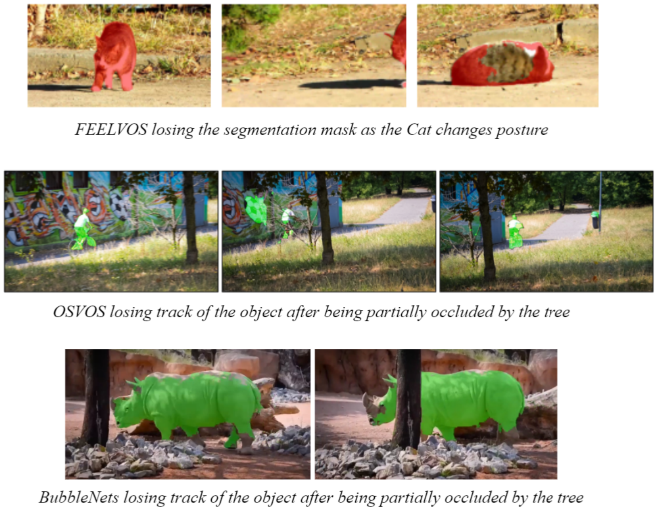
TABLE I
Performance comparison of VOS models on DAVIS-2017 validation dataset

Model	$\mathcal{J}$&$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
XMem [1]	**86.2**	**82.9**	**89.5**
BATMAN [17]	86.2	83.2	89.3
e-OSVOS (2020) [10]	77.2	74.4	80.0
FEELVOS [21]	71.5	69.1	74.0
BubbleNets [5]	62.6	59.7	65.5
OSVOS (2017) [10]	60.3	56.6	63.9

These results indicate that XMem and BATMAN are two highly successful models, and have achieved state-of-the-art performance. This success is supported by the novel features they have implemented, tackling major problems which previous models have suffered from, including:

1) Motion Blur
2) Occlusion
3) Long term dependencies and memory
4) GPU Memory Usage (XMem only)

All of the other models (OSVOS, FEELVOS, and BubbleNets) also suffer from the above issues to a significant extent. Here are some examples.


FEELVOS losing the segmentation mask as the Cat changes posture

OSVOS losing track of the object after being partially occluded by the tree

BubbleNets losing track of the object after being partially occluded by the tree

Fig. 5. Issues in OSVOS, FEELVOS, and BubbleNets.

Although these problems cannot be eliminated completely and exist in every model to a certain degree, the XMem and BATMAN models results prove that they have handled them quite well.

IV. Conclusion

As of 2022, XMem and BATMAN are the state-of-the-art models to be used in the field of VOS for general applications. These models are close enough in accuracy to be used interchangeably, however there may be certain applications where one model performs better than the other.

After accuracy, the most important benchmark in comparing VOS models is their speed. Unfortunately, the speed benchmarks for BATMAN are not included in its research paper, neither is the code for their accuracy benchmarks publicly available. On the other hand, XMem model is rated to perform around an average of 20 FPS [1] (hardware-dependent) and shows considerable improvement over previous and existing models .

In the absence of these results from BATMAN, it is more prudent to declare XMem the better and well-rounded model for Video Object Segmentation.

V. Further Improvements

Despite recent advancements, there are still many areas of improvement in VOS models. The accuracy of the XMem and BATMAN are below 87% which may not be acceptable in critical applications.

There is one key-area of improvement where both XMem and BATMAN have not succeeded in accurately segmenting objects. Both of these models are unable to segment fast moving objects in videos. In the case of the XMem model, it is simply unable to track the fast moving objects. In the BATMAN model, the fast-moving objects are not detected either, but they also disrupt the segmentation masks for nearby objects due to a disruption in the optical flow that the model uses.

References

[1] H. K. Cheng and A. G. Schwing, "XMEM: Long-term video object segmentation with an Atkinson-Shiffrin Memory Model," *arXiv.org*, 18-July-2022. [Online]. Available: https://arxiv.org/abs/2207.07115v2.

[2] R.C Atkinson and R.M Shiffrin, "HUMAN MEMORY: A PROPOSED SYSTEM AND ITS CONTROL PROCESSES!," 1968 [Online]. Available:https://app.nova.edu/toolbox/instructionalproducts/edd8124/articles/1968-Atkinson_and_Shiffrin.pdf.

[3] S. W. Oh, J. Y. Lee, K. Sunkavalli, and S. J. Kim, "Video Object Segmentation using Space-Time Memory Networks," Aug. 2019 [Online]. Available:https://arxiv.org/pdf/1904.00607.pdf.

[4] S. W. Oh, J. Y. Lee, K. Sunkavalli, and S. J. Kim, "Fast Video Object Segmentation by Reference-Guided Mask Propagation," Jun. 2018 [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/papers/Oh_Fast_Video_Object_CVPR_2018_paper.pdf.

[5] B. A. Griffin and J. J. Corso, "BubbleNets: Learning to Select the Guidance Frame in Video Object Segmentation by Deep Sorting Frames," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10718–10727. [Online]. Available: https://arxiv.org/pdf/1903.11779v2.pdf.

[6] H. Zhang, X. Xu, and J. Jia, "Proposal-generation-network-based Video Object Segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6755–6764.

[7] X. Li et al., "FusionSeg: Learning to Combine Motion and Appearance for Fully Automatic Segmentation of Generic Objects in Videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2117–2126.

[8] N. Wang and D.-Y. Yeung, "Siamese Instance Search for Tracking," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1420–1429.

[9] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixe, D. Cremers, and L. Van Gool, "One-Shot Video Object Segmentation," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 221-230. doi: 10.1109/CVPR.2017.36.

[10] T. Meinhardt and L. Leal-Taixé, "Make One-Shot Video Object Segmentation Efficient Again," Dec. 2020 [Online]. Available: https://arxiv.org/pdf/2012.01866v1.pdf.

[11] J. Chang, D. Wei, and J. W. Fisher III, "A video representation using temporal Superpixels," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[12] S. A. Ramakanth and R. V. Babu, "SeamSeg: Video object segmentation using patch seams," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[13] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung. Fully connected object proposals for video segmentation. In ICCV, 2015.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

[15] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. IEEE Conf. on Computer Vision and Pattern Recognition, 2017.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[17] Y. Yu, J. Yuan, G. Mittal, L. Fuxin, and M. Chen, "BATMAN: Bilateral Attention Transformer in Motion-Appearance Neighboring Space for Video Object Segmentation," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Aug. 2022. Available: https://paperswithcode.com/paper/batman-bilateral-attention-transformer-in.

[18] Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9225–9234 (2019).

[19] Xie, H., Yao, H., Zhou, S., Zhang, S., Sun, W.: Efficient regional memory network for video object segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1286–1295 (2021).

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv.org*, 06-Dec-2017. [Online]. Available: https://arxiv.org/abs/1706.03762.

[21] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen, "FEELVOS: Fast End-to-End Embedding Learning for Video Object Segmentation," arXiv:1902.09513v2 [cs.CV], Apr. 2019.

[22] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS:Proposal-generation, refinement and merging for video object segmentation. In ACCV, 2018.

[23] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS:Proposal-generation, refinement and merging for the YouTube-VOS challenge on video object segmentation 2018. The 1st Large-scale Video Object Segmentation Challenge - ECCV Workshops, 2018.

[24] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS:Proposal-generation, refinement and merging for the davis challenge on video object segmentation 2018. The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops, 2018.

[25] Y. Chen, J. Pont-Tuset, A. Montes, and L. Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In CVPR, 2018.

[26] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 Davis Challenge on Video Object Segmentation," *arXiv.org*, 01-Mar-2018. [Online]. Available: https://arxiv.org/abs/1704.00675.