

AI VIET NAM – RESEARCH TEAM

DEEP LEARNING FOR VIDEO OBJECT SEGMENTATION: A REVIEW

April 20, 2024

Date of publication:	08/04/2022
Authors:	Mingqi Gao, Feng Zheng, James J.Q.Yu, Caifeng Shan, Guiguang Ding, Jungong Han
Sources:	Artificial Intelligence Review (2023) 56:457–531
Data sources (if any):	VOS - REVIEW
Keywords:	VOS (Video Object Segmentation), Deep Learning, CNN (Convolutional neural network)
Summary by:	Nhi Ng Thao

1. Intro: Opening summary:

- Detecting object recognition in video is one of the computer vision applications, with object detection in video being one of the most supported and intensively researched applications. This article aims to systematically knowledge and document object detection in videos using deep learning.
- Point out the advantages and disadvantages of each type of method. Through the content introducing definitions, fundamental concepts and basic ideas of algorithms in this field. Next is a summary of the data sets for training and testing object recognition algorithms in videos, as well as the disadvantages of those algorithms and popular evaluation metrics.
- Based on the quantitative and qualitative results of a number of representative methods on a data set, what are the disadvantages to provide and analyze for further discussion on future research directions. Overall, this article aims to quickly grasp current progress in the research field of object recognition in video.

2. Abstract of the article:

- Video object segmentation (VOS) is the task of separating foreground regions from the background in a video sequence. Similar to object tracking, VOS methods establish correspondence of identical objects across frames, but more detailed object representation can be achieved. So VOS, plays an important role in practice, such as visual surveillance, action recognition, summarization and video editing.
- Initially VOS was based on manual features, and objectivity, projective flow, visual saliency were techniques for detecting objects in video. But these are just previous techniques, with modern developments, VOS methods based on deep learning return high results and performance in terms of results and accuracy. So recent methods are based on deep learning neural networks.

- Current VOS methods are grouped and divided into four categories: Unsupervised, semi-supervised, interactive, and language-guided.
 - **Unsupervised segmentation:** This method does not require labeled training data.
 - **Semi-supervised segmentation:** This method combines both labeled and unlabeled training data. It uses information from both data types to create segmented areas.
 - **Interactive segmentation:** This method involves user interaction or additional information.
 - **Segmentation based on linguistic instructions:** This method uses information from linguistic instructions to segment objects.
 - **This article mainly focuses on two widely researched types:** Unsupervised VOS (UVOS) and Semi-supervised VOS (SVOS).
 - UVOS methods perform segmentation without any prior or ground truth labels (unsupervised setting). Objects with prominent motion patterns or visual salience can be recognized.
 - SVOS methods start with ground truth labels available within a few frames (usually just the first frame, semi-supervised setting). These labels are manually annotated to indicate the objects that will be segmented from the remaining frames.
- To avoid conceptual confusion, it is worth mentioning that some recent works refer to unsupervised/semi-supervised VOS as automatic/semi-automatic VOS or zero-shot/one-shot VOS.
- An example comparing these two types of studies: Both methods take raw video as input.

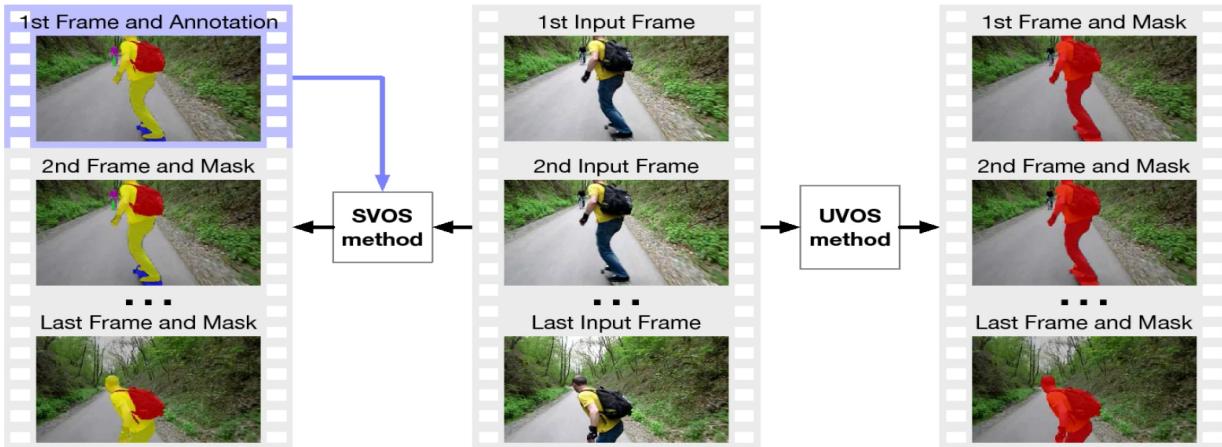


Figure 1: Illustrates the difference between the two VOS methods.

The UVOS method identifies objects with dominant or prominent motion. In contrast, target objects (those that need segmentation) in SVOS depend on human annotation in the first frame (highlighted in purple). Therefore, SVOS methods have more flexibility in identifying target objects.

- **Summary of the purpose of the article's research:**

- Provides evaluation and analysis of datasets beneficial for training and evaluating UVOS and SVOS methods.
- Groups the existing UVOS and SVOS methods into six categories according to the use of spatial and temporal features and provides an in-depth and organized assessment of

their origin, historical development, architecture, advantages, their disadvantages and methods of representation.

- Discusses the performance of the considered methods by analyzing the published evaluation results on several benchmark datasets and testing some representative techniques on other types of challenging video sequences together.
- Summarizes some development trends of the considered methods and draws some predictions about possible future advances.

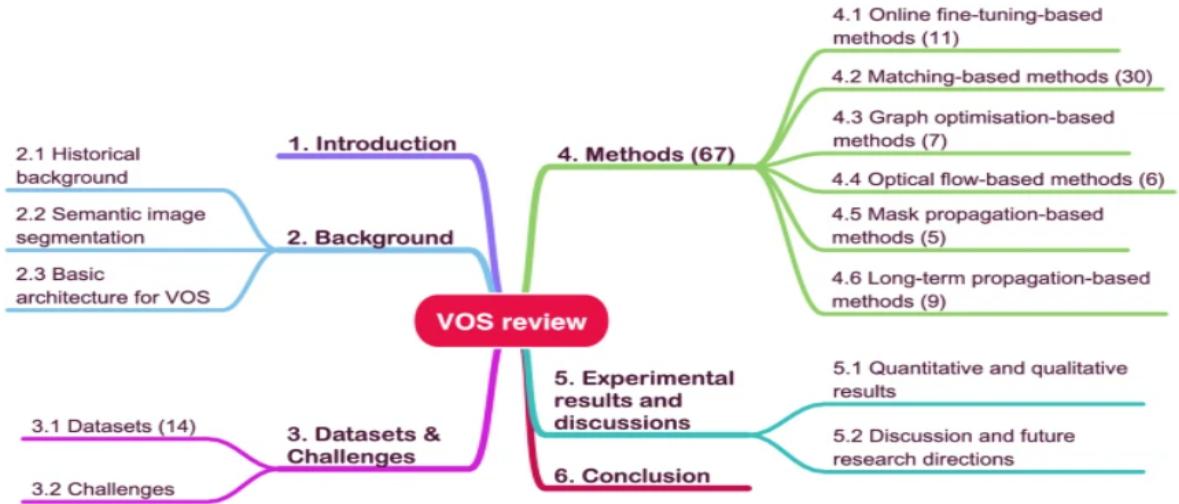


Figure 2: Visual table of contents of the article

3. Context

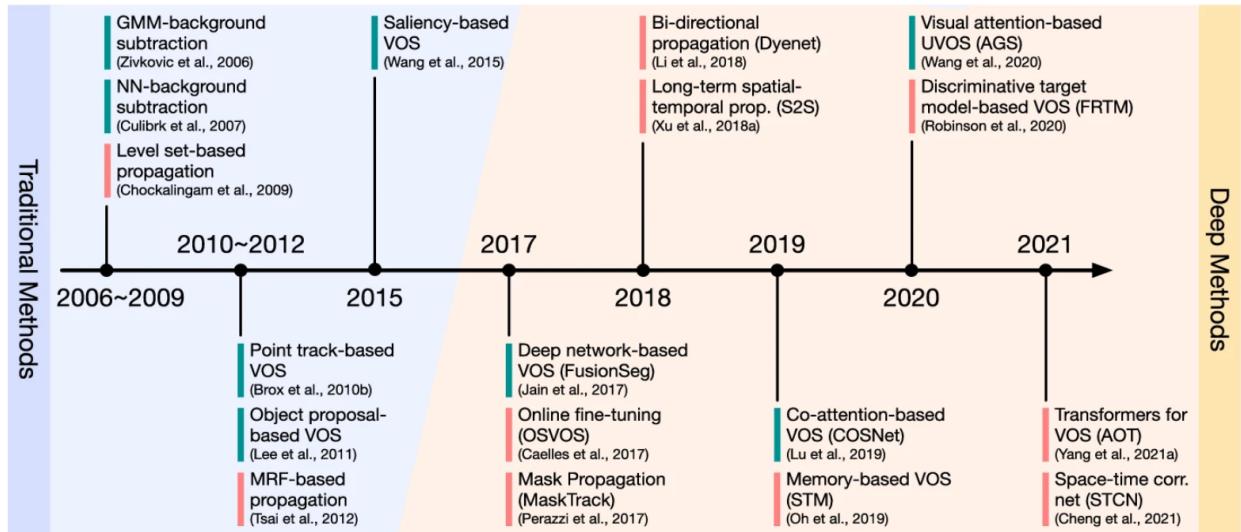


Figure 3: Brief historical context in the field of VOS.

• VOS historical context

- Initially focused on extracting moving objects from video sequences using background extraction techniques, such as building background models and subtracting them from the current frames.

- Developed after successful object proposal generation methods, Proposal-based Object prediction methods involve generating proposals for video frames and ranking them to identify objects. repeats, although slower due to poor performance.
- Realizing the importance of temporal continuity, these Point Trajectory-based Methods construct point trajectories based on motion information and cluster them for segmentation, improving accuracy by looking at consider local and global trajectory information.
- Along with automatic segmentation, SVOS (Semi-supervised VOS) methods aim to propagate annotated masks/indications to other frames, focusing on feature characterization and temporal correlation. time before the emergence of deep learning-based methods.

- **Semantic image segmentation**

- Classifies each pixel into predefined semantic categories based on encoded features.
- To successfully apply CNN to image segmentation, several changes were made, in which a fully convolutional network appeared.
- FCN can take input of arbitrary size and produce output of corresponding size. During training, the FCN is initialized with weights pre-trained on ImageNet and then adjusted on a segmentation dataset such as PASCAL. FCN generates a set of probabilities for each pixel in the input image, indicating the likelihood of the pixel belonging to semantic categories.

- **Basic architecture of VOS:**

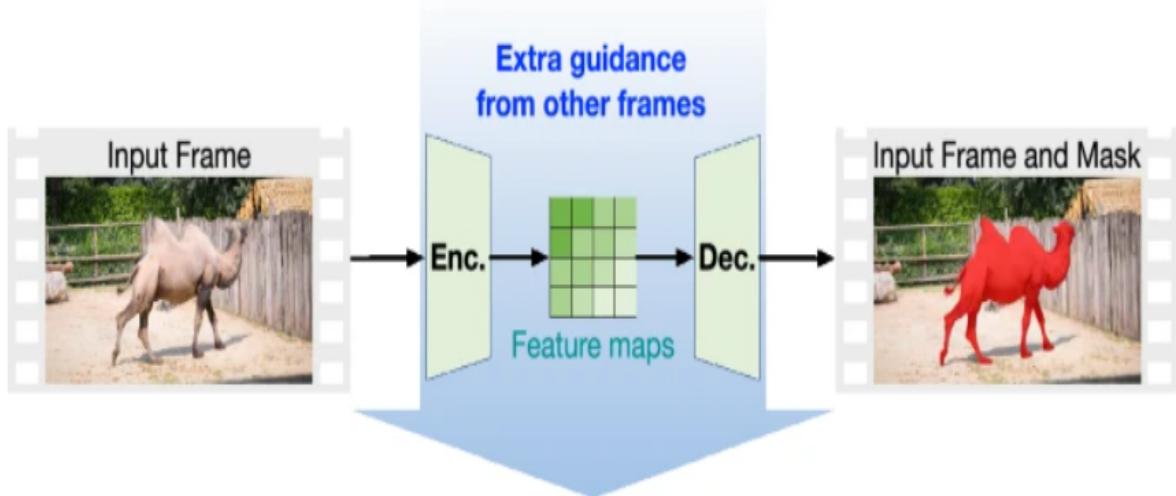


Figure 4: The basic architecture for VOS methods consists of two submodules: encoder and decoder, which perform the tasks of feature extraction and resolution recovery, respectively.

- In most VOS methods, segmentation is achieved by performing frame-by-frame target object extraction, where each frame is segmented according to spatio-temporal clues provided from the another frame in the same sequence.

4. Datasets and Difficulties: Considering the high demand of deep learning systems for data, this section goes through the existing datasets for VOS, followed by the corresponding evaluation metrics and key difficulties.

- **Hopkins-155**

Datasets	D.type				DA	Resolution	Videos	Annotations	Categories	Objects
	R	S	S	M						
Hopkins-155 (Tron and Vidal 2007)	✓		✓	✓	320 × 240–640 × 480	155	4615	-	345*	
BMS-26 (Brox and Malik 2010b)	✓		✓		350 × 288–640 × 480	26	189	2	47	
FBMS-59 (Ochs et al. 2013)	✓		✓		350 × 288–960 × 540	59	720	11	139	
SegTrackV1 (Tsai et al. 2012)	✓	✓		✓	320 × 240–414 × 352	6	244	6	6	
SegTrackV2 (Li et al. 2013)	✓		✓	✓	320 × 240–640 × 360	14	1154	12	24	
YouTube-Objects (Prest et al. 2012)	✓	✓			320 × 240–960 × 540	126	2127	10	126	
JumpCut (Fan et al. 2015)	✓	✓	✓	✓	640 × 400–1280 × 720	22	6331	12	22	
DAVIS-2016 (Perazzi et al. 2016a)	✓	✓		✓	854 × 480	50	3455	-	50	
DAVIS-2017 (Pont-Tuset et al. 2017)	✓		✓	✓	854 × 480	150	10,459	-	376	
DAVIS-2017-U (Caelles et al. 2019)	✓		✓	✓	854 × 480	150	10,731	-	449	
YouTube-VOS-2018 (Xu et al. 2018b)	✓		✓		1280 × 720	4453	197,272	94	7754	
YouTube-VOS-2019 (Xu et al. 2019b)	✓		✓		1280 × 720	4519	>190,000	94	8614	
YouTube-VIS (Yang et al. 2019a)	✓		✓		1280 × 720	2883	>131,000	40	4883	
SAIL-VOS (Hu et al. 2019)		✓	✓	✓	1280 × 800	201	111,654	162	1,896,295	

Figure 5: 14 video datasets and their main properties.

Based on these properties, the listed datasets are discussed in detail, especially in terms of application difficulties and settings, to guide researchers interested in VOS in choosing the appropriate dataset to train and evaluate their own methods.

- designed to evaluate point-based motion segmentation algorithms, where a set of points (from 39 to 550 points per frame) is annotated instead of entire pixels.
- This dataset consists of sequences grouped into three categories:
 - 1) checkerboard: moving objects overlaid with a checkerboard pattern to ensure the number of points tracked;)
 - 2) traffic scenes: outdoor traffic situations;
 - 3) articulated/non-rigid objects: sequences with movements of joints, faces and pedestrians.
- Although Hopkins-155 is useful for evaluating segmentation methods against rotation, translation, and degenerate motion, sparse annotations and limited challenges make it unsuitable for training and evaluation VOS methods are based on deep learning.

- **BMS (Berkeley Motion Segmentation Dataset) series**

- is designed for moving object segmentation and includes two versions: BMS-26 (Brox and Malik, 2010b) and FBMS-59 (Ochs et al., 2013, Freiburg-BMS).
- BMS-26 includes 26 video sequences, in which people and cars are the two most commonly used object types.
- FBMS-59 expands from BMS-26 by increasing the number of video sequences to 59 and including more object types.
- In both datasets, difficulties such as occlusion and motion pattern variation appear, so the disadvantages of VOS methods are evaluated on this dataset. However, with data marked for training because of low spatial resolution and relatively few annotated frames, it will be very difficult to train a good model from this data set.

- **SegTrack series**

- Designed for segmenting and tracking objects in video and includes two versions: SegTrack v1 (Tsai et al. 2012) and SegTrack v2 (Li et al. 2013).
- SegTrack v1 contains only 6 video sequences, but all frames are annotated with pixel-level layers.
- After adding more video sequences and annotated objects, SegTrack v2 expands SegTrack v1.
- Difficulty: frequent occurrence of rapid motion and object deformation, relatively low spatial resolution.

- **YouTube-Objects**

- This dataset included 126 video sequences with 2,127 annotated frames, and became the largest VOS dataset at the time. However, due to the paucity of object annotations and the uneven distribution of the classes, strings are not a suitable dataset for training.

- **JumpCut**

- consists of 22 video sequences with 6,331 frames, all annotated with pixel-level layers. In addition to real-world recorded video sequences, the dataset also includes a small number of frames. Based on the types of participants (mainly humans and animals) and challenges (fast moving and static objects), the dataset is divided into different groups for better organization.

- **DAVIS (Densely Annotated Video Segmentation) series**

- DAVIS (Densely Annotated VIdeo Segmentation), has evolved through versions over the years with three versions: DAVIS-2016 (Perazzi et al., 2016a), DAVIS-2017 (Pont-Tuset et al., 2017) and DAVIS-2017-U (Caelles et al., 2019), corresponding to different types of VOS tasks.

- **YouTube-VOS series**

- includes three versions: YouTube-VOS 2018 (Xu et al. 2018b), YouTube-VOS 2019 (Xu et al. 2019b) and YouTube-VIS (Yang et al. 2019a).
- The first two versions are designed for multi-object SVOS, while the latter caters to multi-object UVOS.
- Each video sequence in the datasets has a larger number of frames than any other dataset, allowing VOS methods to model and exploit long-term temporal dependencies between frames.
- **SAIL-VOS (Semantic Amodal Instance Level Video Object Segmentation)**
 - This is a synthetic dataset for VOS (Hu et al. 2019), where all video frames and corresponding layers are collected from Grand Theft Auto V, an action-adventure game. The images in the game are displayed as realistically as possible, so it is useful for training and evaluating VOS methods. Additionally, since all video sequences are generated by the game emulator, the resulting object layer is completely reliable, even when they are experiencing heavy congestion.
- **Evaluation metrics**
 - In VOS, the metrics commonly used to evaluate performance are the Jaccard index (Everingham et al. 2010), the F-measure (Martin et al. 2004), and their mean:

$$\left\{ \begin{array}{l} \mathcal{J} = \frac{|M \cap G|}{|M \cup G|} \\ \mathcal{F} = \frac{2P_c R_c}{P_c + R_c} \\ \mathcal{J}\&\mathcal{F} = \frac{(\mathcal{J} + \mathcal{F})}{2} \end{array} \right.$$

- Where G and M refer to the ground truth mask and segmentation mask, respectively. J evaluates the regional similarity between these two types of masks. P and R_c are the precision and recall calculated from the points in the contour $c(M)$ and $c(G)$. Therefore F evaluates the accuracy of boundary delineation. J&F measures overall VOS performance.

- **Summary**
- Hopkins-155, BMS series, SegTrack series: These earlier datasets were originally designed to evaluate non-deep learning methods but can still evaluate the performance of VOS methods, especially in handling object distortion and occlusion. However, they are rarely used in recent methods due to limitations in data diversity, number of challenges, and video length.
- YouTube-Objects, JumpCut: These datasets include videos with high range and resolution and are popular for evaluating earlier VOS methods in spatio-temporal feature embedding. However, they also have limitations in terms of data diversity and object recognition challenges, and only a few recent methods have been evaluated on them.
- SAIL-VOS: Different from other datasets, SAIL-VOS includes synthetic videos. Although there is still a gap between the generated and actual video frames, it provides reliable and controlled occlusions, which can improve the robustness of VOS methods facing occlusions. . However, it has not been widely used in the considered methods.
- DAVIS series, YouTube-VOS series: These are the most commonly used datasets for training and evaluating recent VOS methods. They provide large-scale video sequences, diverse object types, variety of challenges, and high-quality annotations. The DAVIS

series and the YouTube series differ in annotation, allowing assessment of different VOS properties. DAVIS is favored for evaluating temporal stability, while the YouTube series is favored for general performance due to the larger number of videos and object types.

5. Common difficulties and challenges:

- Introduces some of the challenges for UVOS and SVOS fields, including attribute changes, occlusions, conflicts between similar instances, unknown backgrounds, temporal consistency, and the trade-off between efficiency and robustness. Exactly.

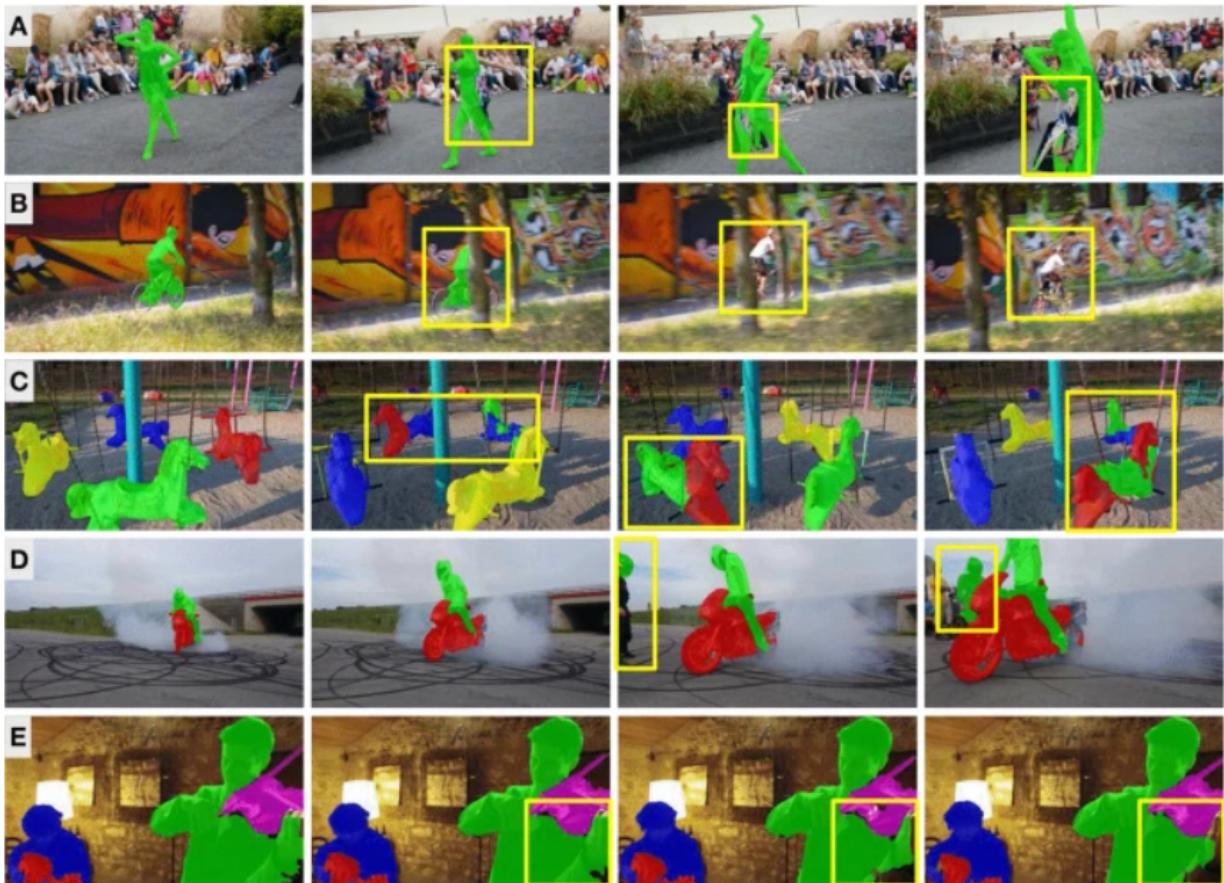


Figure 6: Incorrect result

Each row shows the influence of an obstacle on existing VOS methods.

- A. A variable property;
- B. occlusion;
- C. discrimination between similar objects;
- D. ambiguous background image;
- E. The VOS is temporally consistent (this row consists of continuous frames without rapid motion, occlusion, and significant appearance changes). box highlights erroneous results

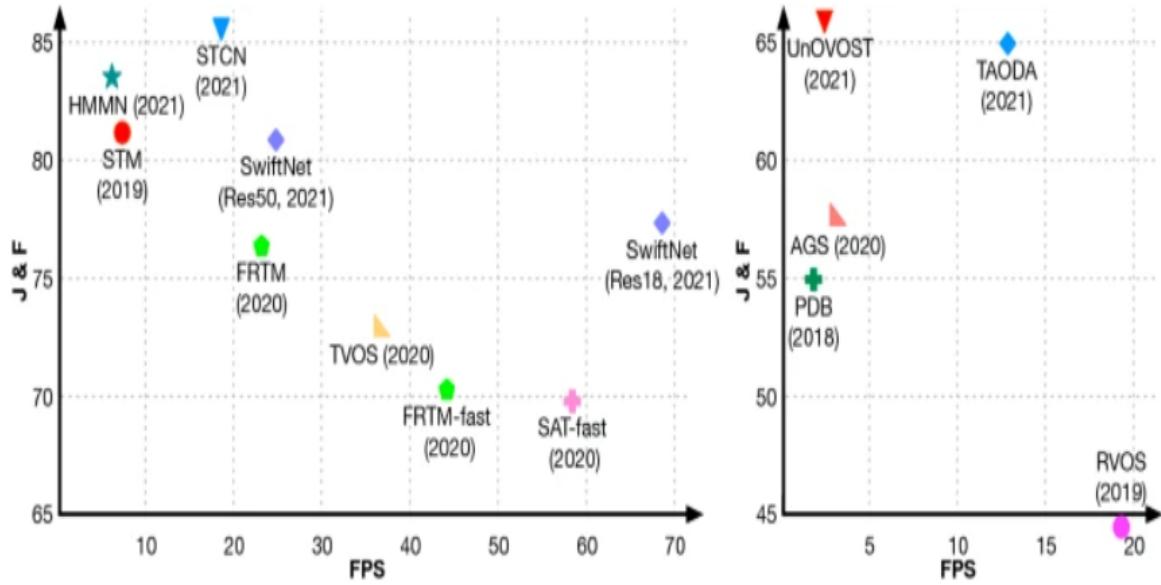


Figure 7: Accuracy and segmentation efficiency of recent SVOS (left) and UVOS (right) methods on the DAVIS-2017 validation set

- **Change object properties**
 - This obstacle mainly affects VOS methods based on visual similarity. During inference, these methods segment regions with similar image features to the annotated (mostly SVOS methods) or predicted (mostly UVOS methods) target objects.
- **Occluded by confounding factors**
 - This obstacle mainly affects propagation-based VOS methods, which consider objects predicted in the previous frame for estimation current frame segment.
- **Disturbed by background/similar objects**
 - This obstacle mainly affects VOS methods based on visual similarity, saliency, or motion patterns.
- **Consistency in time**
 - This obstacle mainly affects VOS methods that use less motion information. During inference, these methods essentially perform image segmentation for each video frame. Therefore, it is difficult to maintain temporal consistency of segmented objects, i.e. the evolution of the object mask predicted from consecutive frames is not smooth (in case no occlusion, rapid movement, or significant property change).
- **Balance between VOS accuracy and efficiency**
 - This obstacle mainly affects VOS methods serving real-time applications. In general, these methods should perform at least 24 FPS (Frames Per Second) segmentation while achieving high quality object class.

6. Method

- **Online refinement-based methods**
 - There are three main stages to transfer the output domain of the segmentation network from general knowledge to annotated objects:
 - (1) Initialize the network (gray) with pre-trained parameters on ImageNet (Russakovsky

et al. 2015;

(2) pre-train the network (green) on object segmentation datasets (e.g., MS-COCO (Lin et al. 2014) and DAVIS (Perazzi et al. . 2016a));

(3) Fine-tuning the network (yellow) on the annotation frame. Since pre-training and fine-tuning are performed before and during inference, we call them "external" processes online" and "online", respectively.

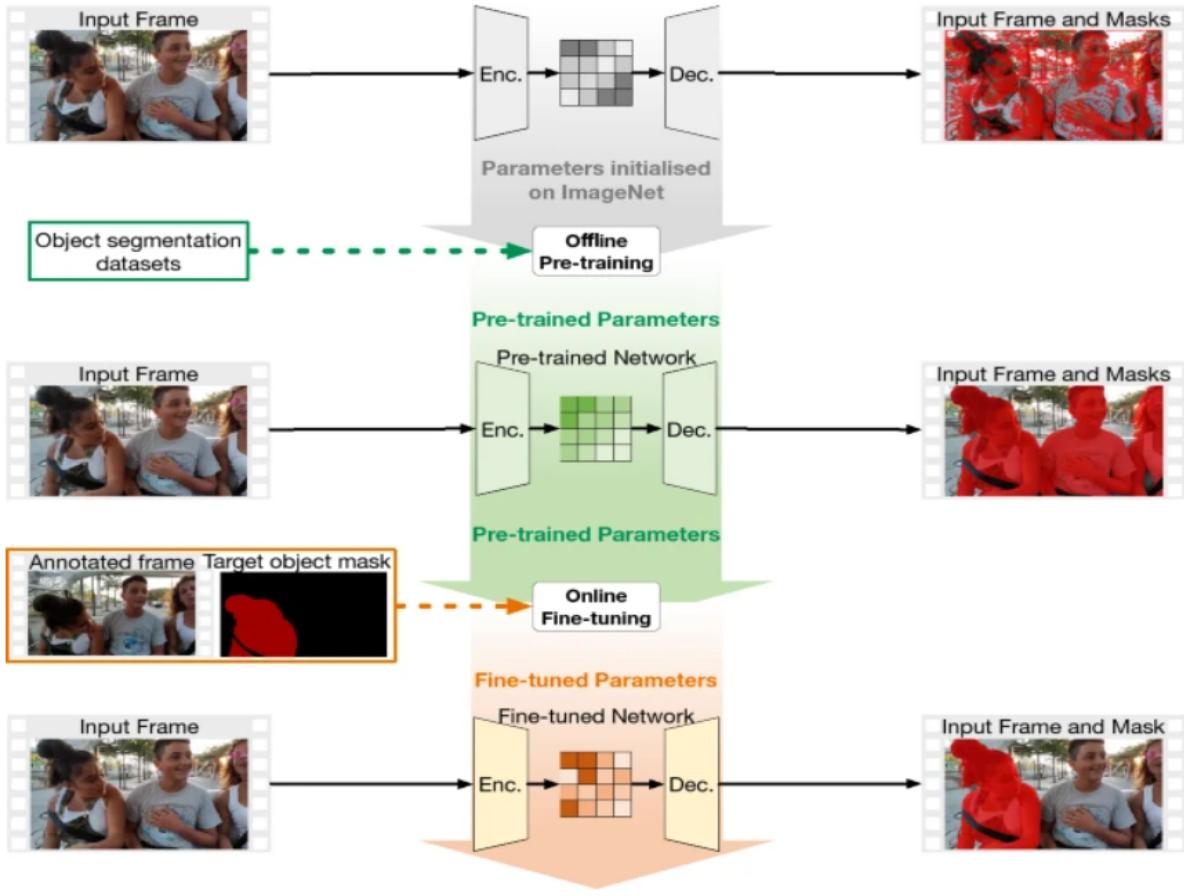


Figure 8: Diagram of VOS method based on online refinement

- OSVOS is the first method to use online fine-tuning for SVOS. Then some methods are derived from OSVOS. Their major modifications to OSVOS are marked in bold words with the prefix "+". Recently, several variations of online refinement have been proposed for effective VOS.

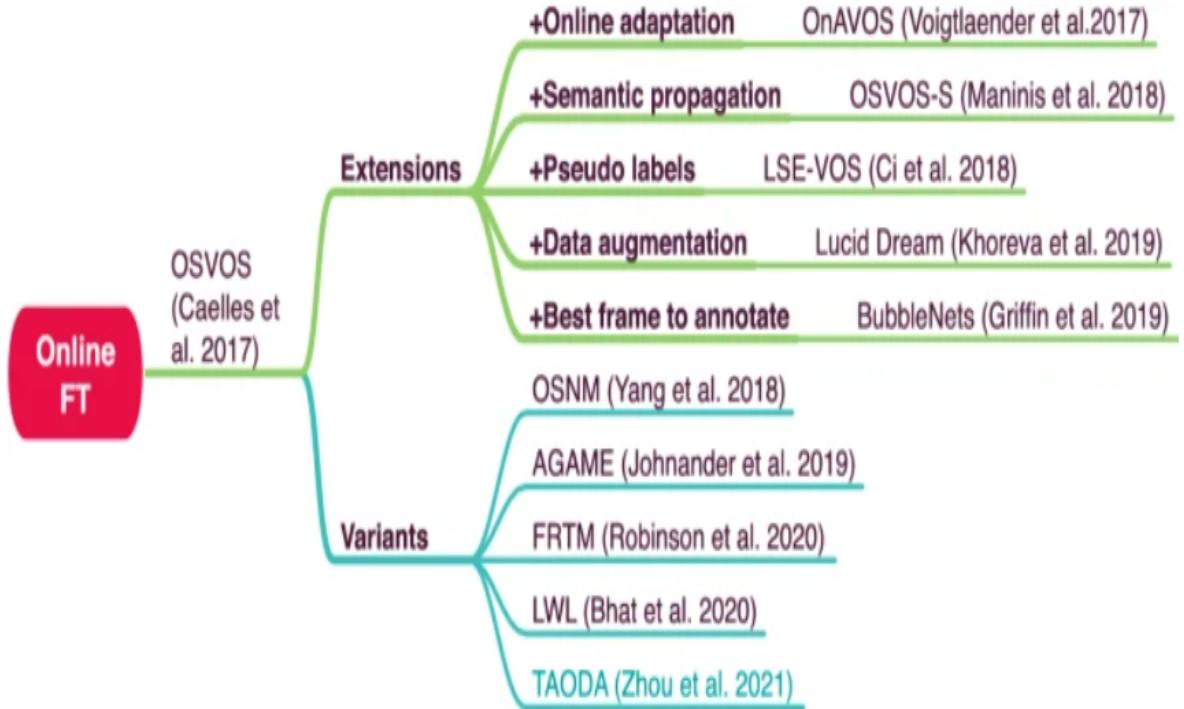


Figure 9: Roadmap for developing online fine-tuning-based methods.

- The VOS method is based on online refinement and their variations, derived from OSVOS (Caelles et al. 2017). These methods transform the network output from general knowledge to specific objects by fine-tuning the network parameters with first frame annotations. However, problems in OSVOS have motivated the development of extension and variant methods: Considering only the first frame annotations, reduces VOS efficiency.
- **Extensions that work to improve OSVOS in terms of adaptability and robustness,**
 - * OnAVOS (Voigtlaender and Leibe 2017) and LSE-VOS (Ci et al. 2018): Improve adaptation by combining high-confidence results from past frames. However, multiple online fine-tuning is required, which reduces VOS efficiency.
 - * OSVOS-S (Maninis et al. 2018) and LucidTracker (Khoreva et al. 2019): Increasing segmentation robustness. OSVOS-S combines knowledge from general audience segmentation to refine VOS results. LucidTracker achieves this by generating diverse patterns from annotations. These methods boost VOS performance but are still less efficient due to the need for additional computation or data augmentation.
 - * BubbleNets (Griffin and Corso 2019): Is an incremental method, which can be integrated into the above methods to predict the optimal frame for annotation, helping to generate an optimal frame in a cost-effective manner .
- **Variations focus on more VOS efficiency.** OSNM, A-GAME, FRTM, LWL and TAODA: Developed to change the network output domain with more efficient algorithms. Although better efficiency is achieved, there is still an accuracy gap between the variants.
- **Combination-based method** This method performs VOS by measuring the correspondence between the target frame and the reference frame

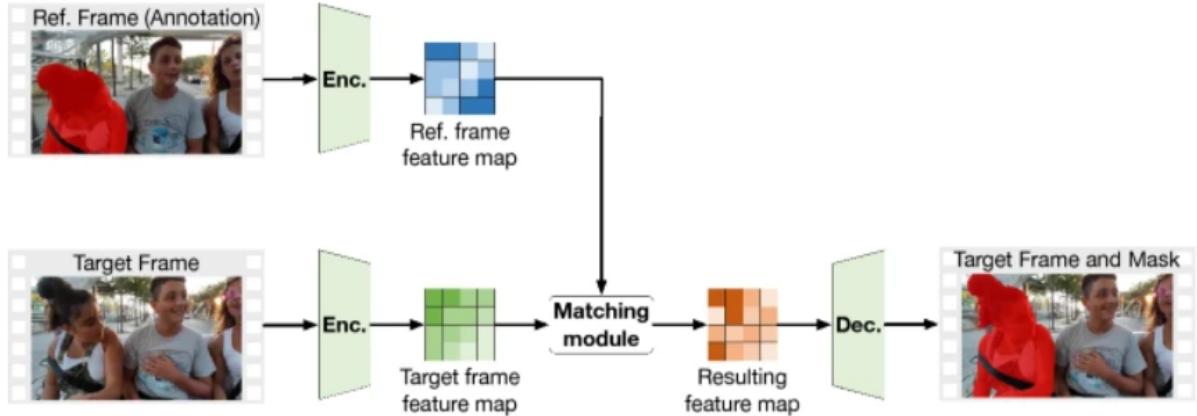


Figure 10: Diagram of combination-based VOS methods

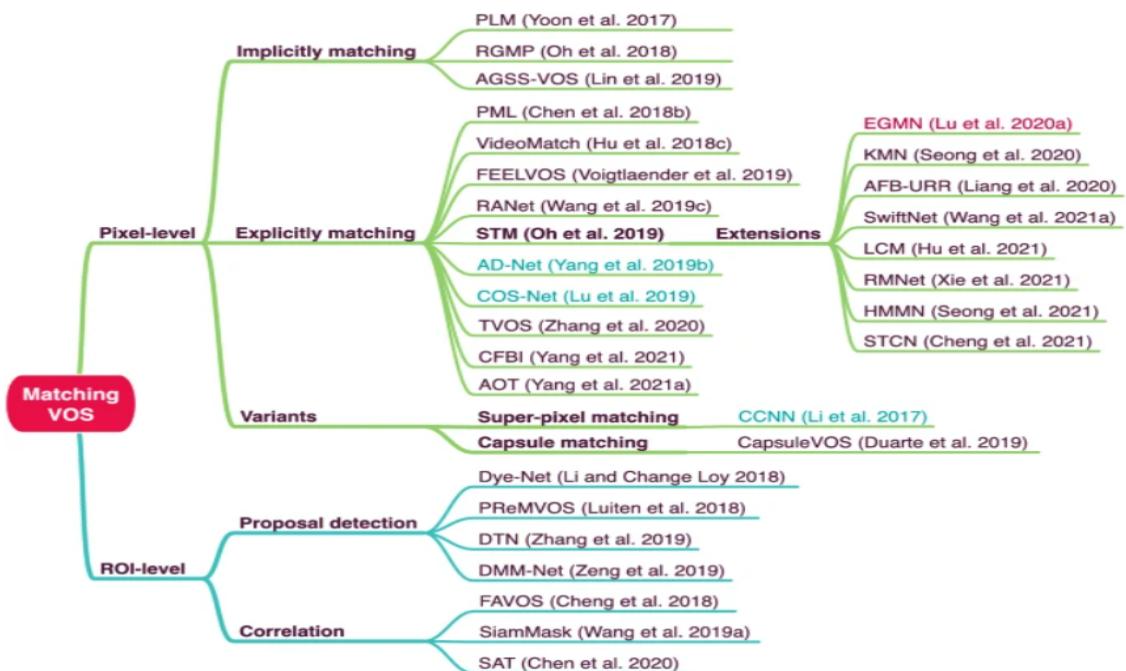


Figure 11: Roadmap for developing combination-based methods

- Most of today's leading SVOS methods have been developed based on feature matching, and these methods have achieved high performance in experiments. Reliable correspondence also provides high-quality results for UVOS methods.
- Feature matching-based SVOS methods typically incorporate pixel-level matching, measuring dense correspondence between frames. This method can produce robust results without online fine-tuning, while also helping to reduce computational time.
- The explicit histogram-based method directly measures the similarity between the pixels of the target frame and the reference frame, without the need for an additional module for correspondence. These methods improve segmentation performance by combining similarity with fine-grained features in the target frame.
- Their differences mainly lie in the relevant entities and the approach to locating the target audience. Since correlation-based methods do not require any additional network to generate ROI, they perform VOS much faster than detection-based networks. However, the segmentation performance of this method is limited because fine-grained details are ignored in the ROI-based segmentation module.
- **Graph optimization based methods** there are two types of approaches to organize and analyze graph nodes: track selection and information propagation

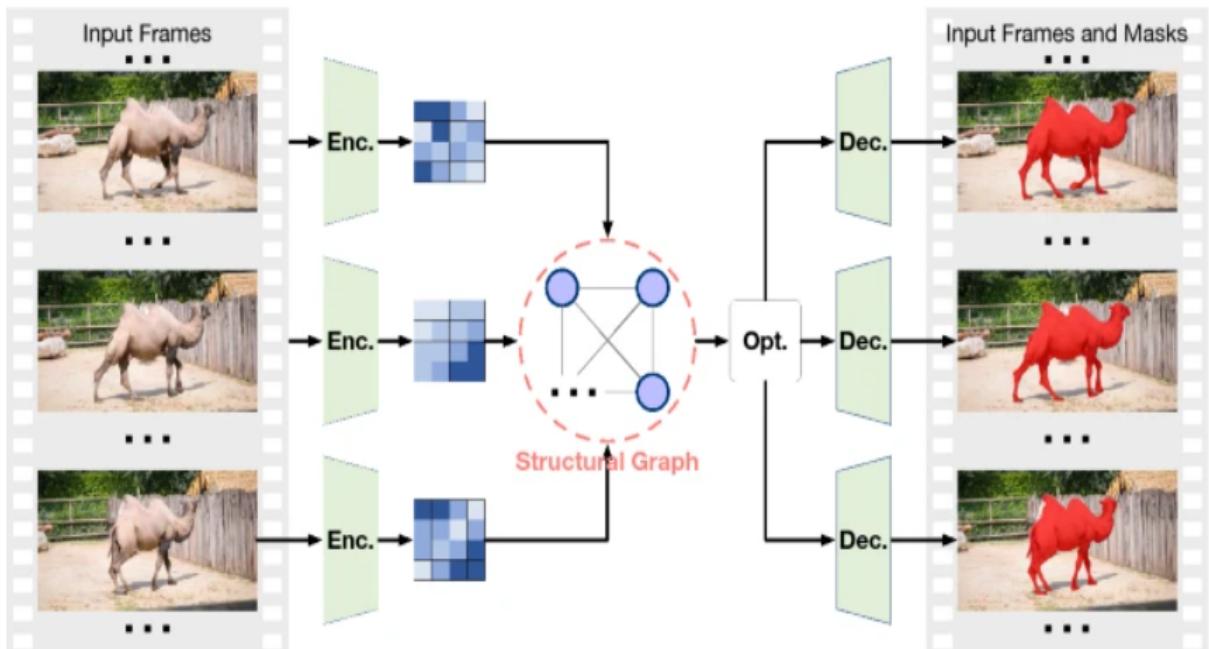


Figure 12: Diagram of VOS methods based on graph optimization, applicable to both SVOS and UVOS.

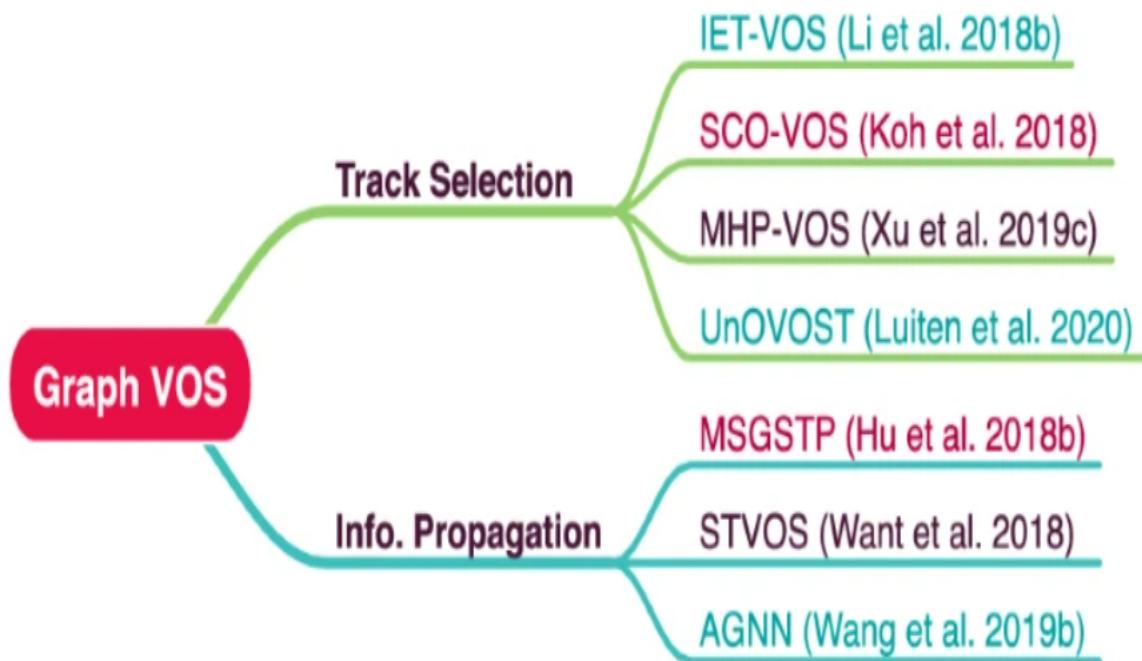


Figure 13: Development roadmap of methods based on graph optimization

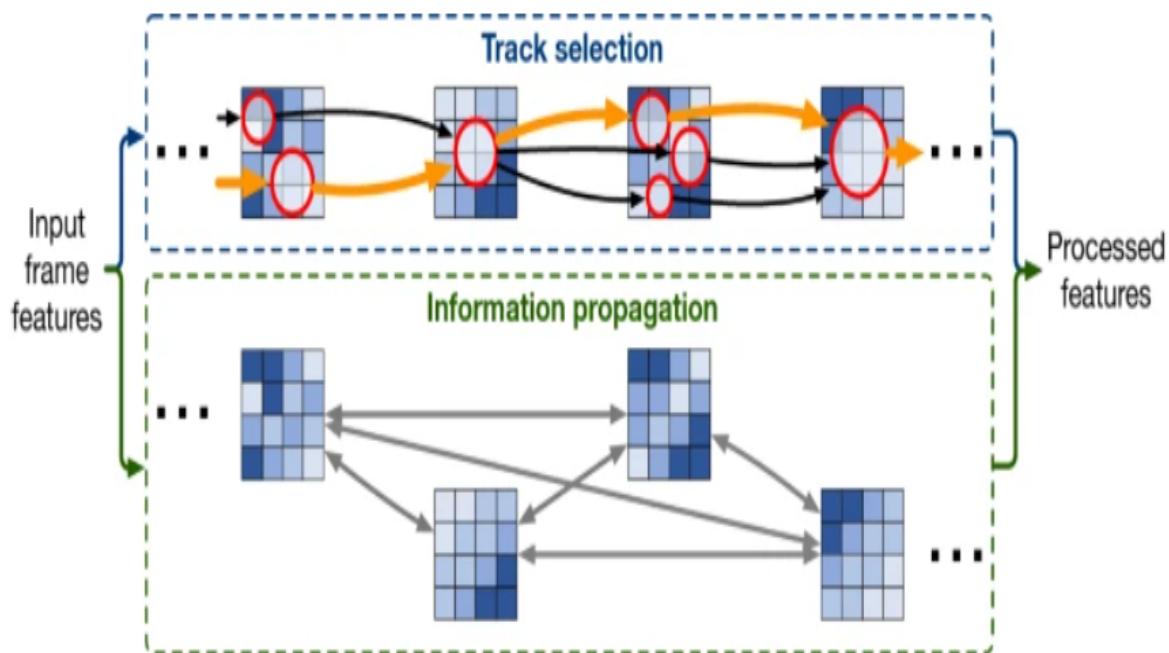


Figure 14: Diagram of two techniques for organizing and analyzing graph nodes

- There are two main options for organizing and analyzing graph data in VOS: tracking options and information dissemination. The track selection-based scheme organizes nodes into a set of tracks or trees, thereby creating object masks through optimal retrieval of tracks. These methods often require additional networks to generate object proposals, increasing the computational cost and number of network parameters.
- The information propagation-based scheme supports VOS by transmitting information repeatedly between nodes. These methods focus on propagating label-related information, making the segmentation results sensitive to the connections between nodes. Recent work has mitigated this problem by considering deep features and frame-level information propagation.

In summary, graph optimization provides a broader correspondence with VOS methods, allowing to achieve high-quality segmentation results, especially in UVOS. However, this technique has high computational cost, making it unsuitable for real-time and resource-limited VOS tasks.

- **Optical flow-based method** Optical flow has been a widely used technique in VOS due to its pixel-level motion patterns. This technique assumes that the target object and the background have different motion patterns. Therefore, integrating optical flow into VOS can provide segmented networks with reasonable priorities

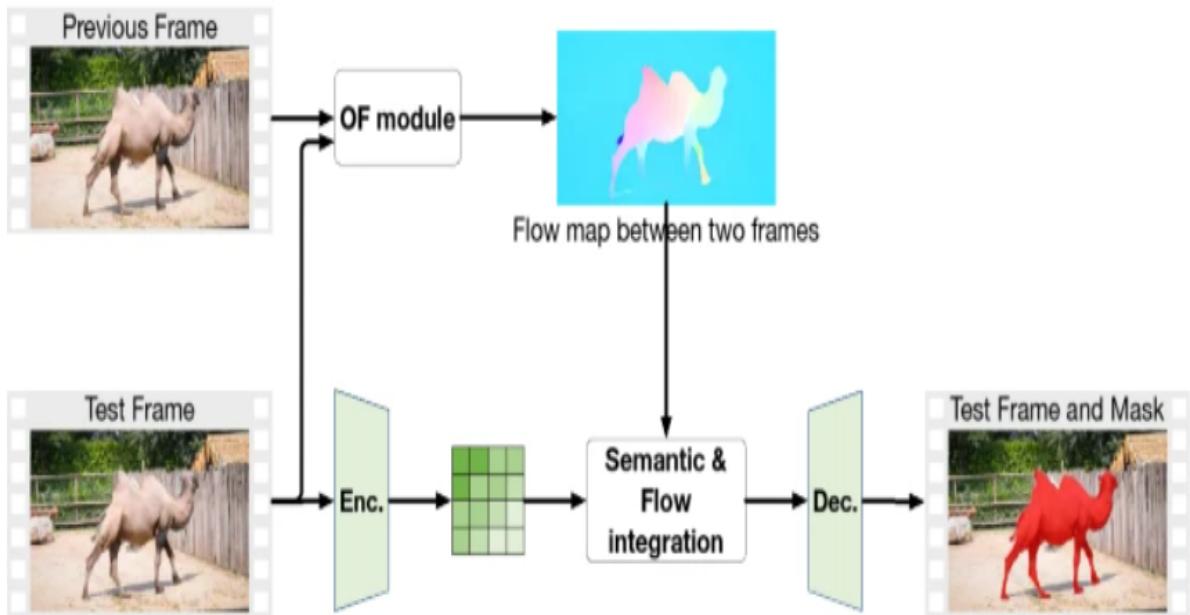


Figure 15: Diagram of optical flow-based VOS methods, applicable to both SVOS and UVOS

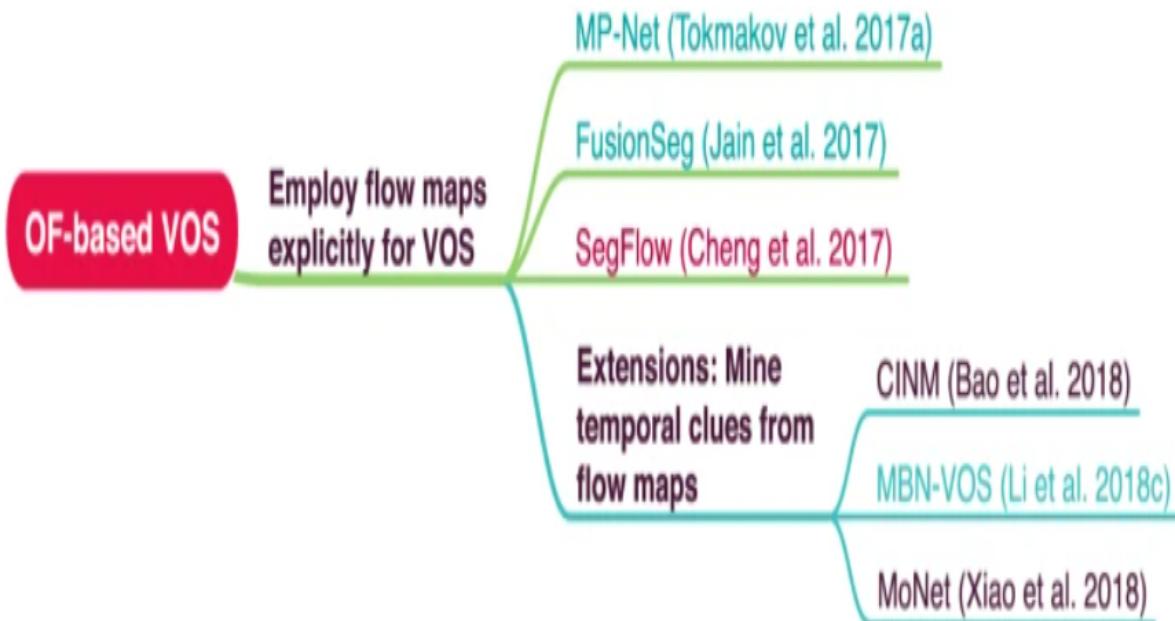


Figure 16: Development roadmap of optical flow-based VOS methods

- This section focuses on optical flow-based VOS methods, which assume that the object and background move in different patterns. These methods use flow maps to estimate the shape and location of the target object. Previous methods have explicitly integrated flow maps with spatial features to create object classes. However, flow estimation is not always reliable due to lack of training data and challenges such as dynamic backgrounds. Therefore, recent methods have been proposed to exploit optical flow more effectively and avoid these risks. *Although good results have been achieved on many challenging sequences, the use of optical flow has become less popular in recent VOS systems due to several limitations such as indistinguishability object-to-background treatment in some cases, as well as requiring additional deep networks during inference. Therefore, new methods using layer propagation have gradually replaced optical flow.*
- Method based on mark propagation

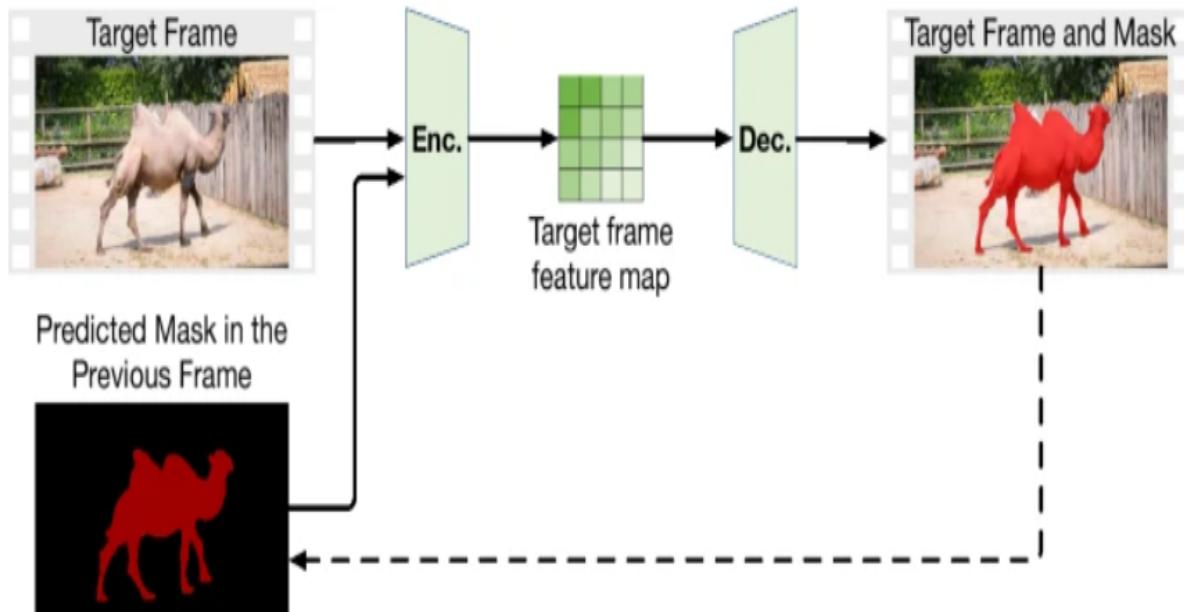


Figure 17: Diagram of VOS method based on mark propagation

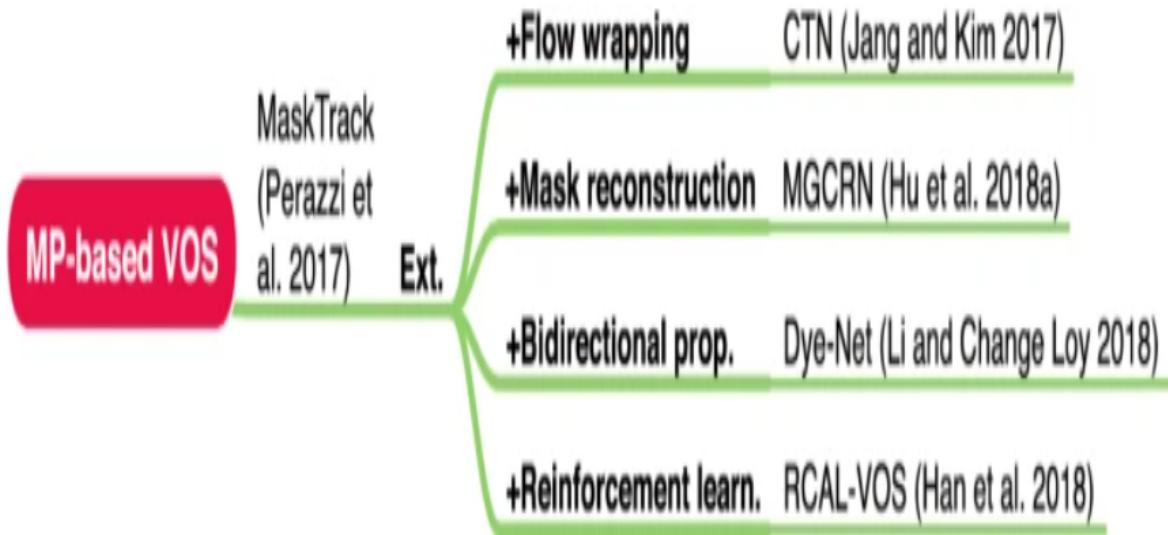


Figure 18: VOS development roadmap based on mark propagation

– this method uses the existing layer to estimate the position and shape of objects in subsequent frames. Although they provide high-quality results, these methods face challenges when dealing with strong deformations and sudden motions. To improve reliability, methods such as CTN and MGCRN integrate optical flow information, while methods such as DyeNet and RCAL-VOS use two-dimensional layer propagation techniques to cope with the challenges. Overall, the use of layer propagation has contributed greatly to VOS methods by providing information about the position and shape of objects from previous frames.

- Method based on long-term breeding

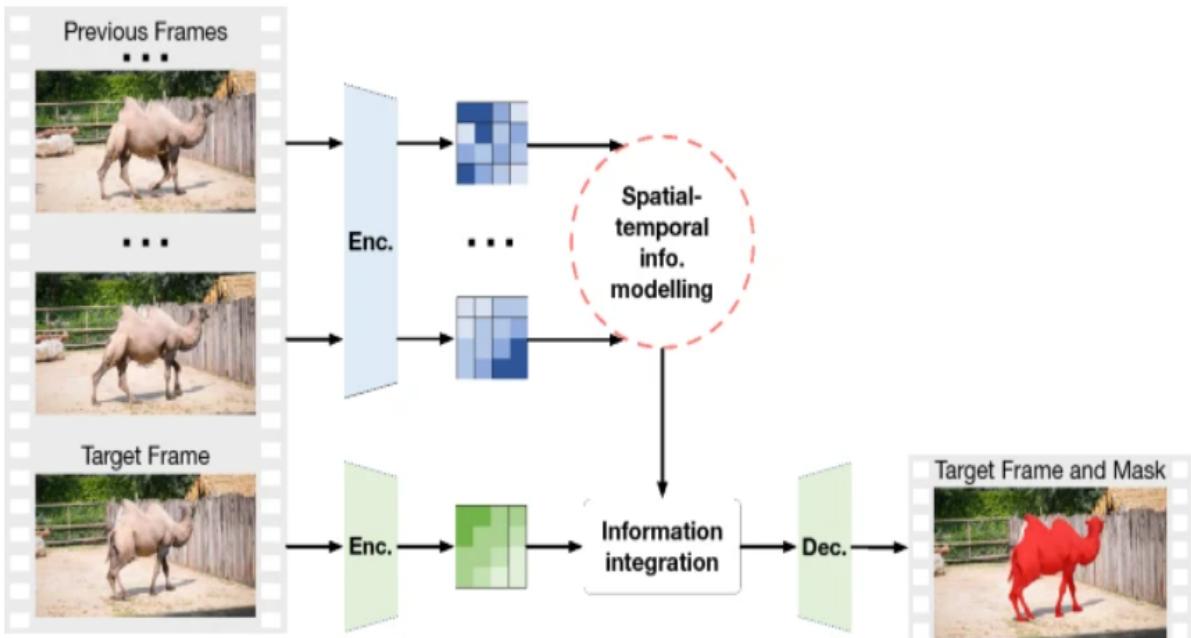


Figure 19: Diagram of VOS methods based on long-term time propagation, applicable to both SVOS and UVOS

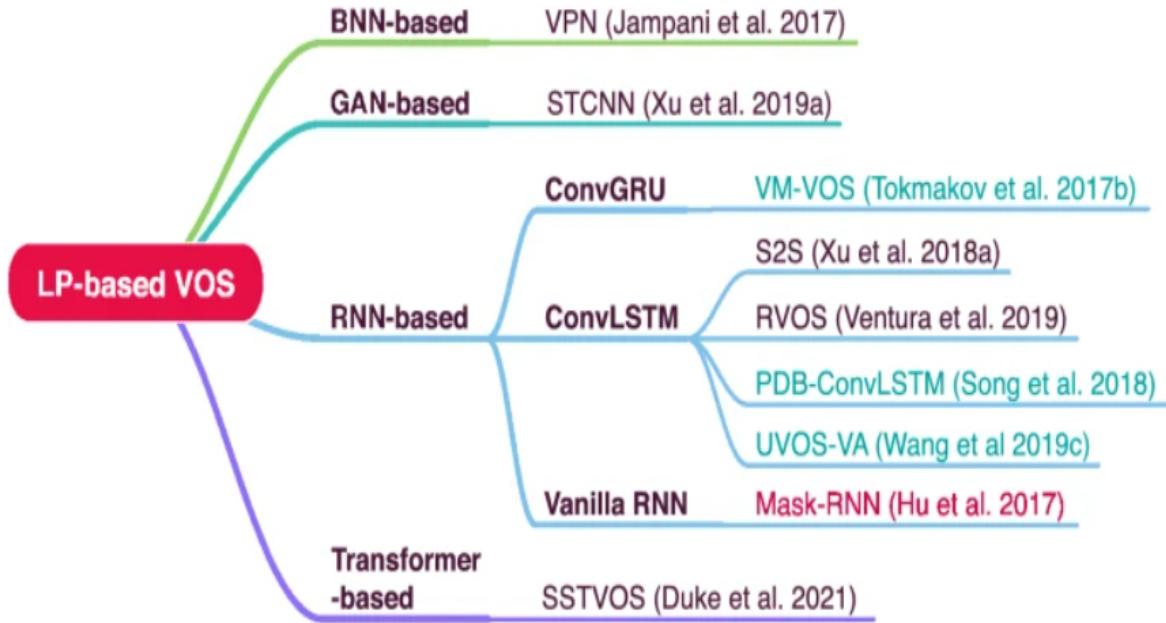


Figure 20: Follow the development of VOS methods based on long-term time propagation (abbreviated as 'LP-based VOS')

- This section discusses VOS methods based on long-term temporal propagation, which accumulates spatio-temporal information over many frames to estimate the shape and location of objects. There are four main types of techniques to achieve this, including BNN, GAN-based methods, RNN-based VOS methods, and hybrid methods to enhance performance.
- Although the use of long-term breeding offers many benefits, it is currently rarely used and does not produce better results than other methods. Some challenges encountered were slow computation and the use of short frames during training. Future proposals focus on how to transmit long-term information under limited resource conditions.

7. Experimental results and discussion

- Quantitative and qualitative results

Table 1: Đánh giá và hiệu quả phân đoạn của các phương pháp SVOS đã được trên базa DAVIS-2016

Methods	S. techs			T. techs			Frames	Resolutions	J&T	FPS ↑
	O	M	G	O	P	L				
OSVOS	x						1	480 x 854	80.2	0.38
MSKTrack	x			x	x		1, t-1	480 x 854	77.6	0.29
OSMN					x		1, t-1	480 x 854	73.3	1.87
RGMP		x		x		x	1, t-1	480 x 854	81.8	12.4
SiamMask		x			x		1, t-1	480 x 854	69.8	79.6
A-GAME					x		1, t-1	480 x 854	81.9	12.6
RVOS						x	[1, t-1, 1]	240 x 427	72.3	193.7
RANet		x			x		1, t-1	480 x 854	87.1	41.3
STM		x			x		[1, t-1, 5]	480 x 854	89.4	11.9

- **Discussion and future research directions**

- Large-scale and densely annotated training datasets: Large-scale and densely annotated training datasets are required to fully inform the VOS model. Currently, DAVIS is one of the main datasets for this purpose, but more similar datasets need to be developed to improve performance.
- Long-term temporal information analysis: There is a need to research and develop VOS methods capable of analyzing long-term temporal information from video sequences, helping to address challenges such as occlusion, out-of-sight, and transitions. move quickly.
- Balancing between accuracy and efficiency of VOS: It is necessary to develop VOS models that are able to balance between accuracy and efficiency, by designing networks that are lightweight but still ensure performance.
- Multi-object UVOS: There is a need to develop UVOS methods that allow the discrimination of multiple objects in a video sequence, instead of just identifying a single object. This will facilitate more practical application in VOS.

AI VIET NAM – COURSE 2024

HỌC SÂU TRONG NHẬN DIỆN ĐỐI TƯỢNG TRONG VIDEO: TỔNG QUAN

Ngày 20 tháng 4 năm 2024

Ngày công bố:	08/04/2022
Tác giả:	Mingqi Gao, Feng Zheng, James J.Q.Yu, Caifeng Shan, Guiguang Ding, Jungong Han
Nguồn:	Artificial Intelligence Review (2023) 56:457–531
Nguồn dữ liệu (nếu có):	VOS - REVIEW
Từ khóa:	VOS (Video Object Segmentation), Deep Learning, CNN (Convolutional neural network)
Người tóm tắt:	Nhi Ng Thảo

1. Mở đầu: Phần tóm tắt mở đầu:

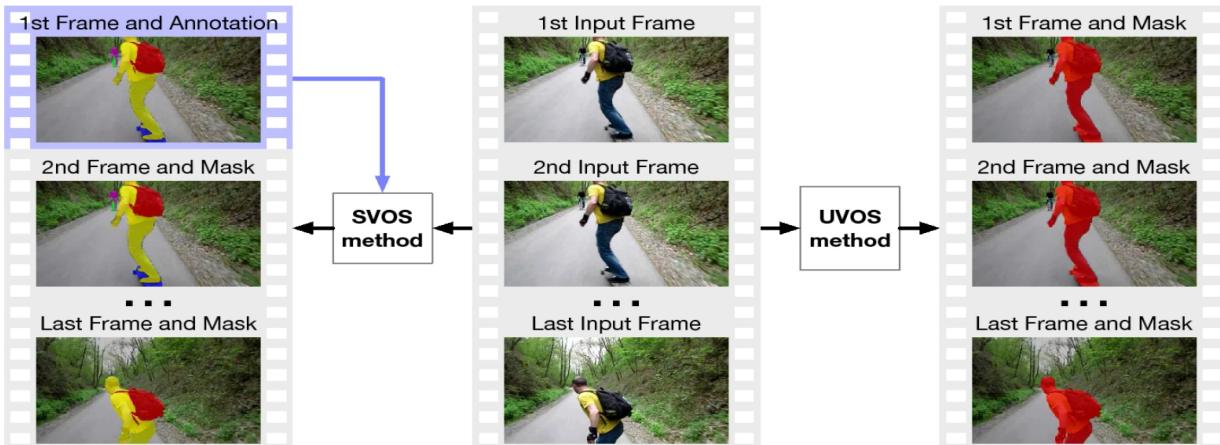
- Phát hiện nhận diện đối tượng trong video là một trong những ứng dụng thị giác máy tính, với phát hiện đối tượng trong video là một trong những ứng dụng được ủng hộ và nghiên cứu chuyên sâu nhất. Bài báo này nhằm mục đích hệ thống về kiến thức, tài liệu phát hiện đối tượng trong video bằng deep learning.
- Nêu ra những ưu điểm và nhược điểm của từng loại phương pháp. Thông qua những nội dung giới thiệu định nghĩa, khái niệm nền tảng và ý tưởng cơ bản của các thuật toán trong lĩnh vực này. Sau đó là tóm tắt các bộ dữ liệu để huấn luyện và thử nghiệm các thuật toán nhận diện đối tượng trong video và cũng như các nhược điểm của thuật toán ấy và các số liệu đánh giá phổ biến.
- Dựa trên những kết quả định lượng và định tính của một số phương pháp đại diện trên một tập dữ liệu có những nhược điểm gì để cung cấp cũng như phân tích nhằm thảo luận thêm về hướng nghiên cứu tương lai. Nhìn chung bài báo này nhằm mục đích giúp nhanh chóng nắm bắt tiến độ hiện tại trong lĩnh vực nghiên cứu nhận diện đối tượng trong video.

2. Sơ lược về bài báo:

- Phân đoạn đối tượng video (VOS) là nhiệm vụ tách các vùng tiền cảnh khỏi nền trong chuỗi video. Tương tự như theo dõi đối tượng, các phương pháp VOS thiết lập sự tương ứng của các đối tượng giống hệt nhau trên các khung, nhưng có thể đạt được biểu diễn đối tượng chi tiết hơn. Vậy nên VOS, có vai trò quan trọng trong thực tế, như giám sát trực quan, nhận dạng hành động, tóm tắt và chỉnh sửa video.
- Ban đầu VOS dựa trên tính năng thủ công, và sự khán quan, luồng ảnh xạ ảnh (optical flow), vùng chủ thể chính (visual saliency) là những kỹ thuật phát hiện đối tượng trong video. Nhưng đây chỉ là những kỹ thuật trước đây, với sự phát triển hiện đại, các phương

pháp VOS dựa trên deep learning trả về những kết quả và hiệu suất cao về kết quả cũng như độ chính xác. Vậy nên những phương pháp gần đây được thực hiện dựa trên mạng nơ ron học sâu.

- Các phương pháp VOS hiện tại được nhóm và chia thành 4 loại: Không giám sát, bán giám sát, tương tác và hướng dẫn bằng ngôn ngữ.
 - **Phân đoạn không giám sát:** Phương pháp này không yêu cầu dữ liệu huấn luyện được gán nhãn.
 - **Phân đoạn bán giám sát:** Phương pháp này kết hợp cả dữ liệu huấn luyện được gán nhãn và không gán nhãn. Nó sử dụng thông tin từ cả hai loại dữ liệu để tạo ra các khu vực phân đoạn.
 - **Phân đoạn tương tác:** Phương pháp này liên quan đến việc tương tác với người dùng hoặc các thông tin bổ sung.
 - **Phân đoạn dựa trên hướng dẫn bằng ngôn ngữ:** Phương pháp này sử dụng thông tin từ hướng dẫn bằng ngôn ngữ để phân đoạn đối tượng.
 - **Bài báo này chủ yếu tập trung vào hai loại được nghiên cứu rộng rãi là:** VOS không giám sát (UVOS) và VOS bán giám sát(SVOS).
 - Các phương pháp UVOS thực hiện phân đoạn mà không có bất kỳ nhãn sự thật cơ bản hoặc trước nào (unsupervised setting). Các đối tượng có mẫu chuyển động nổi bật hoặc độ nổi bật thị giác có thể được nhận dạng.
 - các phương pháp SVOS bắt đầu với các nhãn sự thật cơ bản có sẵn trong một khung hình (thường chỉ là khung hình đầu tiên, semi-supervised setting). Các nhãn này được chú thích thủ công để cho biết các đối tượng sẽ được phân đoạn từ các khung còn lại.
- Dể tránh nhầm lẫn về khái niệm, điều đáng nói là một số tác phẩm gần đây gọi unsupervised/semi-supervised VOS là automatic/semi-automatic VOS hoặc zero-shot/one-shot VOS.
- Một ví dụ so sánh 2 loại nghiên cứu này: Cả hai phương pháp đều lấy video thô làm đầu

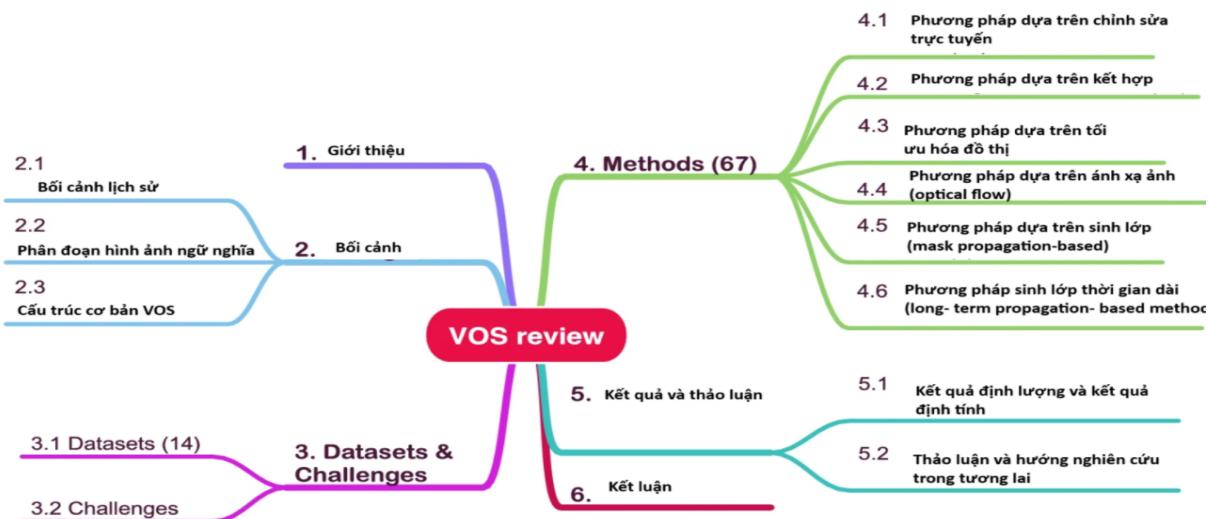


Hình 1: Minh họa sự khác biệt giữa hai phương pháp VOS.

vào. Phương pháp UVOS nhận dạng đối tượng có chuyển động chi phối hoặc nổi bật đối tượng. Ngược lại, các đối tượng mục tiêu (những đối tượng cần phân đoạn) trong SVOS phụ thuộc vào chú thích của con người trong khung đầu tiên (được tô sáng màu tím). Do đó, các phương pháp SVOS có tính linh hoạt hơn trong việc xác định các đối tượng mục tiêu.

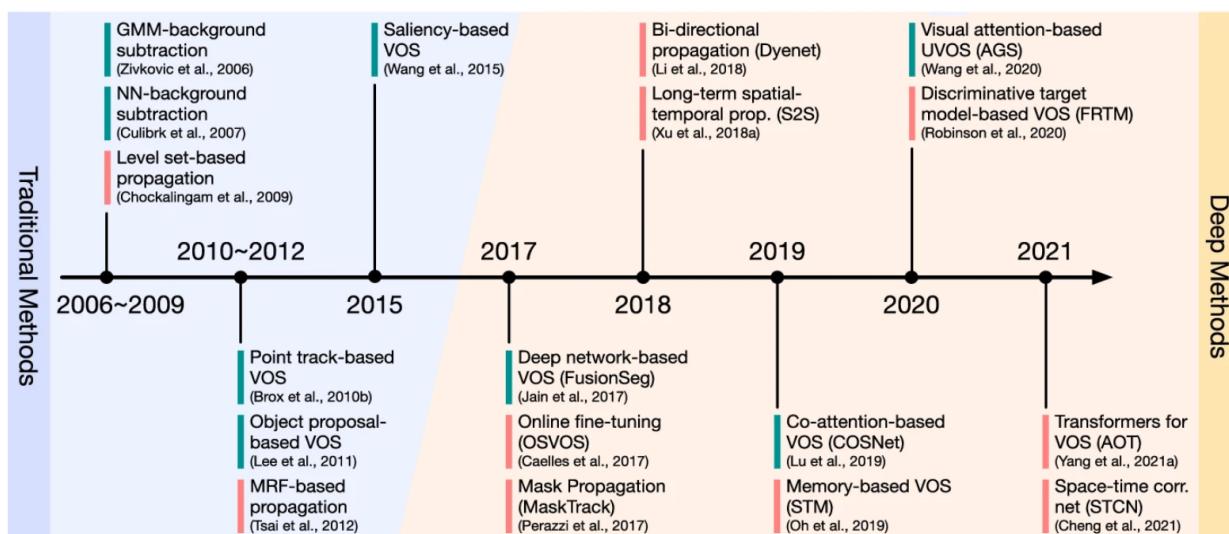
- **Tóm lại mục đích nghiên cứu bài báo:**

- Cung cấp đánh giá và phân tích các bộ dữ liệu có lợi cho việc đào tạo và đánh giá các phương pháp UVOS và SVOS.
- Nhóm các phương pháp UVOS và SVOS hiện có thành sáu loại theo việc sử dụng tính năng không gian và thời gian và cung cấp đánh giá chuyên sâu và có tổ chức về nguồn gốc, lịch sử phát triển, kiến trúc, ưu, nhược điểm và phương pháp đại diện của chúng.
- Thảo luận về hiệu suất của các phương pháp được xem xét bằng cách phân tích kết quả đánh giá được phát hành trên một số bộ dữ liệu điểm chuẩn và thử nghiệm một số kỹ thuật đại diện trên các loại chuỗi video thử thách khác nhau.
- Tóm tắt một số xu hướng phát triển của các phương pháp được xem xét và rút ra một số dự báo về những tiến bộ có thể có trong tương lai.



Hình 2: Mục lục trực quan của bài báo

3. Bối cảnh



Hình 3: Ngắn gọn bối cảnh lịch sử trong lĩnh vực VOS.

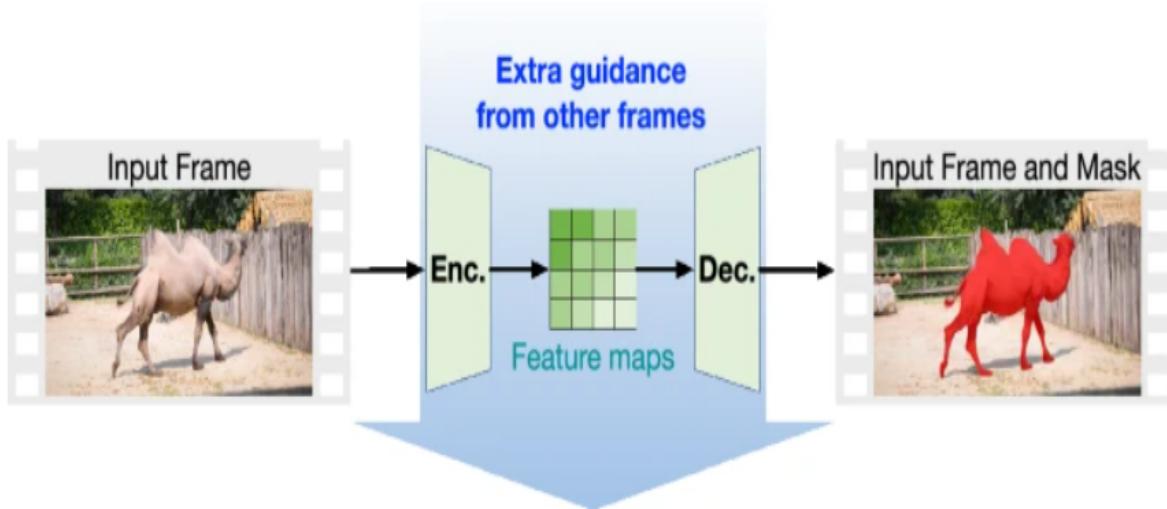
- **Bối cảnh lịch sử VOS**

- Ban đầu tập trung vào việc trích xuất các đối tượng di chuyển từ chuỗi video bằng các kỹ thuật tách nền, như xây dựng các mô hình nền và trừ chúng từ các khung hình hiện tại.
- Phát triển sau các phương pháp tạo ra đề xuất đối tượng thành công, các phương pháp dự đoán Đối tượng dựa trên Đề xuất liên quan đến việc tạo ra đề xuất cho các khung hình video và xếp hạng chúng để xác định các đối tượng lặp lại, mặc dù chậm hơn do hiệu suất kém.
- Nhận ra tầm quan trọng của liên tục thời gian, các Phương pháp dựa trên Quỹ đạo Điểm này xây dựng các quỹ đạo điểm dựa trên thông tin chuyển động và phân cụm chúng cho phân đoạn, cải thiện độ chính xác bằng cách xem xét thông tin quỹ đạo cục bộ và toàn cầu.
- Cùng với phân đoạn tự động, các phương pháp SVOS (VOS Bán giám sát) nhằm lan truyền các mặt nạ/chỉ dẫn đã được chú thích sang các khung hình khác, tập trung vào mô tả đặc điểm và sự tương quan thời gian trước sự xuất hiện của các phương pháp dựa trên học sâu.

- **Phân vùng ảnh theo ngữ nghĩa**

- Phân loại từng pixel thành các danh mục ngữ nghĩa được xác định trước dựa trên các đặc trưng đã được mã hóa.
- Để áp dụng thành công CNN vào phân đoạn hình ảnh, một số thay đổi đã được thực hiện, trong đó mạng tích chập hoàn toàn (fully convolutional network) đã xuất hiện.
- FCN có thể lấy đầu vào có kích thước tùy ý và tạo ra đầu ra với kích thước tương ứng. Trong quá trình đào tạo, FCN được khởi tạo với trọng số được đào tạo trước trên ImageNet và sau đó được điều chỉnh trên bộ dữ liệu phân đoạn như PASCAL. FCN tạo ra một tập hợp xác suất cho mỗi pixel trong hình ảnh đầu vào, chỉ ra khả năng pixel thuộc về các danh mục ngữ nghĩa.

- **Kiến trúc cơ bản của VOS:**



Hình 4: Kiến trúc cơ bản cho các phương thức VOS bao gồm hai mô-đun con: bộ mã hóa và bộ giải mã, thực hiện các nhiệm vụ trích xuất tính năng và khôi phục độ phân giải, tương ứng.

- Trong hầu hết các phương pháp VOS, việc phân đoạn đạt được bằng cách thực hiện trích xuất đối tượng mục tiêu theo từng khung, trong đó mỗi khung được phân đoạn theo các manh mối không gian-thời gian được cung cấp từ các khung khác trong cùng một chuỗi.
4. **Bộ dữ liệu và Khó khăn:** Xem xét nhu cầu cao của các hệ thống học sâu đối với dữ liệu, phần này đi qua các bộ dữ liệu hiện có cho VOS, tiếp theo là các số liệu đánh giá tương ứng và những khó khăn chính.

Datasets	D.type				DA	Resolution	Videos	Annotations	Categories	Objects
	R	S	S	M						
Hopkins-155 (Tron and Vidal 2007)	✓		✓	✓	320 × 240 – 640 × 480	155	4615	-	345*	
BMS-26 (Brox and Malik 2010b)	✓		✓		350 × 288 – 640 × 480	26	189	2	47	
FBMS-59 (Ochs et al. 2013)	✓		✓		350 × 288 – 960 × 540	59	720	11	139	
SegTrackV1 (Tsai et al. 2012)	✓	✓		✓	320 × 240 – 414 × 352	6	244	6	6	
SegTrackV2 (Li et al. 2013)	✓		✓	✓	320 × 240 – 640 × 360	14	1154	12	24	
YouTube-Objects (Prest et al. 2012)	✓	✓			320 × 240 – 960 × 540	126	2127	10	126	
JumpCut (Fan et al. 2015)	✓	✓	✓	✓	640 × 400 – 1280 × 720	22	6331	12	22	
DAVIS-2016 (Perazzi et al. 2016a)	✓		✓	✓	854 × 480	50	3455	-	50	
DAVIS-2017 (Pont-Tuset et al. 2017)	✓		✓	✓	854 × 480	150	10,459	-	376	
DAVIS-2017-U (Caelles et al. 2019)	✓		✓	✓	854 × 480	150	10,731	-	449	
YouTube-VOS-2018 (Xu et al. 2018b)	✓		✓		1280 × 720	4453	197,272	94	7754	
YouTube-VOS-2019 (Xu et al. 2019b)	✓		✓		1280 × 720	4519	>190,000	94	8614	
YouTube-VIS (Yang et al. 2019a)	✓		✓		1280 × 720	2883	>131,000	40	4883	
SAIL-VOS (Hu et al. 2019)		✓	✓	✓	1280 × 800	201	111,654	162	1,896,295	

Hình 5: 14 bộ dữ liệu video và các thuộc tính chính của chúng.

Dựa trên các thuộc tính này, các bộ dữ liệu được liệt kê được thảo luận chi tiết, đặc biệt là về các khó khăn và cài đặt áp dụng, để hướng dẫn các nhà nghiên cứu quan tâm đến VOS chọn bộ dữ liệu phù hợp để đào tạo và đánh giá các phương pháp của riêng họ.

• Hopkins-155

- thiết kế để đánh giá các thuật toán phân đoạn chuyển động dựa trên điểm, trong đó một tập hợp các điểm (từ 39 đến 550 điểm mỗi khung hình) được chú thích thay vì toàn bộ pixel.
- Tập dữ liệu này bao gồm các chuỗi được nhóm thành ba loại:
 - 1) bàn cờ: các đối tượng chuyển động được phủ bởi một mô hình bàn cờ để đảm bảo số lượng điểm được theo dõi;
 - 2) cảnh giao thông: các tình huống giao thông ngoài trời;
 - 3) các đối tượng khớp nối/không cứng nhắc: các chuỗi có chuyển động của các khớp, khuôn mặt và người đi bộ.

- Mặc dù Hopkins-155 có ích để đánh giá các phương pháp phân đoạn chống lật xoay, dịch và chuyển động suy biến, nhưng chú thích thưa thớt và thách thức hạn chế làm cho nó không phù hợp để đào tạo và đánh giá các phương pháp VOS dựa trên deep learning.

- **BMS (Berkeley Motion Segmentation Dataset) series**

- được thiết kế để phân đoạn đối tượng di chuyển và bao gồm hai phiên bản: BMS-26 (Brox và Malik, 2010b) và FBMS-59 (Ochs và cộng sự, 2013, Freiburg-BMS).
- BMS-26 bao gồm 26 chuỗi video, trong đó con người và ô tô là hai loại đối tượng được sử dụng phổ biến nhất.
- FBMS-59 mở rộng từ BMS-26 bằng cách tăng số lượng chuỗi video lên 59 và bao gồm nhiều loại đối tượng hơn.
- Trong cả hai bộ dữ liệu, những khó khăn như tắc nhẽn và biến đổi mô hình chuyển động đều xuất hiện, do đó những nhược điểm của các phương pháp VOS đều được đánh giá trên bộ dữ liệu này. Tuy nhiên với những dữ liệu được đánh dấu để huấn luyện vì độ phân giải không gian thấp và các khung được chú thích khá ít, điều này sẽ rất khó nếu muốn huấn luyện được một mô hình tốt từ bộ dữ liệu này

- **SegTrack series**

- Thiết kế cho việc phân đoạn và theo dõi đối tượng trong video và bao gồm 2 phiên bản: SegTrack v1 (Tsai et al. 2012) và SegTrack v2 (Li et al. 2013).
- SegTrack v1 chỉ chứa 6 chuỗi video, nhưng tất cả các khung hình đều được chú thích với các lớp cấp pixel.
- Sau khi thêm nhiều chuỗi video và đối tượng được chú thích hơn, SegTrack v2 mở rộng phiên bản SegTrack v1.
- Khó khăn: sự xuất hiện thường xuyên của chuyển động nhanh và biến dạng đối tượng, phân giải không gian tương đối thấp.

- **YouTube-Objets**

- Tập dữ liệu này bao gồm 126 chuỗi video với 2.127 khung hình được chú thích, và đã trở thành tập dữ liệu VOS lớn nhất vào thời điểm đấy. Tuy nhiên, do chú thích đối tượng ít và phân phối không đồng đều của các loại, đây không phải là một tập dữ liệu phù hợp cho việc huấn luyện.

- **JumpCut**

- bao gồm 22 chuỗi video với 6.331 khung hình, tất cả đều được chú thích với lớp cấp pixel. Ngoài các chuỗi video được ghi lại trong thế giới thực, tập dữ liệu còn bao gồm một lượng nhỏ các khung hình. Dựa trên các loại đối tượng tham gia (chủ yếu là con người và động vật) và các thử thách (chuyển động nhanh và các đối tượng tĩnh), tập dữ liệu được chia thành các nhóm khác nhau để tổ chức tốt hơn.

- **DAVIS (Densely Annotated Video Segmentation) series**

- DAVIS (Densely Annotated VIdeo Segmentation), đã phát triển qua các phiên bản trong nhiều năm với ba phiên bản: DAVIS-2016 (Perazzi và cộng sự, 2016a), DAVIS-2017 (Pont-Tuset và cộng sự, 2017) và DAVIS-2017-U (Caelles và cộng sự, 2019), tương ứng với các loại nhiệm vụ VOS khác nhau.

- **YouTube-VOS series**

- bao gồm ba phiên bản: YouTube-VOS 2018 (Xu et al. 2018b), YouTube-VOS 2019 (Xu et al. 2019b) and YouTube-VIS (Yang et al. 2019a).
- Hai phiên bản đầu tiên được thiết kế cho SVOS đa đối tượng, trong khi phiên bản sau phục vụ cho UVOS đa đối tượng.

- Mỗi chuỗi video trong các tập dữ liệu có số lượng khung hình lớn hơn bất kỳ tập dữ liệu nào khác, cho phép các phương pháp VOS mô hình hóa và khai thác sự phụ thuộc thời gian dài hạn giữa các khung hình.

- **SAIL-VOS (Semantic Amodal Instance Level Video Object Segmentation)**

- Đây là bộ dữ liệu tổng hợp cho VOS (Hu et al. 2019), nơi tất cả các khung hình video và lớp tương ứng được thu thập từ Grand Theft Auto V, một trò chơi phiêu lưu hành động. Các hình ảnh trong trò chơi được hiển thị chân thực nhất có thể, do đó nó rất hữu ích cho việc đào tạo và đánh giá các phương pháp VOS. Ngoài ra, vì tất cả các chuỗi video được tạo bởi trình mô phỏng trò chơi, lớp đối tượng thu được là hoàn toàn đáng tin cậy, ngay cả khi chúng đang gấp phải sự tắc nghẽn nặng nề.

- **Evaluation metrics**

- Trong VOS, các số liệu thường được sử dụng để đánh giá hiệu suất là chỉ số Jaccard (Everingham et al. 2010), F-measure (Martin et al. 2004), và giá trị trung bình của chúng:

$$\left\{ \begin{array}{l} \mathcal{J} = \frac{|M \cap G|}{|M \cup G|} \\ \mathcal{F} = \frac{2P_cR_c}{P_c + R_c} \\ \mathcal{J}\&\mathcal{F} = \frac{(\mathcal{J} + \mathcal{F})}{2} \end{array} \right.$$

- Trong đó G và M lần lượt đề cập đến mặt nạ chân lý mặt đất và mặt nạ phân đoạn. J đánh giá sự tương đồng khu vực giữa hai loại khẩu trang này. P và R_c là độ chính xác và khả năng nhớ lại được tính toán từ các điểm trong đường viền $c(M)$ và $c(G)$. Do đó F đánh giá tính chính xác của việc khoanh vùng ranh giới. J&F đo lường hiệu suất VOS tổng thể.

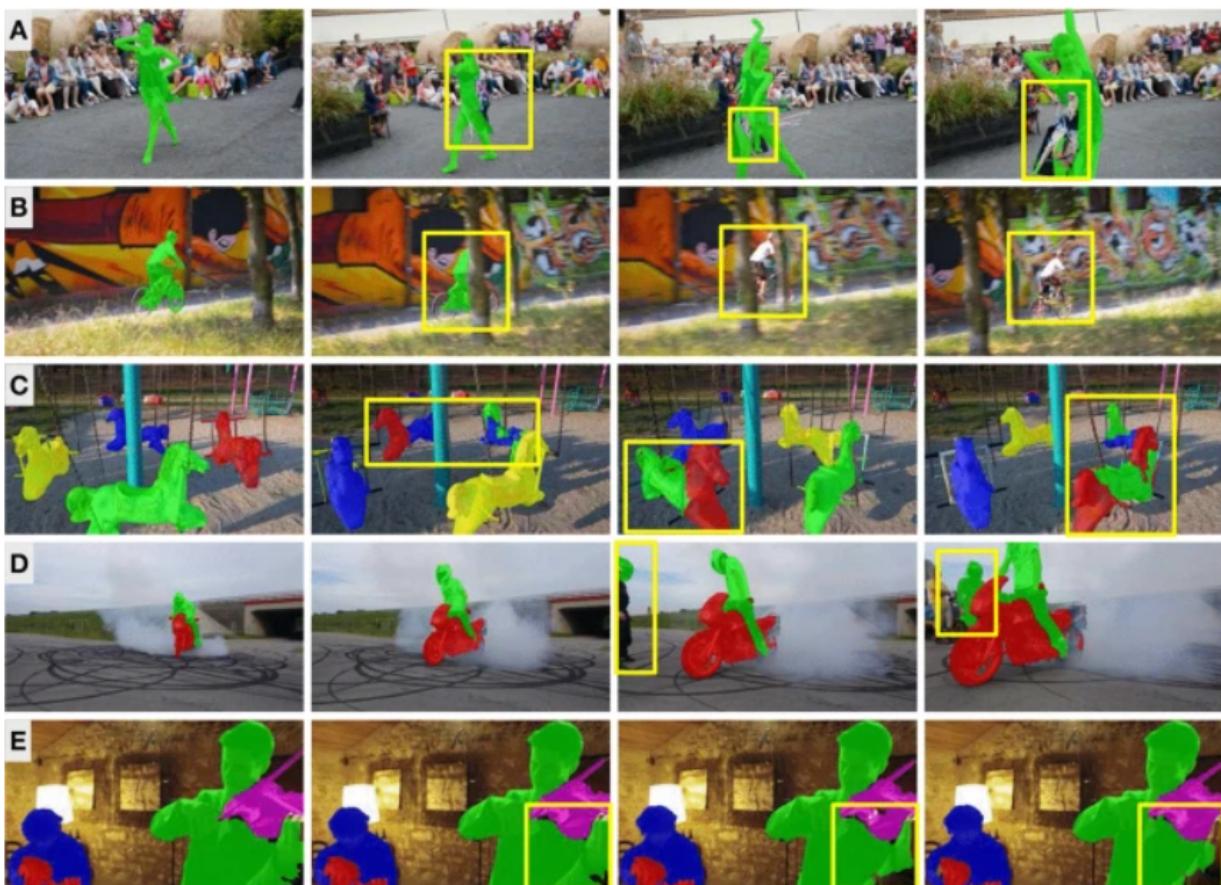
- **Summary**

- Hopkins-155, BMS series, SegTrack series: Các bộ dữ liệu sớm này ban đầu được thiết kế để đánh giá các phương pháp không sâu học nhưng vẫn có thể đánh giá hiệu suất của các phương pháp VOS, đặc biệt là trong việc xử lý biến dạng đối tượng và che khuất. Tuy nhiên, chúng ít được sử dụng trong các phương pháp gần đây do hạn chế về đa dạng dữ liệu, số lượng thách thức và độ dài video.
- YouTube-Objects, JumpCut: Các bộ dữ liệu này bao gồm video có phạm vi và độ phân giải cao và được ưa chuộng để đánh giá các phương pháp VOS sớm hơn trong việc nhúng đặc trưng không gian-thời gian. Tuy nhiên, chúng cũng có hạn chế về đa dạng dữ liệu và thử thách nhận diện đối tượng, và chỉ có một số phương pháp gần đây được đánh giá trên chúng.
- SAIL-VOS: Khác với các bộ dữ liệu khác, SAIL-VOS bao gồm các video tổng hợp. Mặc dù vẫn còn khoảng cách giữa các khung hình video được tạo ra và thực tế, nó cung cấp các che khuất đáng tin cậy và được kiểm soát, có thể cải thiện tính mạnh mẽ của các phương pháp VOS đối mặt với các che khuất. Tuy nhiên, nó chưa được sử dụng rộng rãi trong các phương pháp được xem xét.
- DAVIS series, YouTube-VOS series: Đây là các bộ dữ liệu được sử dụng phổ biến nhất để huấn luyện và đánh giá các phương pháp VOS gần đây. Chúng cung cấp chuỗi video quy mô lớn, các loại đối tượng đa dạng, nhiều thách thức và các chủ thích chất lượng cao. Loạt DAVIS và loạt YouTube khác nhau về chủ thích, cho phép đánh giá các thuộc

tính VOS khác nhau. DAVIS được ưa chuộng để đánh giá tính ổn định thời gian, trong khi loạt YouTube được ưa chuộng để hiệu suất tổng quát do số lượng video và loại đối tượng lớn hơn.

5. Những khó khăn, thử thách phổ biến:

- Giới thiệu một số thách thức đối với các trường UVOS và SVOS, bao gồm thay đổi thuộc tính, tắc, xung đột giữa các trường hợp tương tự, nền không rõ, tính nhất quán tạm thời và sự cân bằng giữa hiệu quả và độ chính xác.

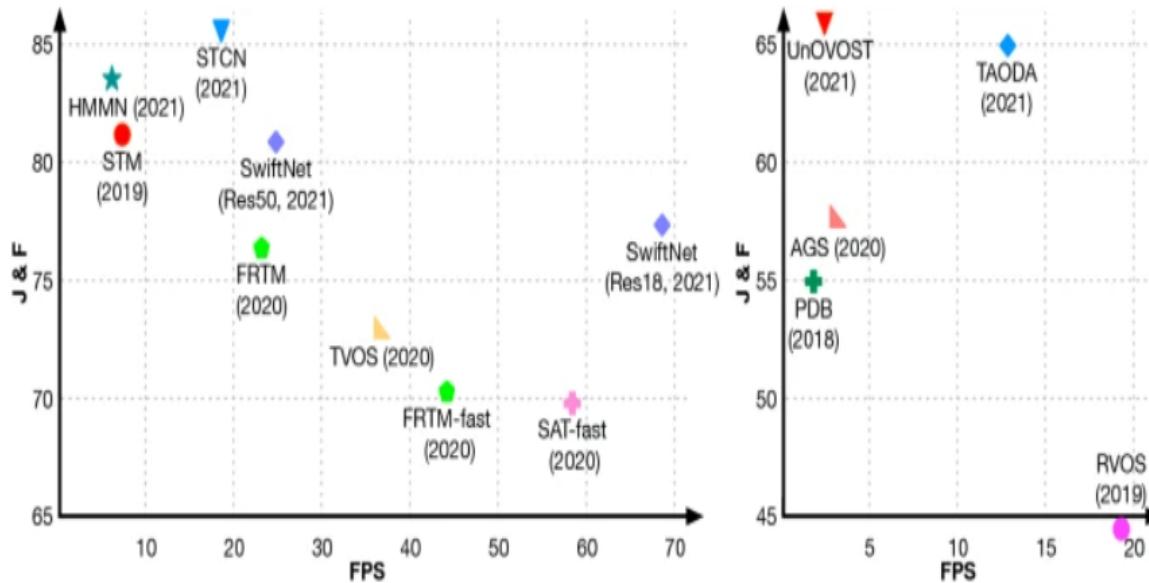


Hình 6: Kết quả sai

Mỗi hàng cho thấy ảnh hưởng của một yếu tố trở ngại đối với các phương pháp VOS hiện có.

- Một tài sản thay đổi;*
- tắc;*
- phân biệt đối xử giữa các đối tượng tương tự;*
- hình nền mơ hồ;*
- VOS nhất quán về mặt thời gian (hàng này bao gồm các khung hình liên tục không có chuyển động nhanh, tắc nghẽn và thay đổi diện mạo đáng kể).*

Hộp màu vàng làm nổi bật kết quả sai



Hình 7: Độ chính xác và hiệu quả phân đoạn của các phương pháp SVOS (trái) và UVOS (phải) gần đây trên bộ xác thực DAVIS-2017

- **Thay đổi thuộc tính đối tượng**

- Trở ngại này chủ yếu ảnh hưởng đến các phương pháp VOS dựa trên sự tương đồng về hình ảnh. Trong quá trình suy luận, các phương pháp này phân đoạn các vùng có đặc điểm hình ảnh tương tự với các đối tượng đích được chú thích (hầu hết là phương pháp SVOS) hoặc dự đoán (hầu hết là phương pháp UVOS).

- **Bị che khuất bởi những yếu tố gây nhiễu**

- Trở ngại này chủ yếu ảnh hưởng đến các phương pháp VOS dựa trên lan truyền, xem xét các đối tượng được dự đoán trong khung trước đó để ước tính phân đoạn khung hình hiện tại.

- **Bị nhiễu bởi nền/ đối tượng tương tự**

- Trở ngại này chủ yếu ảnh hưởng đến các phương pháp VOS dựa trên sự tương đồng về hình ảnh, độ nổi bật hoặc mô hình chuyển động.

- **Nhất quán về thời gian**

- Trở ngại này chủ yếu ảnh hưởng đến các phương pháp VOS sử dụng ít thông tin chuyển động hơn. Trong quá trình suy luận, các phương pháp này về cơ bản thực hiện phân đoạn hình ảnh cho từng khung hình video. Do đó, rất khó để duy trì tính nhất quán về mặt thời gian của các đối tượng được phân đoạn, tức là sự phát triển của mặt nạ đối tượng được dự đoán từ các khung hình liên tục không trơn tru (trong trường hợp không có tắc, chuyển động nhanh hoặc thay đổi thuộc tính đáng kể).

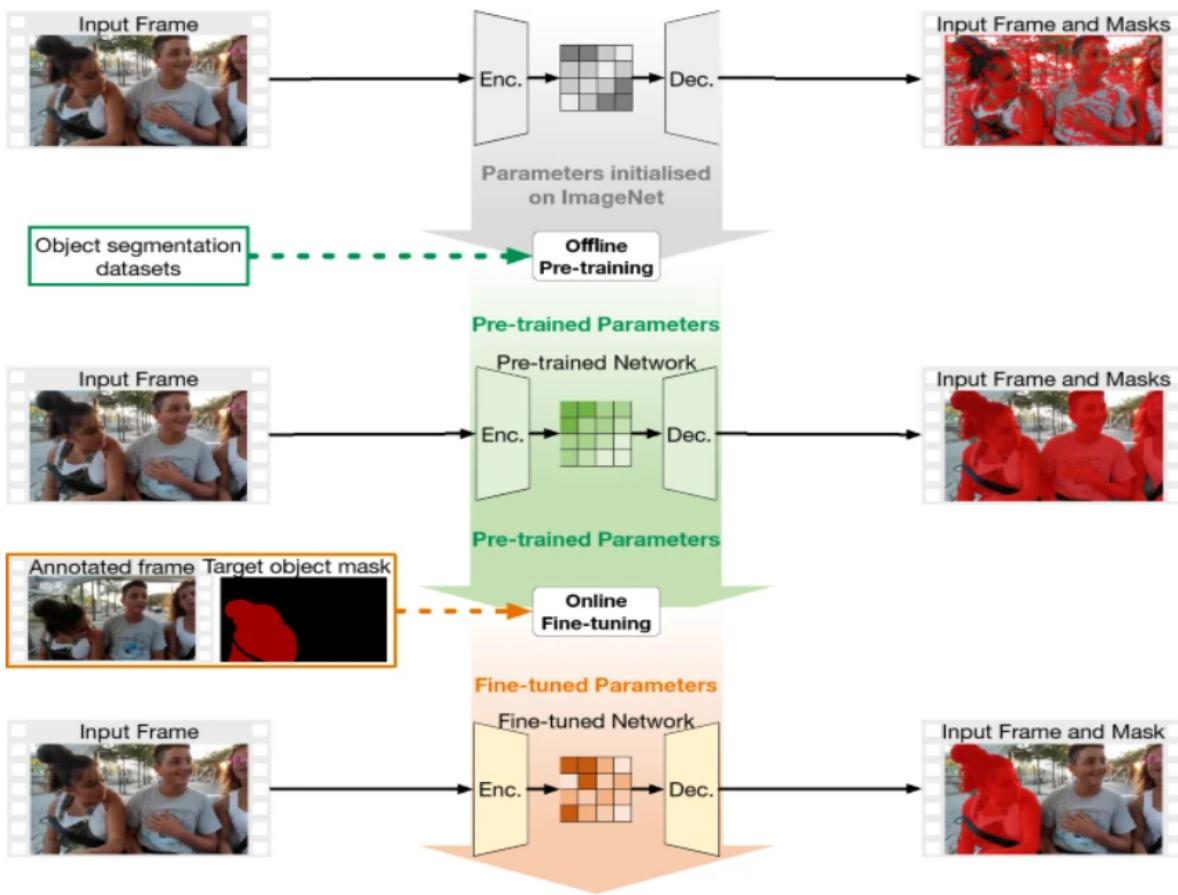
- **Cân bằng giữa độ chính xác và hiệu quả của VOS**

- Trở ngại này chủ yếu ảnh hưởng đến các phương pháp VOS phục vụ các ứng dụng thời gian thực. Nói chung, các phương pháp này nên thực hiện phân đoạn ít nhất 24 FPS (Khung hình mỗi giây) trong khi đạt được lớp đối tượng chất lượng cao.

6. Phương pháp

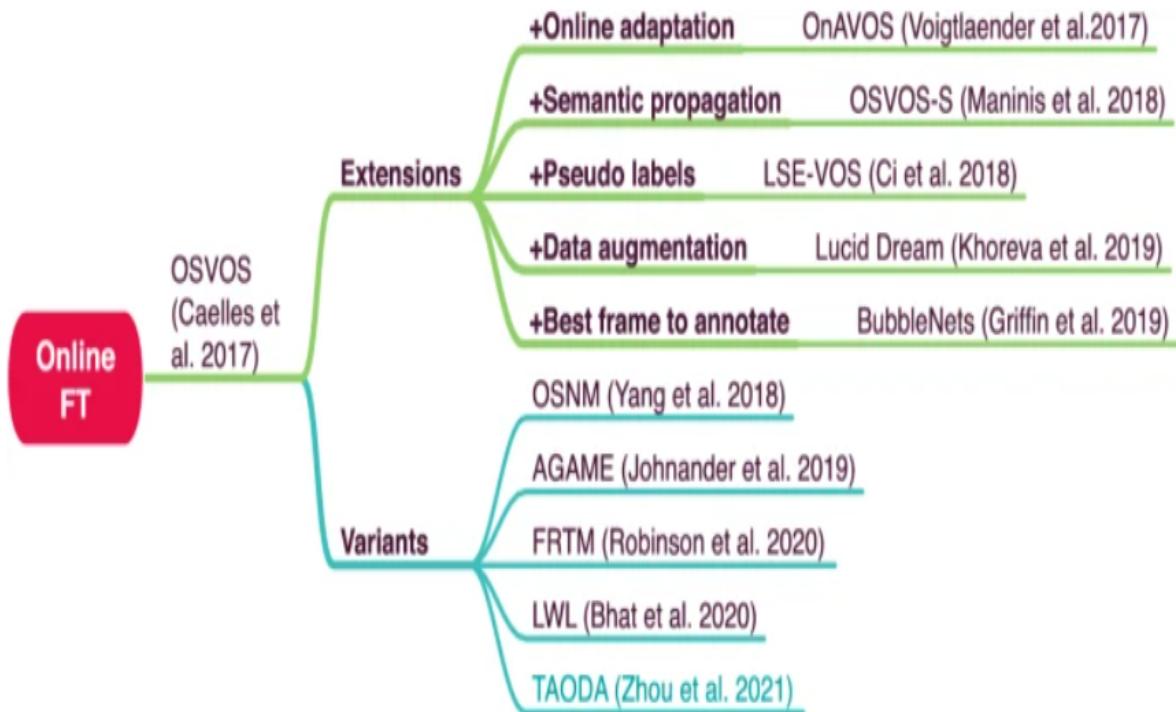
- **Các phương pháp dựa trên tinh chỉnh trực tuyến**

- Có ba giai đoạn chính để chuyển miền đầu ra của mạng phân đoạn từ kiến thức chung sang đối tượng được chú thích:
 - (1) Khởi tạo mạng (màu xám) với các tham số được đào tạo trước trên ImageNet (Russakovsky et al. 2015);
 - (2) đào tạo trước mạng (màu xanh lá cây) trên các bộ dữ liệu phân đoạn đối tượng (ví dụ: MS-COCO (Lin et al. 2014) và DAVIS (Perazzi et al. 2016a));
 - (3) Tinh chỉnh mạng (màu vàng) trên khung chú thích. Vì đào tạo trước và tinh chỉnh được thực hiện trước và trong quá trình suy luận, chúng tôi gọi chúng là các quy trình "ngoại tuyến" và "trực tuyến", tương ứng.



Hình 8: Sơ đồ phương pháp VOS dựa trên tinh chỉnh trực tuyến

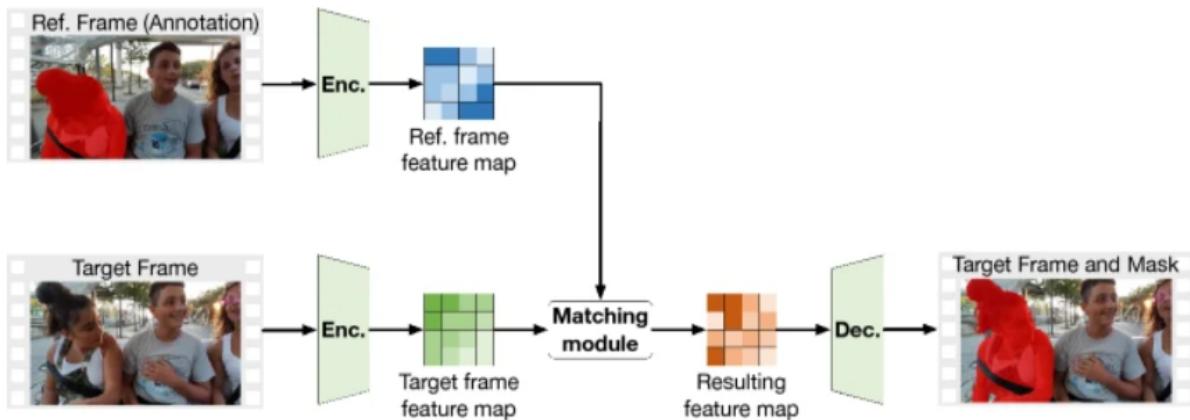
- OSVOS là phương pháp đầu tiên sử dụng tinh chỉnh trực tuyến cho SVOS. Sau đó, một số phương pháp có nguồn gốc từ OSVOS. Các sửa đổi chính của họ đối với OSVOS được đánh dấu bằng các từ in đậm với tiền tố "+". Gần đây, một số biến thể của tinh chỉnh trực tuyến được đề xuất cho VOS hiệu quả.



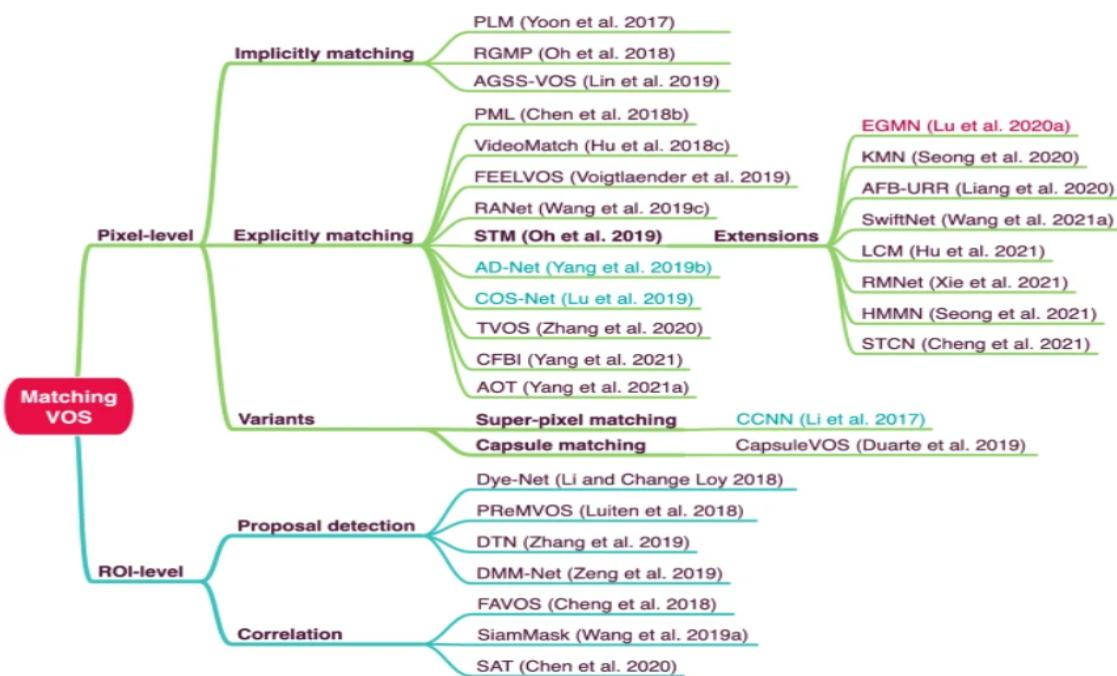
Hình 9: Lộ trình phát triển các phương pháp dựa trên tinh chỉnh trực tuyến.

- Phương pháp VOS dựa trên tinh chỉnh trực tuyến và các biến thể của chúng, xuất phát từ OSVOS (Caelles et al. 2017). Các phương pháp này chuyển đổi đầu ra của mạng từ kiến thức chung sang các đối tượng cụ thể bằng cách tinh chỉnh các tham số mạng với các chú thích khung đầu tiên. Tuy nhiên, các vấn đề trong OSVOS đã thúc đẩy sự phát triển của các phương pháp mở rộng và biến thể: Chỉ xem xét các chú thích khung đầu tiên, làm giảm hiệu quả VOS.
- **Phản mở rộng hoạt động nhằm cải thiện OSVOS về khả năng thích ứng và mạnh mẽ,**
 - * OnAVOS (Voigtlaender và Leibe 2017) và LSE-VOS (Ci et al. 2018): Cải thiện khả năng thích ứng bằng cách kết hợp kết quả có độ tin cậy cao từ các khung hình trong quá khứ. Tuy nhiên, cần phải tinh chỉnh nhiều lần trực tuyến, làm giảm hiệu quả VOS.
 - * OSVOS-S (Maninis et al. 2018) và LucidTracker (Khoreva et al. 2019): Tăng cường sự mạnh mẽ của phân đoạn. OSVOS-S kết hợp kiến thức từ phân đoạn đối tượng chung để tinh chỉnh kết quả VOS. LucidTracker đạt được điều này bằng cách tạo ra các mẫu đa dạng từ các chú thích. Các phương pháp này thúc đẩy hiệu suất VOS nhưng vẫn kém hiệu quả hơn do cần tính toán thêm hoặc tăng cường dữ liệu.
 - * BubbleNets (Griffin và Corso 2019): Là một phương pháp gia tăng, có thể được tích hợp vào các phương pháp trên để dự đoán khung hình tối ưu để chú thích, giúp tạo ra một khung tối ưu với chi phí hiệu quả.
- **Các biến thể tập trung vào hiệu quả VOS hơn.** OSNM, A-GAME, FRTM, LWL và TAODA: Phát triển để thay đổi miền đầu ra mạng với các thuật toán hiệu quả hơn. Mặc dù đạt được hiệu quả tốt hơn, vẫn còn khoảng cách chính xác giữa các biến thể.
- **Phương pháp dựa trên kết hợp** Phương pháp này thực hiện VOS bằng cách đo sự tương

ứng giữa khung mục tiêu và khung tham chiếu

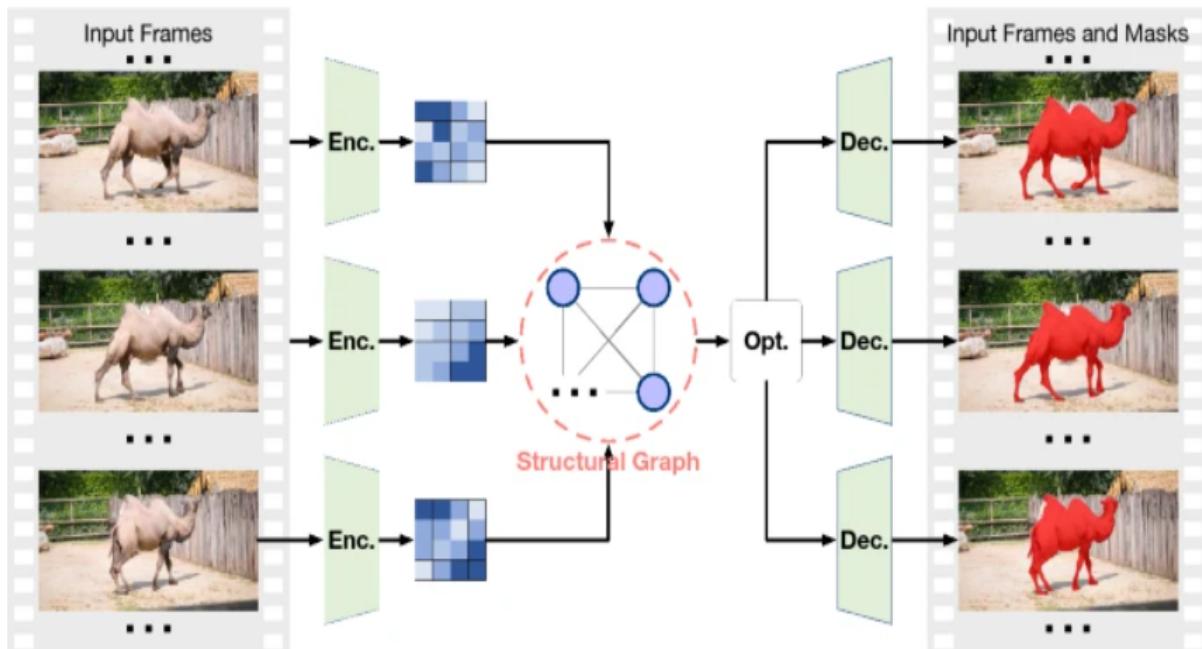


Hình 10: Sơ đồ các phương pháp VOS dựa trên kết hợp

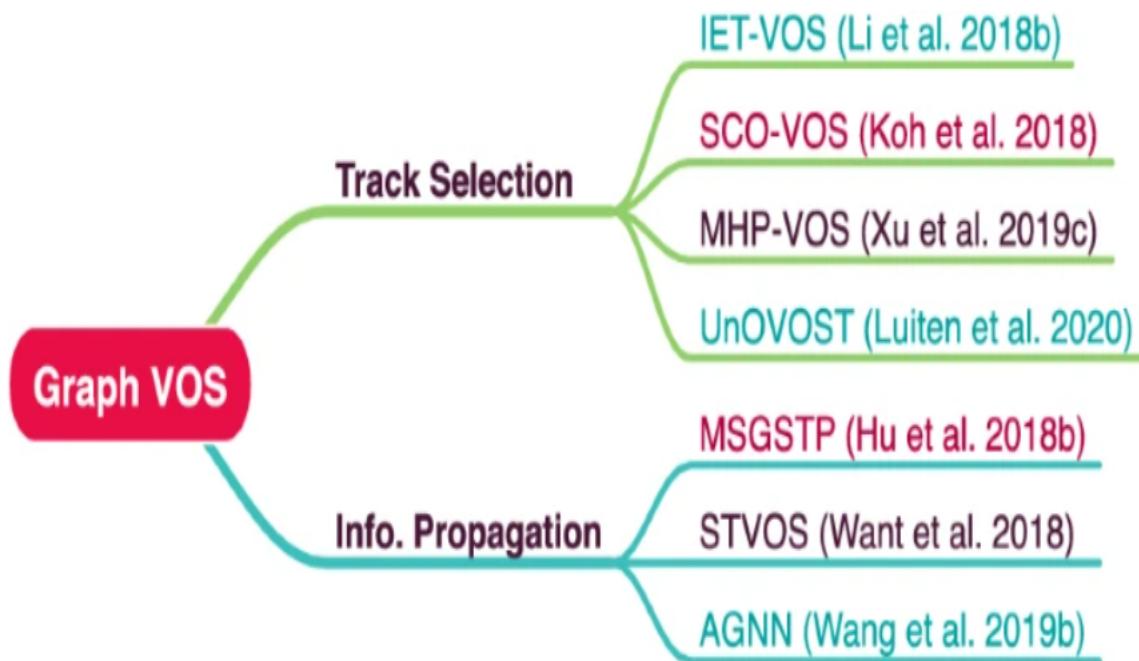


Hình 11: Lộ trình phát triển các phương pháp VOS dựa trên kết hợp

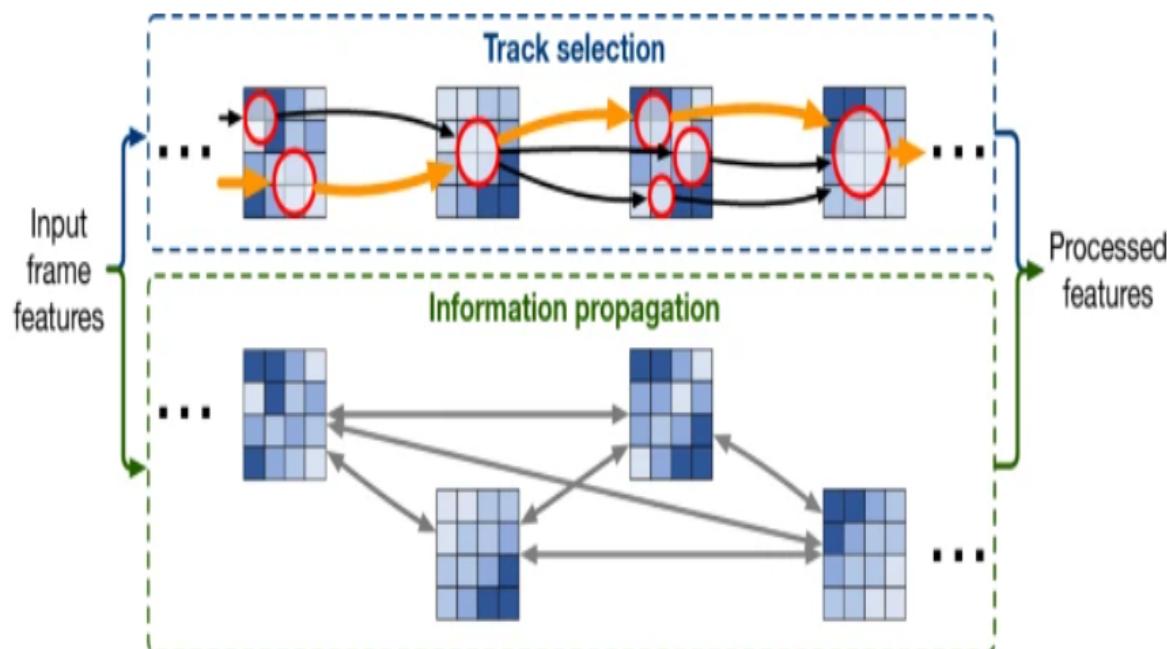
- Hầu hết các phương pháp SVOS hàng đầu hiện nay đã được phát triển dựa trên tính năng phù hợp, và các phương pháp này đã đạt được hiệu quả cao trong các thử nghiệm. Sự tương ứng đáng tin cậy cũng mang lại kết quả chất lượng cao cho các phương pháp UVOS.
 - Các phương pháp SVOS dựa trên tính năng phù hợp thường kết hợp cấp pixel, đo lường sự tương ứng dày đặc giữa các khung hình. Phương pháp này có thể tạo ra các kết quả mạnh mẽ mà không cần tinh chỉnh trực tuyến, đồng thời giúp giảm thiểu thời gian tính toán.
 - Phương pháp dựa trên sơ đồ rõ ràng đo trực tiếp sự tương đồng giữa pixel của khung mục tiêu và khung tham chiếu, mà không cần thêm mô-đun cho thư từ. Các phương pháp này cải thiện hiệu suất phân đoạn bằng cách kết hợp sự tương đồng với các tính năng chi tiết trong khung mục tiêu.
 - Sự khác biệt của chúng chủ yếu nằm ở các thực thể phù hợp và cách tiếp cận để định vị đối tượng mục tiêu. Vì các phương pháp dựa trên tương quan không yêu cầu bất kỳ mạng bổ sung nào để tạo ROI, chúng thực hiện VOS nhanh hơn nhiều so với các mạng dựa trên phát hiện. Tuy nhiên, hiệu suất phân đoạn của phương pháp này bị hạn chế vì các chi tiết chi tiết bị bỏ qua trong mô-đun phân đoạn dựa trên ROI.
- **Phương pháp dựa trên tối ưu hóa đồ thị** có hai loại cách tiếp cận để tổ chức và phân tích các nút đồ thị: lựa chọn theo dõi và truyền bá thông tin



Hình 12: Sơ đồ các phương pháp VOS dựa trên tối ưu hóa đồ thị, áp dụng cho cả SVOS và UVOS.



Hình 13: Lộ trình phát triển của các phương pháp dựa trên tối ưu hóa đồ thị

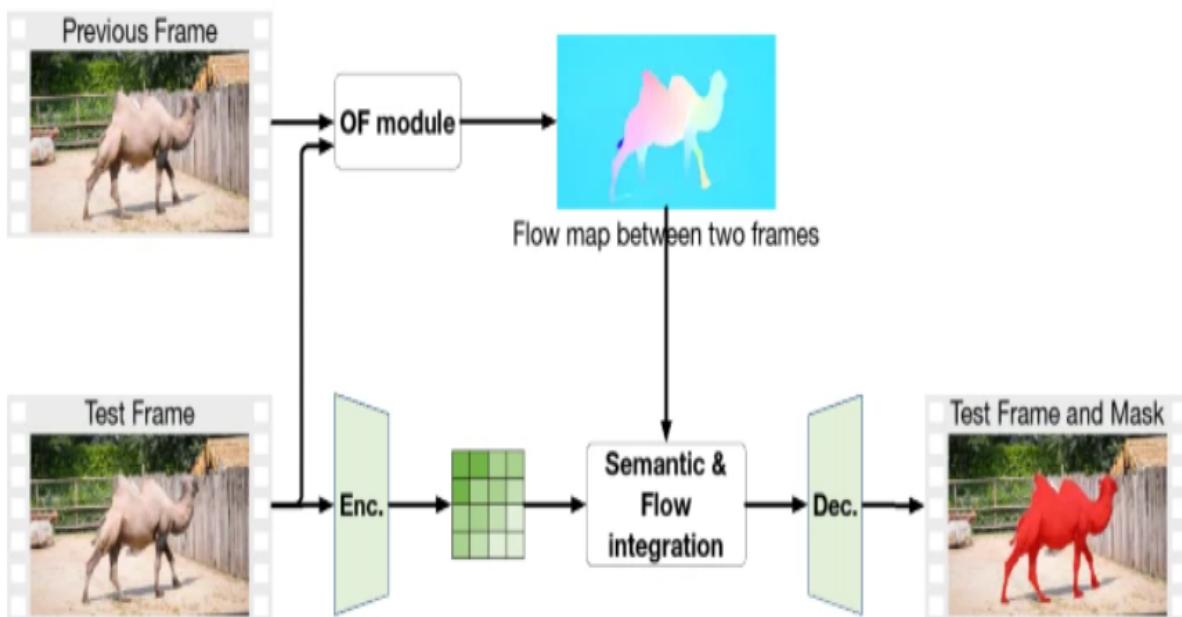


Hình 14: Sơ đồ hai kỹ thuật để tổ chức và phân tích nút đồ thị

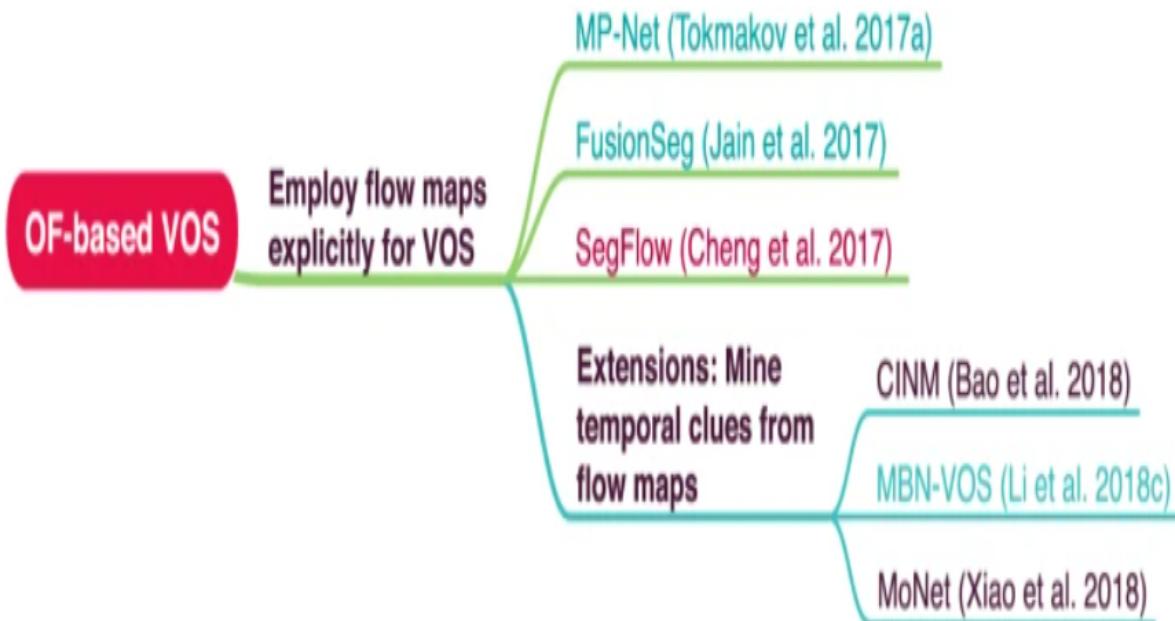
- Có hai phương án chính để tổ chức và phân tích dữ liệu đồ thị trong VOS: lựa chọn theo dõi và truyền bá thông tin. Sơ đồ dựa trên lựa chọn theo dõi tổ chức các nút thành một tập hợp các rãnh hoặc cây, từ đó tạo mặt nạ đối tượng thông qua việc truy xuất các bản nhạc tối ưu. Các phương pháp này thường yêu cầu các mạng bổ sung để tạo các đề xuất đối tượng, làm tăng chi phí tính toán và số lượng tham số mạng.
- Sơ đồ dựa trên truyền bá thông tin hỗ trợ VOS bằng cách truyền thông tin lặp đi lặp lại giữa các nút. Các phương pháp này tập trung vào việc truyền bá thông tin liên quan đến nhãn, làm cho kết quả phân đoạn nhạy cảm với các kết nối giữa các nút. Công việc gần đây đã giảm thiểu vấn đề này bằng cách xem xét các tính năng sâu sắc và truyền bá thông tin cấp khung.

Tóm lại, tối ưu hóa đồ thị mang lại sự tương ứng rộng rãi hơn với các phương pháp VOS, cho phép đạt được kết quả phân đoạn chất lượng cao, đặc biệt là trong UVOS. Tuy nhiên, kỹ thuật này có chi phí tính toán cao, không phù hợp cho các tác vụ VOS giới hạn tài nguyên và thời gian thực.

- **Phương pháp dựa trên dòng chảy quang học** Luồng quang học đã là một kỹ thuật được sử dụng rộng rãi trong VOS do các mẫu chuyển động ở cấp độ pixel. Kỹ thuật này giả định rằng đối tượng mục tiêu và nền có các kiểu chuyển động khác nhau. Do đó, việc tích hợp luồng quang vào VOS có thể cung cấp các mạng phân đoạn với các ưu tiên hợp lý

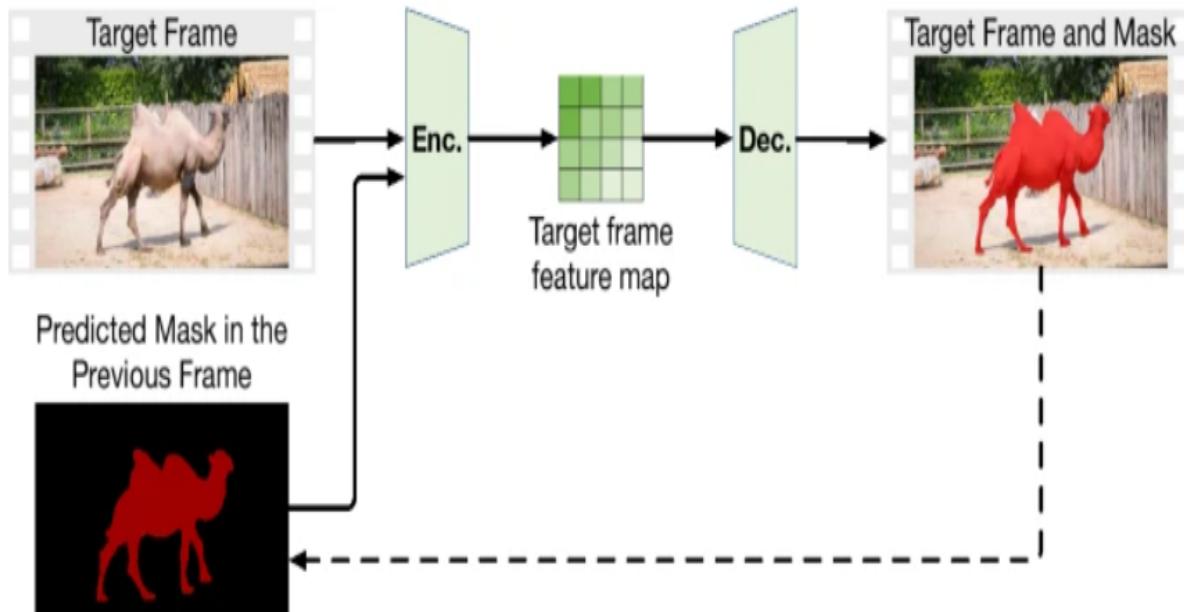


Hình 15: Sơ đồ các phương pháp VOS dựa trên luồng quang, áp dụng cho cả SVOS và UVOS

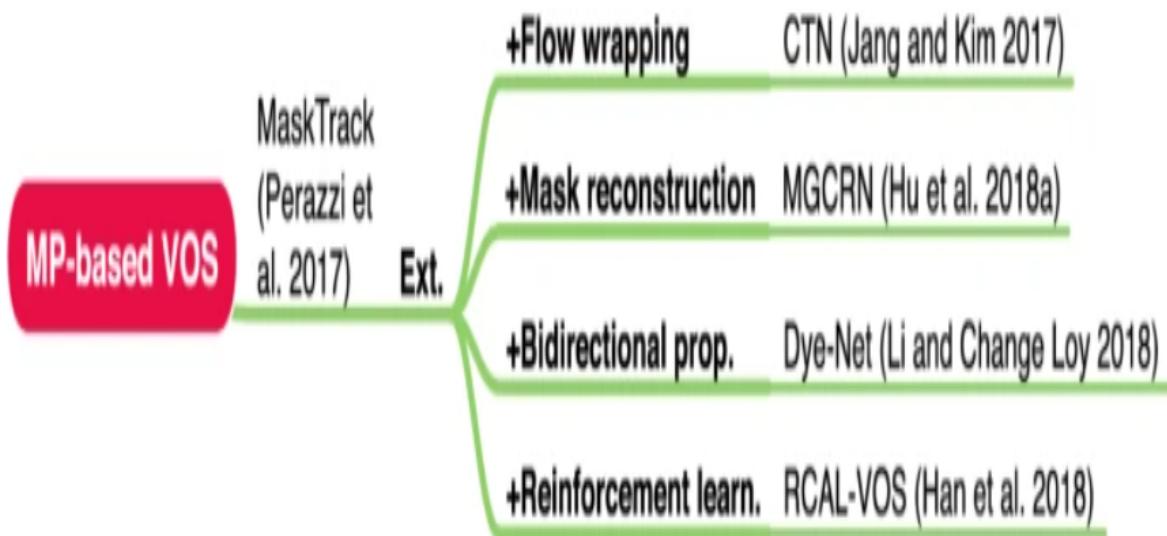


Hình 16: Lộ trình phát triển của các phương pháp VOS dựa trên luồng quang

- Phần này tập trung vào các phương pháp VOS dựa trên luồng quang, giả định rằng đối tượng và nền di chuyển theo các mẫu khác nhau. Các phương pháp này sử dụng các bản đồ dòng chảy để ước tính hình dạng và vị trí của đối tượng mục tiêu. Các phương pháp trước đó đã tích hợp rõ ràng các bản đồ dòng chảy với các tính năng không gian để tạo lớp đối tượng. Tuy nhiên, việc ước tính luồng không luôn đáng tin cậy do thiếu dữ liệu đào tạo và các thách thức như nền động. Vì vậy, các phương pháp gần đây đã được đề xuất để khai thác luồng quang hiệu quả hơn và tránh những rủi ro này. *Mặc dù đã đạt được kết quả tốt trên nhiều chuỗi thách thức, việc sử dụng luồng quang hiện đã ít phổ biến trong các hệ thống VOS gần đây do một số hạn chế như sự không phân biệt đối xử giữa đối tượng và nền trong một số trường hợp, cũng như yêu cầu các mạng sâu bổ sung trong quá trình suy luận. Do đó, các phương pháp mới sử dụng nhân giống lớp đã thay thế dần luồng quang.*
- Phương pháp dựa trên nhân giống lớp



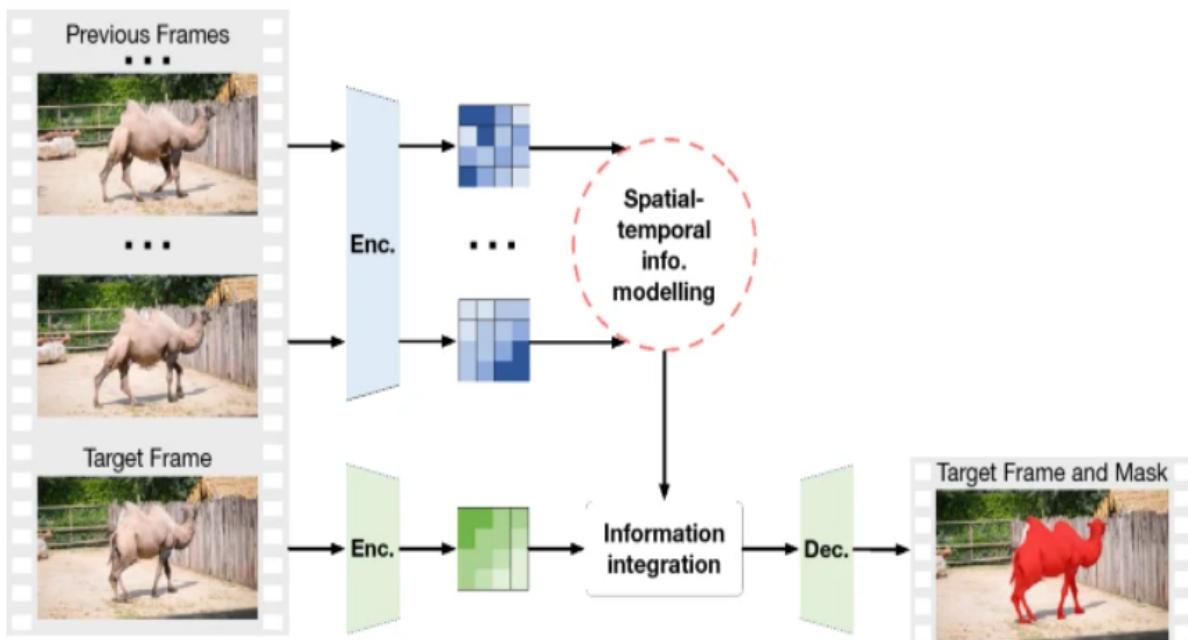
Hình 17: Sơ đồ phương pháp VOS dựa trên nhân giống lớp



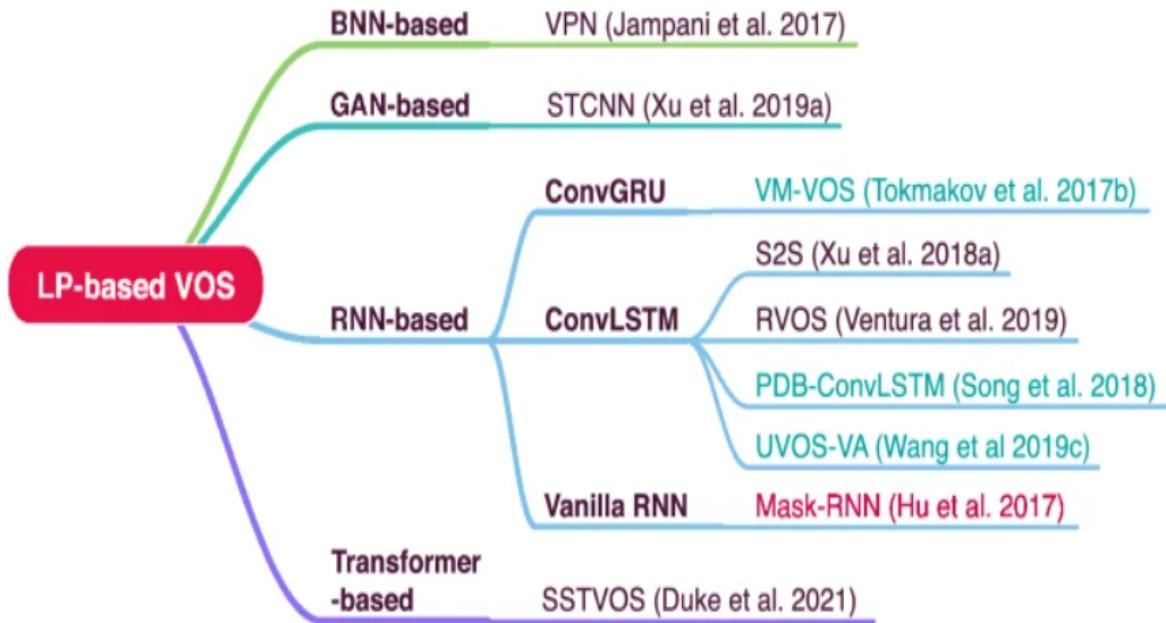
Hình 18: Lộ trình phát triển VOS dựa trên tuyên truyền lớp

– phương pháp này sử dụng lớp đà có để ước tính vị trí và hình dạng của các đối tượng trong các khung tiếp theo. Mặc dù đem lại kết quả chất lượng cao, nhưng các phương pháp này gặp thách thức khi xử lý biến dạng mạnh và chuyển động đột ngột. Để cải thiện độ tin cậy, các phương pháp như CTN và MGCRN tích hợp thông tin về luồng quang, trong khi các phương pháp như DyeNet và RCAL-VOS sử dụng các kỹ thuật nhân giống lớp hai chiều để đối phó với các thách thức này. Tổng quan, việc sử dụng nhân giống lớp đà đóng góp rất nhiều cho các phương pháp VOS bằng cách cung cấp thông tin về vị trí và hình dạng của các đối tượng từ các khung hình trước.

- **Phương pháp dựa trên nhân giống thời gian dài hạn**



Hình 19: Sơ đồ các phương pháp VOS dựa trên sự lan truyền thời gian dài hạn, áp dụng cho cả SVOS và UVOS



Hình 20: Theo dõi phát triển các phương pháp VOS dựa trên sự lan truyền thời gian dài hạn (viết tắt là 'VOS dựa trên LP')

- Phần này bàn về các phương pháp VOS dựa trên sự lan truyền thời gian dài hạn, tích lũy thông tin không gian-thời gian qua nhiều khung hình để ước tính hình dạng và vị trí của các đối tượng. Có bốn loại kỹ thuật chính để đạt được điều này, bao gồm BNN, phương pháp dựa trên GAN, các phương pháp VOS dựa trên RNN, và các phương pháp kết hợp để tăng cường hiệu suất.
- Mặc dù việc sử dụng nhãn giống dài hạn mang lại nhiều lợi ích, nhưng hiện tại ít được sử dụng và không đem lại kết quả tốt hơn các phương pháp khác. Một số thách thức gặp phải là tính toán chậm và việc sử dụng các khung ngắn hạn trong quá trình đào tạo. Đề xuất tương lai tập trung vào cách truyền thông tin dài hạn trong điều kiện tài nguyên hạn chế.

7. Kết quả thí nghiệm và thảo luận

- Kết quả định lượng và định tính

Bảng 1: Độ chính xác và hiệu quả phân đoạn của các phương pháp SVOS đại diện trên bộ xác thực DAVIS-2016

Methods	S. techs			T. techs			Frames	Resolutions	J&T	FPS ↑
	O	M	G	O	P	L				
OSVOS	x						1	480 x 854	80.2	0.38
MSKTrack	x			x	x		1, t-1	480 x 854	77.6	0.29
OSMN					x		1, t-1	480 x 854	73.3	1.87
RGMP		x		x		x	1, t-1	480 x 854	81.8	12.4
SiamMask		x			x		1, t-1	480 x 854	69.8	79.6
A-GAME					x		1, t-1	480 x 854	81.9	12.6
RVOS						x	[1, t-1, 1]	240 x 427	72.3	193.7
RANet		x			x		1, t-1	480 x 854	87.1	41.3
STM		x			x		[1, t-1, 5]	480 x 854	89.4	11.9

- **Thảo luận và định hướng nghiên cứu trong tương lai**

- Bộ dữ liệu đào tạo quy mô lớn và chú thích dày đặc: Cần có các bộ dữ liệu đào tạo có chú thích dày đặc và quy mô lớn để cung cấp thông tin đầy đủ cho mô hình VOS. Hiện nay, DAVIS là một trong những bộ dữ liệu chính phục vụ cho mục đích này, nhưng cần phát triển thêm các bộ dữ liệu tương tự để cải thiện hiệu suất.
- Phân tích thông tin thời gian dài hạn: Cần nghiên cứu và phát triển các phương pháp VOS có khả năng phân tích thông tin thời gian dài hạn từ chuỗi video, giúp giải quyết các thách thức như tắc, khuất tầm nhìn và chuyển động nhanh.
- Cân bằng giữa độ chính xác và hiệu quả của VOS: Cần phát triển các mô hình VOS có khả năng cân bằng giữa độ chính xác và hiệu quả, bằng cách thiết kế các mạng nhẹ nhàng nhưng vẫn đảm bảo hiệu suất.
- UVOS đa đối tượng: Cần phát triển các phương pháp UVOS cho phép phân biệt nhiều đối tượng trong chuỗi video, thay vì chỉ xác định một đối tượng duy nhất. Điều này sẽ tạo điều kiện cho ứng dụng thực tế hơn trong VOS.