Article

Not peer-reviewed version

Segment Anything Model (SAM 2) Unveiled: Functionality, Applications, and Practical Implementation Across Multiple Domains

Jahanggir Hossain Setu *, Mahmudul Islam *, Syed Tangim Pasha *, Nabarun Halder *, Ekram Hossain, Asif Mahmud, Ashraful Islam *, Md. Zahangir Alam, M. Ashraful Amin

Posted Date: 26 August 2024

doi: 10.20944/preprints202408.1790.v1

Keywords: Segment Anything Model 2; SAM 2; Video Segmentation; Real-Time Segmentation; Video Processing; Memory Attention; Streaming Memory; Interactive Segmentation; Spatio-Temporal Segmentation; Promptable Segmentation

 doi:10.20944/preprints202408.1790.v1

Article

Segment Anything Model (SAM 2) Unveiled: Functionality, Applications, and Practical Implementation across Multiple Domains

Jahanggir Hossain Setu *, Mahmudul Islam *, Syed Tangim Pasha *, Nabarun Halder *, Ekram Hossain, Asif Mahmud, Ashraful Islam, Md. Zahangir Alam and M. Ashraful Amin

Center for Computational & Data Sciences, Independent University Bangladesh, Dhaka, Bangladesh

* Correspondence: jahanggir15-10533sets@iub.edu.bd (J.H.S.); mahmud@iub.edu.bd (M.I.); syed15-8631sets@iub.edu.bd (S.T.P.); nabarun.earthsets@iub.edu.bd (N.H.)

Abstract: Segment Anything Model 2 (SAM 2) is a state-of-the-art development by Meta AI Research, designed to address the limitations of its predecessor, SAM, particularly in the realm of video segmentation. SAM 2 employs a transformer-based architecture enhanced with streaming memory, enabling real-time processing for both images and videos. This advancement is important given the exponential growth of multimedia content and the subsequent demand for efficient video analysis. Utilizing the SA-V dataset, SAM 2 excels in handling the intricate spatio-temporal dynamics inherent in video data, ensuring accurate and efficient segmentation. Key features of SAM 2 include its ability to provide real-time segmentation with minimal user interaction, maintaining robust performance even in dynamic and cluttered visual environments. This study provides a comprehensive overview of SAM 2, detailing its architecture, functionality, and diverse applications. It further explores the model's potential in improving practical implementations across various domains, emphasizing its significance in advancing real-time video analysis.

Keywords: Segment Anything Model 2; SAM 2; video segmentation; real-time segmentation; video processing; memory attention; streaming memory; interactive segmentation; spatio-temporal segmentation; promptable segmentation

1. Introduction

Image segmentation is a crucial component in several systems that aim to comprehend visual data. The technique divides images—or video frames—into multiple elements or sections [1]. Image segmentation refers to assigning semantic labels to pixels, also known as semantic segmentation. It can also involve the separation of distinct objects, which is called instance segmentation [2]. Image categorization is a less difficult task than semantic segmentation. It involves labeling each pixel in an image with a specific item category, e.g., human, car, tree, or sky. On the other hand, image classification involves predicting a solitary label that represents the complete image [3]. Instance segmentation extends the capabilities of semantic segmentation by detecting several items in an image and precisely delineating and isolating each specific object of interest, e.g., discriminating between different individuals in the image [4]. Initial efforts have been to combine the two segmentation tasks to comprehend scenes thoroughly [5–8]. Segmentation plays a vital role in a wide range of applications, including driverless vehicles, medical image analysis, augmented reality, and video surveillance, among others [9]. A multitude of image segmentation algorithms have been devised in the literature, ranging from basic techniques, e.g., histogram-based bundling, thresholding [10], k-means clustering [11], region-growing [12], and watersheds [13], to more sophisticated methods such as graph cuts [14], active contours [15], conditional and Markov random fields [16], and sparsity-based approaches [17,18]. Currently acknowledged as the next generation of image segmentation algorithms, Deep Learning (DL) models exhibit significant performance gains in recent years. These models (e.g., Universal Network (U-Net) [19], High-Resolution Network (HRNet) [20], Mask Region-based Convolutional Neural Network (Mask R-CNN) [21], Fully Convolutional Networks (FCN) [22], Segmented Neural

 doi:10.20944/preprints202408.1790.v1

Network (SegNet) [23]), Segment Anything Model (SAM) [24]) often achieve achieving the highest levels of accuracy on well-known standards, resulting in a significant shift in the sector.

Video segmentation, which involves identifying the key objects in a video scene based on their specific properties or semantics, is a fundamental and difficult problem in Computer Vision (CV). It possesses numerous potential uses, including robotics, autonomous driving, social media, automated surveillance, augmented reality, movie creation, and video conferencing [25]. The problem has been tackled using conventional CV and Machine Learning (ML) methods. These techniques include hand-crafted features, e.g., optical flow and histogram statistics, heuristic prior knowledge, e.g., motion boundaries [26] and visual attention mechanism [27], low/mid-level visual representations, e.g., super-voxel [28], trajectory [29], and object proposal [30], as well as classical ML models, e.g., graph models [31], clustering [32], support vector machines (SVM) [33], random decision forests [34], random walks [35], Markov random fields [36], and conditional random fields [37]. Recently, DL models, namely FCN [22], You Only Look Once (YOLO) v5,v7,v8 models [38–40], Mask R-CNN [21], and SAM [24] have significantly progressed in video segmentation. DL-based video segmentation algorithms exhibit significantly higher precision and, at times, greater efficiency than traditional approaches, e.g., SAM [24], Mask R-CNN [21], etc.

The uprising of foundation models has caused a substantial change in different fields, e.g., Natural Language Processing (NLP), CV, Reinforcement Learning (RL), etc. These models achieve spectacular outcomes because of their thorough pre-training on large datasets and exceptional ability to apply their knowledge to a wide range of specific tasks, e.g., Machine Translation, Image Segmentation, Autonomous Driving, Healthcare, etc. [41]. The Generative Pre-trained Transformer (GPT) [42] developed by OpenAI has achieved significant advancements in various language tasks within the field of NLP. It has also facilitated the development of successful commercial applications such as ChatGPT [43], renowned for its ability to generate coherent language in real-time and engage in meaningful interactions with users. Nevertheless, in the field of CV, researchers are still pursuing the construction of foundation models that are both powerful and adaptable. This pursuit is driven by the need to solve the distinct obstacles and complexities in the visual domain.

The creation of Contrastive Language-Image Pre-training (CLIP) [44], a model that successfully integrates image and text modalities, has shown the ability to generalize to new visual concepts without prior exposure. Despite this, the generalization capacity of vision tasks for AI models is still limited due to the scarcity of complete training data, especially when compared to NLP models. Last year, Meta AI Research unveiled the Segment Anything Model (SAM) [24], a highly adaptable and responsive model that can accurately segment any item in images or videos without requiring extra training. In CV, this approach is referred to as zero-shot transfer. SAM is a unique CV model trained using the SA-1B dataset [24], which includes more than 11 million images and one billion masks. This makes SAM the first foundation model of its kind. SAM is designed to provide precise segmentation outcomes by utilizing several cues, including points, boxes, or a combination. It has consistently demonstrated excellent generalization capabilities across various images and objects. Despite its long achievements, SAM had several limitations. The primary purpose of SAM was to do static picture segmentation, rendering it inefficient for video analysis [45,46]. The processing speed and efficiency of the system were not optimized for handling massive datasets, making it less appropriate for real-time applications [24]. In addition, SAM had a deficiency in its capability to retain memory across several frames, resulting in a restricted capacity to monitor the movement of objects over a period of time [45,46]. SAM's training on static picture datasets also limited its effectiveness and capacity to apply to various video scenarios [47], e.g., object tracking, action recognition, lighting variations, etc.

Under consideration of the limitations of the SAM model, Meta AI Research recently introduced a new model, Segment Anything Model 2 (SAM 2) [47]. It is designed for prompt object segmentation in both images and videos. It runs in real-time using a transformer-based architecture with streaming memory. SAM 2 builds on the success of the original SAM model, which was designed to facilitate flexible, prompt image segmentation. There is a need for a model that can handle images and

 doi:10.20944/preprints202408.1790.v1

videos seamlessly with the rapid growth of multimedia content and the increasing demand for video analysis. SAM 2 meets this need by incorporating advanced features like streaming memory and a robust training SA-V dataset. SAM 2 is introduced to address the limitations of the original SAM model, specifically in handling video data. The new model extends segmentation capabilities from static images to videos and accommodates the complex spatio-temporal dynamics involved. SAM 2 addresses several challenges in visual segmentation. It provides efficient real-time video segmentation and quick and accurate video frame processing. The model enhances accuracy while requiring fewer interactions. This makes it more user-friendly and effective in practical applications. Additionally, SAM 2 improves the ability to segment objects in dynamic and cluttered environments. This ensures robust performance even in complex visual scenes. The objective of this study is to provide an overview of SAM 2.

2. Literature Review

There are various segmentation techniques available and each of them has its own unique approach and applications. These techniques can be broadly categorized into three main types e.g. traditional segmentation techniques, ML-based segmentation techniques, and DL-Based techniques. Traditional techniques depend on basic image processing methods. ML-based techniques use statistical models and algorithms while DL-based techniques use neural networks for advanced segmentation tasks. In this section, we will describe various segmentation techniques across these three categories.

2.1. Traditional Segmentation Techniques

Global Thresholding: Global thresholding uses one threshold value for the whole image. Pixels above the threshold are marked as foreground, and those below are marked as background. This method is quick and easy but may struggle with uneven lighting or complex backgrounds. Global Thresholding can be used for simple image processing tasks like document scanning [48].

Adaptive Thresholding: Adaptive thresholding sets a threshold for smaller parts of the image and adjusts for different lighting conditions. This method calculates the threshold for each pixel based on its neighbours which makes it better for unevenly illuminated images. Though this method is slower than global thresholding but it is useful in fields like medical imaging and document processing [49].

Otsu's Method: Otsu's method automatically finds the best threshold by minimizing the variation within the foreground and background pixel intensities. It is particularly useful for bimodal images where the histogram has two distinct peaks. This method does not work well on images with overlapping intensity distributions. It is commonly used in automated industrial inspection and medical imaging [50].

Canny Edge Detector: The Canny edge detector is a multi-step process for detecting edges in images. It includes Gaussian smoothing which finds intensity gradients, non-maximum suppression, and edge tracking using hysteresis. This method is sensitive to noise and requires careful parameter tuning while widely used in edge detection and video processing [51].

Sobel Operator: The Sobel operator applies convolution masks to determine the gradient magnitude of image intensity. It emphasizes edges and transitions that provide a simple way to detect edges. This method is susceptible to noise and is mainly used in real-time edge detection for video surveillance [52].

Prewitt Operator: Similar to the Sobel operator, the Prewitt operator calculates image intensity gradients using convolution masks. It is used for edge detection but is less sensitive to noise compared to the Sobel operator. This method is used in basic image processing tasks [53].

Region Growing: Region growing begins with seed points and adds neighbouring pixels that match similarity criteria. It works well for images with distinct regions but can be sensitive to noise and initial seed selection. This method is mainly applied in medical imaging e.g. tumor detection [54].

 doi:10.20944/preprints202408.1790.v1

Region Splitting and Merging: This technique splits an image into regions and merges adjacent ones that meet similarity criteria. It is an iterative process that combines both splitting and merging steps. This method is computationally expensive and it is widely used in satellite image analysis and remote sensing [55].

K-means Clustering: K-means clustering divides the image into K groups based on pixel intensity. It iteratively assigns pixels to clusters and updates cluster centroids until convergence. This method requires pre-specifying the number of clusters and is used in colour quantization and image compression [56].

Watershed Algorithm: The watershed algorithm views a grayscale image as a topographic map and identifies the lines that divide regions. This method does over-segmentation in noisy images but is effective in separating overlapping objects [57].

Normalized Cuts: Normalized cuts segment the image by minimizing a criteria that balances the edge weights between and within segments. This method is computationally expensive and requires graph construction but is useful for object recognition [58].

2.2. Machine Learning-Based Segmentation Techniques

K-Nearest Neighbors (KNN): KNN classifies each pixel by the majority vote of its closest neighbours in the feature space. It is simple but can be computationally expensive for large datasets and it is used in small-scale image segmentation tasks [59].

Support Vector Machines (SVM): SVM finds the best hyperplane to separate pixel classes in the feature space which maximizes the margin between classes. This method requires a well-defined feature space and is applied in texture segmentation. [60].

Random Forests: Random forests use multiple decision trees to classify pixels based on features which provides robust and accurate segmentation. This method is prone to overfitting on noisy datasets and is used in multi-spectral image segmentation [61].

Logistic Regression: Logistic regression models the probability of a pixel belonging to a class using a logistic function. It is a linear model but can be extended to multi-class segmentation. This method is limited to linear decision boundaries and can be applied in binary segmentation tasks like object detection[62].

Bayesian Networks: Bayesian networks use probabilistic models to predict the likelihood of a pixel belonging to a segment which incorporates prior knowledge and observed data. This method requires prior knowledge and is used in probabilistic image segmentation in medical imaging [63].

Conditional Random Fields (CRF): CRFs model the conditional probability of the pixel labels given the image that considers the dependencies between neighbouring pixels for improved segmentation accuracy. This method is difficult to optimize and effective in refining segmentation maps [64].

2.3. Deep Learning-Based Techniques

Fully Convolutional Networks (FCNs): FCNs convert traditional CNNs into fully convolutional networks for pixel-wise prediction that allows for end-to-end training and segmentation. This method struggles with fine details and is widely used in real-time semantic segmentation tasks [65].

U-Net: U-Net has an encoder-decoder structure with skip connections that is designed for biomedical image segmentation. It is known for its ability to work with very few training images. This method requires large amounts of labelled data and is used in biomedical image segmentation [66].

SegNet: SegNet uses an encoder-decoder architecture for pixel-wise classification that transfers indices from the encoder to the decoder to improve segmentation accuracy and reduce computation. This method may lose spatial details and is applied in road scene segmentation for autonomous driving [67].

 doi:10.20944/preprints202408.1790.v1

Mask R-CNN: Mask R-CNN builds on Faster R-CNN by adding a branch to predict segmentation masks for each Region of Interest (RoI) that allows for both object detection and segmentation at the same time. This method requires a large dataset and is Used in instance segmentation [68].

DeepLab: DeepLab is a semantic segmentation technique that uses atrous convolutions to capture multi-scale context while preserving resolution. It also incorporates fully connected Conditional Random Fields (CRFs) to improve object edges and small details. The limitation of this method is slow inference time but it is effective in semantic segmentation for autonomous vehicles [69].

Despite the advancements in existing segmentation models, there are several limitations such as difficulty handling complex video data, high computational demands, and a lack of flexibility in real-time applications. These models often can not maintain accuracy across diverse and dynamic environments which leads to poor performance in practical scenarios. The development of SAM2 addresses these challenges by introducing a more robust architecture capable of effective real-time segmentation in both images and videos. This makes SAM2 optimal for a wide range of applications across various domains.

3. Architecture Comparison of SAM and SAM 2

The primary goal of this chapter is to provide a comprehensive comparison of the SAM and SAM 2 architectures. By examining the architectural differences between these two models, this chapter aims to shed light on the specific advancements and innovations that SAM 2 introduces over its predecessor, SAM.

3.1. Architecture of SAM

Inspired by techniques from NLP, the SAM is a foundational model for segmentation tasks [70]. Promptable segmentation is its main objective. To indicate the objects to segment in an image, different input forms are used in this context, including masks, bounding boxes, foreground/background points, and free-form text [70]. Even in undetermined situations, e.g., overlapping objects, partial occlusion, small objects, unfamiliar objects etc., SAM attempts to produce trustworthy segmentation masks in response to any given prompt. Unlike interactive segmentation models that are primarily intended for human interaction, this model's versatility enables its application to a broad range of scenarios, e.g., medical imaging, autonomous driving, aerial and satellite imagery, robotics navigation, underwater exploration, traffic management etc. [70].

SAM was developed to address the growing need for a robust segmentation model capable of handling diverse and complex datasets. Inspired by advances in NLP and pre-trained Vision Transformer (ViT), SAM was designed to influence the strengths of these technologies to achieve high performance in segmentation tasks. Key milestones in its development include the integration of promptable segmentation, the use of ViT as an image encoder, and the creation of the extensive SA-1B dataset containing over 1.1 billion masks. SAM's contributions to the field of CV include its innovative approach to segmentation and its ability to generalize across various prompts and applications.

3.1.1. The Segment Anything Task

Specifically, the next token prediction task [70] serves as a source of inspiration for SAM's methodology. Using the context that the words before it have provided, the task is to predict the word that will come after it in a sequence. Similarly, SAM uses prompts to guide the segmentation process, predicting the appropriate segmentation masks based on the input prompts [70]. The concept of promptable segmentation is central to SAM. A prompt can take various forms, e.g., foreground/background points, bounding boxes, masks, or free-form text, and indicates what objects to segment in an image [70]. This flexibility allows SAM to be used in a wide range of applications and makes it adaptable to different types of input [71]. The primary objective of SAM is to generate valid segmentation masks based on any given prompt, even in ambiguous scenarios. This capability enables SAM to perform zero-shot transfer to downstream segmentation tasks, as it can adapt to new prompts and datasets without

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 August 2024 doi:10.20944/preprints202408.1790.v1

requiring additional training [70]. SAM employs a natural pre-training algorithm that influnces the concept of promptable segmentation. This approach facilitates zero-shot transfer, allowing SAM to perform well on downstream segmentation tasks by engineering appropriate prompts. The model's ability to generalize from pre-training to new tasks makes it highly versatile and efficient [70,71].

3.1.2. Network Design of SAM

Three key elements are central to SAM's design, as Figure 1 illustrates.

- The task of promptable segmentation to facilitate zero-shot generalization [70].
- This model architecture consists of a mask decoder, prompt encoder, and image encoder [70].
- The dataset that powers the task and model, specifically the SA-1B dataset containing over 1.1 billion masks [70].

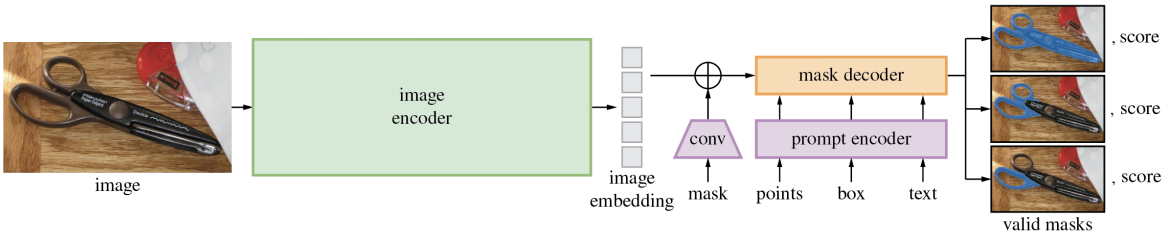

Figure 1. Overview of the architecture of SAM

Image Encoder

The image encoder is responsible for generating one-time image embeddings, which are important for the segmentation process. These embeddings capture the essential features of the image, providing the necessary information for the prompt encoder and mask decoder [70]. SAM utilizes a Masked Auto-Encoder (MAE) pre-trained ViT as its image encoder. The ViT processes high-resolution inputs efficiently, creating robust embeddings that are used in subsequent stages of the segmentation process. This pre-trained ViT allows SAM to impact powerful pretraining methods, enhancing its performance and scalability [71]. The image encoder is designed to process high-resolution inputs efficiently. By running once per image, it creates embeddings that can be seamlessly integrated into the segmentation process. This efficiency is important for real-time performance and scalability [70].

Prompt Encoder

In real-time, the prompt encoder is in charge of encoding various prompt types, e.g., text, masks, boxes, and points, into an embedding vector [70]. Because of this adaptability, SAM can respond to a greater variety of input prompts. Dense prompts are summed element-wise with the image embedding and embedded using convolutions, same as masks. By using these methods, SAM is guaranteed to be able to interpret different kinds of prompts efficiently [70]. An off-the-shelf text encoder from CLIP is used to encode free-form text prompts. This integration allows SAM to handle textual input effectively, further expanding its range of applications, e.g., content-based image retrieval, image captioning, visual question answering, sentiment analysis in images, medical report generation, remote sensing analysis etc. [70].

Mask Decoder

Predicting segmentation masks using the image and prompt encoder embeddings is the main function of the mask decoder. In order to produce precise segmentation masks, it maps these embeddings effectively [70]. The mask decoder's design draws inspiration from pre-existing Transformer decoder blocks. To efficiently update all embeddings, it integrates cross-attention and prompt self-attention mechanisms. According to the input prompts and image embeddings, this design guarantees that the mask decoder can predict segmentation masks with accuracy [70]. Real-time performance

 doi:10.20944/preprints202408.1790.v1

optimization of the mask decoder allows for smooth interactive prompting of the model [71]. Because of this optimization, SAM can complete segmentation tasks quickly and effectively, which makes it appropriate for real-time applications.

3.2. *Architecture of SAM 2*

Since the SAM-2 was introduced, Meta has expanded the parameters of AI [72]. The remarkable powers of its predecessor, SAM, are further enhanced by this ground-breaking development in CV. SAM-2 accurately identifies and segments objects, revolutionizing real-time image and video segmentation [72]. By raising the bar for what is possible in CV, this advancement in visual understanding creates new opportunities for AI applications in a variety of industries, e.g., healthcare, automotive, agriculture, finance, entertainment, logistics, real estate, manufacturing, aerospace, sports, tourism, environmental science, energy, security etc.

SAM-2 expands the use of Meta AI's models beyond static image segmentation to dynamic video tasks with new features and improved performance. SAM 2 [72] was developed in response to the following requirements: introduce memory components; unify the architecture for image and video tasks; support video segmentation; and improve efficiency and occlusion handling. SAM-2 performs better than benchmarks thanks to its multiple mask predictions, occlusion prediction, user-guided refinement, zero-shot segmentation for novel objects, and real-time video segmentation. Future studies are necessary because even with these improvements, SAM-2 still has issues with long-term memory tracking, object disambiguation, fine detail preservation, and temporal consistency [72].

3.2.1. Network Design of SAM 2

Enhanced Encoder

For effective, real-time video frame processing, SAM 2 makes use of a pre-trained Hiera model [72]. As shown in Figure 2, the improvements include a memory encoder, memory bank, and memory attention module that store and utilize object information to facilitate user interaction throughout the video.

Within a single architecture, SAM 2 can now handle tasks related to image and video segmentation due to new techniques and improvements. When the image or video is uncertain, SAM 2 provides multiple potential masks and processes video frames one at a time, enabling real-time segmentation of long videos.

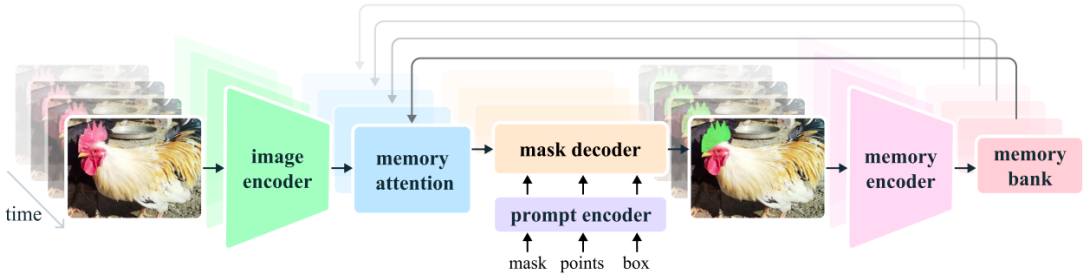

Figure 2. Overview of the architecture of SAM 2

Advanced Decoder

A new head to detect object presence in frames has been added to the decoder, which has been modified for the video context [72]. It can now predict multiple masks for ambiguous prompts. Segmentation masks are produced by the mask decoder, which effectively maps the image and prompts embeddings. To effectively update all embeddings and ensure accurate segmentation masks based on input prompts and image embeddings, it integrates prompt self-attention and cross-attention mechanisms [72].

 doi:10.20944/preprints202408.1790.v1

Enhanced Feature Extraction

Using a streaming method, SAM 2 processes video frames one after the other, enabling real-time segmentation of videos of any length [72]. The model's ability to handle dynamic scenes is improved by temporal conditioning, which integrates information from previous frames and prompts. The model's adaptability is increased by the memory attention mechanism, which uses transformer blocks with self-attention and cross-attention mechanisms to condition current frame features on previous frame information and new prompts [72].

3.2.2. Enhancements in SAM 2 Architecture

SAM 2 adds a number of noteworthy enhancements while still building on the framework established by SAM. SAM 2 can segment objects in videos, but SAM could only segment images. Therefore, SAM is image-specific, while SAM 2 is task-specific and uses a single model for both images and videos. A feature missing from SAM, object tracking across video frames is now possible with SAM 2. SAM 2 can predict object visibility, that SAM [72] does not have. This is made possible by its occlusion head. In image segmentation tasks, SAM 2 is 6 times faster than SAM [72]. Even in image segmentation, it performs better than the original SAM on a number of benchmarks [72].

References

1. Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.
2. Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021.
3. Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020.
4. Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.
5. Stephen Gould, Tianshi Gao, and Daphne Koller. Region-based segmentation and object detection. *Advances in neural information processing systems*, 22, 2009.
6. Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019.
7. Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63:113–140, 2005.
8. Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 702–709. IEEE, 2012.
9. David A Forsyth and Jean Ponce. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002.
10. Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.
11. Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015.
12. Richard Nock and Frank Nielsen. Statistical region merging. *IEEE Transactions on pattern analysis and machine intelligence*, 26(11):1452–1458, 2004.
13. Laurent Najman and Michel Schmitt. Watershed of a continuous function. *Signal processing*, 38(1):99–112, 1994.
14. Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001.
15. Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

 doi:10.20944/preprints202408.1790.v1

16. Nils Plath, Marc Toussaint, and Shinichi Nakajima. Multi-class image segmentation using conditional random fields and global classification. In *Proceedings of the 26th annual international conference on machine learning*, pages 817–824, 2009.
17. J-L Starck, Michael Elad, and David L Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE transactions on image processing*, 14(10):1570–1582, 2005.
18. Shervin Minaee and Yao Wang. An admm approach to masked signal decomposition using subspace representation. *IEEE Transactions on Image Processing*, 28(7):3192–3204, 2019.
19. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
20. Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020.
21. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
22. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
23. Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.
24. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
25. Tianfei Zhou, Fatih Porikli, David J Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7099–7122, 2022.
26. Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE international conference on computer vision*, pages 1777–1784, 2013.
27. Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3395–3402, 2015.
28. Chenliang Xu and Jason J Corso. Evaluation of super-voxel methods for early video processing. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1202–1209. IEEE, 2012.
29. Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *European conference on computer vision*, pages 282–295. Springer, 2010.
30. Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *2011 International conference on computer vision*, pages 1995–2002. IEEE, 2011.
31. Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa. Efficient hierarchical graph-based video segmentation. In *2010 ieee computer society conference on computer vision and pattern recognition*, pages 2141–2148. IEEE, 2010.
32. Chen-Ping Yu, Hieu Le, Gregory Zelinsky, and Dimitris Samaras. Efficient video segmentation using parametric graph partitioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3155–3163, 2015.
33. Federico Perazzi, Oliver Wang, Markus Gross, and Alexander Sorkine-Hornung. Fully connected object proposals for video segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 3227–3234, 2015.
34. Vijay Badrinarayanan, Ignas Budvytis, and Roberto Cipolla. Semi-supervised video segmentation using tree structured graphical models. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2751–2764, 2013.
35. Naveen Shankar Nagaraja, Frank R Schmidt, and Thomas Brox. Video segmentation with just a few strokes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3235–3243, 2015.

 doi:10.20944/preprints202408.1790.v1

36. Won-Dong Jang and Chang-Su Kim. Streaming video segmentation via short-term hierarchical segmentation and frame-by-frame markov random field optimization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 599–615. Springer, 2016.
37. Buyu Liu and Xuming He. Multiclass semantic video segmentation with object-level active inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4286–4294, 2015.
38. Yu Zhang, Zhongyin Guo, Jianqing Wu, Yuan Tian, Haotian Tang, and Xinming Guo. Real-time vehicle detection based on improved yolo v5. *Sustainability*, 14(19):12274, 2022.
39. Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.
40. Gang Wang, Yanfei Chen, Pei An, Hanyu Hong, Jinghu Hu, and Tiange Huang. Uav-yolov8: A small-object-detection model based on improved yolov8 for uav aerial photography scenarios. *Sensors*, 23(16):7190, 2023.
41. Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
42. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
43. Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
44. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
45. Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022.
46. Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22290–22300, 2023.
47. Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
48. Mohamed Abdel-Basset, Victor Chang, and Reda Mohamed. A novel equilibrium optimization algorithm for multi-thresholding image segmentation problems. *Neural Computing and Applications*, 33:10685–10718, 2021.
49. Juan Liao, Yao Wang, Dequan Zhu, Yu Zou, Shun Zhang, and Huiyu Zhou. Automatic segmentation of crop/background based on luminance partition correction and adaptive threshold. *IEEE Access*, 8:202611–202622, 2020.
50. Jonathan T Barron. A generalization of otsu's method and minimum error thresholding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 455–470. Springer, 2020.
51. Yong Woon Kim and Addapalli VN Krishna. A study on the effect of canny edge detection on downscaled images. *Pattern Recognition and Image Analysis*, 30:372–381, 2020.
52. Fei Hao, Dashuai Xu, Delin Chen, Yuntao Hu, and Chaohan Zhu. Sobel operator enhancement based on eight-directional convolution and entropy. *International Journal of Information Technology*, 13(5):1823–1828, 2021.
53. Saeed Balochian and Hossein Baloochian. Edge detection on noisy images using prewitt operator and fractional order differentiation. *Multimedia Tools and Applications*, 81(7):9759–9770, 2022.
54. J Dafni Rose, K VijayaKumar, Laxman Singh, and Sudhir Kumar Sharma. Computer-aided diagnosis for breast cancer detection and classification using optimal region growing segmentation with mobilenet model. *Concurrent Engineering*, 30(2):181–189, 2022.

 doi:10.20944/preprints202408.1790.v1

55. Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 663–679. Springer, 2020.
56. Sandra Jardim, João António, and Carlos Mora. Graphical image region extraction with k-means clustering and watershed. *Journal of Imaging*, 8(6):163, 2022.
57. Yongan Xue, Jinling Zhao, and Mingmei Zhang. A watershed-segmentation-based improved algorithm for extracting cultivated land boundaries. *Remote Sensing*, 13(5):939, 2021.
58. Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE transactions on pattern analysis and machine intelligence*, 2023.
59. Malti Bansal, Apoorva Goyal, and Apoorva Choudhary. A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal*, 3:100071, 2022.
60. Derek A Pisner and David M Schnyer. Support vector machine. In *Machine learning*, pages 101–121. Elsevier, 2020.
61. Robin Genuer, Jean-Michel Poggi, Robin Genuer, and Jean-Michel Poggi. *Random forests*. Springer, 2020.
62. Patrick Schober and Thomas R Vetter. Logistic regression in medical research. *Anesthesia & Analgesia*, 132(2):365–366, 2021.
63. Neville Kenneth Kitson, Anthony C Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. A survey of bayesian network structure learning. *Artificial Intelligence Review*, 56(8):8721–8814, 2023.
64. Bengong Yu and Zhaodi Fan. A comprehensive review of conditional random fields: variants, hybrids and applications. *Artificial Intelligence Review*, 53(6):4289–4333, 2020.
65. Maria Baldeon Calisto and Susana K Lai-Yuen. Adaen-net: An ensemble of adaptive 2d–3d fully convolutional networks for medical image segmentation. *Neural Networks*, 126:76–94, 2020.
66. Nahian Siddique, Sidike Paheding, Colin P Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE access*, 9:82031–82057, 2021.
67. Karuna Kumari Eerapu, Shyam Lal, and AV Narasimhadhan. O-segnet: Robust encoder and decoder architecture for objects segmentation from aerial imagery data. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(3):556–567, 2021.
68. Xiuli Bi, Jinwu Hu, Bin Xiao, Weisheng Li, and Xinbo Gao. Iemask r-cnn: Information-enhanced mask r-cnn. *IEEE Transactions on Big Data*, 9(2):688–700, 2022.
69. Fangfang Liu and Ming Fang. Semantic segmentation of underwater images based on improved deeplab. *Journal of Marine Science and Engineering*, 8(3):188, 2020.
70. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
71. Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024.
72. Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.