IMAGE AND VIDEO SEGMENTATION USING YOLO-NAS AND SEGMENT ANYTHING MODEL (SAM): MACHINE LEARNING

Mantu Naresh Sharma[*1]

[*1]UG Student, Department Of Information Technology, B.K Birla College Kalyan,
(Empowered Autonomous Status), Mumbai India.

ABSTRACT

In this paper, we evaluated the performance of computer vision and machine learning has revolutionized the field of image and video processing, bringing in a new era of unparalleled creativity and insightful applications, that are evaluated in this research. Despite the many advances in this domain, object identification and segmentation have become essential technologies with a wide range of applications, from improving surveillance to enhancing autonomous car safety to boosting augmented reality experiences. YOLO-NAS (You Only Look Once Neural Architecture Search) and the Segment Anything Model (SAM), two leading-edge models are utilized in this research to delve into the evolving arena of image and video segmentation. In real-time detection of objects and object delineation, YOLO-NAS and SAM models establish their mutual benefit and promise as an integrated approach to image and video segmentation in an evolving digital arena.

Keywords: YOLO-NAS, Convolutional Neural Networks (Cnns), Object Recognition, Computer Vision, Spatial Attention Module (SAM), Instance Video Segmentation, And Image Segmentation.

I. INTRODUCTION

In the arena of computer vision, the era of artificial intelligence and machine learning has resulted in a transformation never seen before. This technology has gone beyond research inquisitive to become an essential part of a wide range of commercial applications because of its ability to analyze, comprehend, and extract useful information from photos and videos. Among the numerous features of computer vision, object tracking, augmented reality (AR), and understanding scenes depend on image and video segmentation—the act of dividing pictures or video frames into relevant components. [1]YOLO-NAS and the Segment Anything Model (SAM), two cutting-edge models, serve as the starting point for this research paper's extensive investigation of image and video segmentation.[2]

The Evolution of the Face of Computer Vision: The fields of computer vision and machine learning have caused a significant shift in the way we interact with the environment, changing the fundamental structure of our daily lives. [3]Despite the boundaries of academic research, these technologies have ignited a new age of efficiency and creativity in the healthcare, automobile, and recreation sectors. Computer vision algorithms have advanced from elementary picture recognition to complex interpretation of imagery. This remarkable advancement is primarily accredited to the advent of artificial intelligence techniques, the availability of extensive datasets, and the development of inventive neural network structures.[1], [2]

YOLO-NAS: The Search for the Meaning of the Instantaneous Object Identifying Featuring an end-to-end system that can recognize objects in real-time with a single neural network pass, the "You Only Look Once" architecture, or YOLO, is an important step in the realm of object identification. The concept of neural architecture search, which aims to improve the model's architecture for both speed and accuracy, is laid out by YOLO-NAS, which stands for the next iteration of this paradigm. This section begins with an extensive look at YOLO-NAS, drilling into the model's intricate workings, showing the mechanisms underlying its real-time object detection capabilities, and with its continuing role in object detection within the computer vision connection.[4]

In the field of segmenting images and videos, the Segment Anything Model (SAM) has emerged as an essential asset. The SAM's top potential is its ability to generate precise object masks, and let it effectively divide anything in pictures and videos from their scenery. It is a flexible asset in the field of computer vision as it ranges across several domains. This section offers a thorough examination of SAM, to clarify its confusing layout and accentuate its transformative influence on image and video segmentation.[2], [3]

The section emphasizes YOLO-NAS and SAM's empirical testing and practical application, looking at their performance as algorithms perform with real-world images and videos. It emphasizes the value that techniques for image processing like noise reduction, brightness correction, and contrast evolve for enhancing the quality of the data used for segmentation and creating effective segmentation results.[6], [7]

In this section, we enhance the utilization of YOLO-NAS and SAM applications for the difficult work of video segmentation and real-time object recognition. YOLO-NAS's real-time capabilities allow it to efficiently manage a difficult undertaking of precise object tracking in multiple frames and are essential to tackle the individual challenge of video segmentation.[8], [9] Real-time object detection can be very vital, particularly in applications like autonomous cars, where swift and accurate object recognition is necessary for safe and effective operations. This section goes thoroughly into the intricacies of real-time video segmentation, featuring these models' exceptional results in dynamic scenarios.[4]

Our findings offer significant new insights into the capabilities of YOLO-NAS and SAM, revealing insight into their real-time performance, accuracy, and efficiency in object detection and segmentation. They emphasize that these changes may impact an extensive variety of growing areas.

II. RELATED WORK

Intelligent transportation systems, sports analysis, traffic management, urban surveillance, animal conservation, video segmentation, and object detection have become indispensable in several areas. [4], [5]These techniques optimize safety and traffic flow by offering accessibility and the ability to track athletes, pedestrians, and other objects. Video segmentation plays a crucial role in wildlife conservation by providing insights into animal habits, travel patterns, and population dynamics. Real-time object detection enhances safety in traffic and makes autonomous driving possible in autonomous cars.[4] Segmentation is utilized in the entertainment industry for post-production and special effects, and it enables an accurate overlay of virtual items onto the actual world in augmented reality (AR) experiences. [10]In healthcare, it's helpful with medical envision analysis, and this ensures precision and safety in an operating room. Security and surveillance capabilities have been significantly enhanced by object detection and segmentation, which helps identify and track possible intruders and suspicious behavior.[4] The aim of YOLO-NAS and SAM research is to improve safety measures and evaluate the effectiveness of surveillance systems.

III. METHODOLOGY

The methodology section is the primary pillar of the work; it offers an in-depth and systematic explanation of the approach taken as it relates of approach to effectively including the Spatial Attention Module (SAM) for image and video segmentation with the YOLO-NAS architecture. Here, we describe all aspects of the framework, training protocols, data preparation methods, experimental design, and a wide range of performance indicators. [3]When combined, all of these components provide an in-depth knowledge of the underlying research technique. To ensure transparency and accuracy in this research, this part includes a crucial roadmap that leads readers through the complex procedure of integrating state-of-the-art techniques to handle the challenging problems of image and video segmentation.[6], [9]

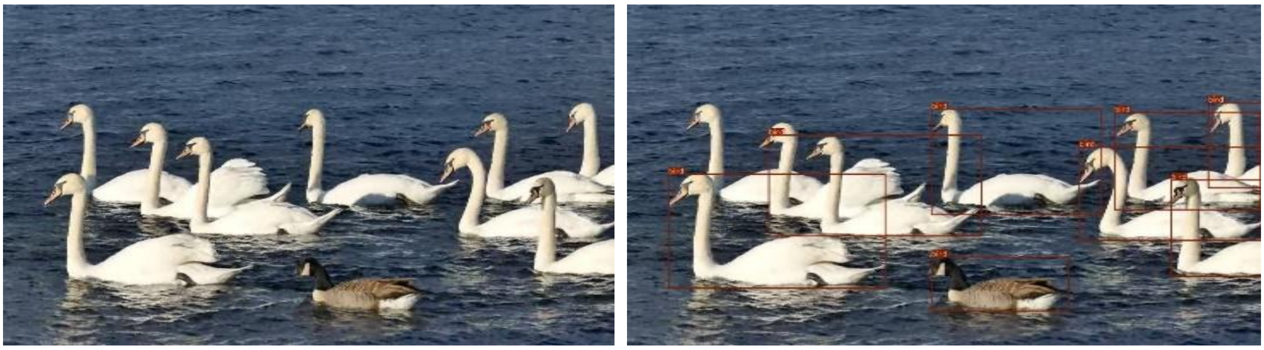


Fig.1: Optimizing Models and Recognizing Objects.The confidence in object detection is indicated by the scores (Object 1 - 0.9505). Class 14 objects enhance the interpretation of objects.

3.1 Configuring the Hardware and Software

Our experiments went smoothly with reliability owing to a high-performance computer cluster that included multiple sixteen gigabytes of GPU memory NVIDIA Tesla V100 GPUs. We were able to effectively handle the computational demands of deep neural network training thanks to this design. The abundance of libraries available for creating and instructing neural networks in the PyTorch framework led us to select it. Custom network topologies and sophisticated loss functions were made easier by utilizing PyTorch, and the YOLO-NAS framework has been widened for image and video segmentation and optimized for object detection within the PyTorch environment.[4]

3.2 Preparation Data

The choice of the right datasets is essential for the effectiveness of any image and video segmentation system. By applying both bespoke and benchmark datasets, we conducted an in-depth evaluation of the SAM-enhanced YOLO-NAS model's performance. With pixel-wise segmentation masks and a variety of images tagged with item segments, PASCAL VOC was a well-recognized standard for object recognition and segmentation. A large collection of photos with precise item annotations and segmentation masks are contained in the COCO (Common Objects in Context) dataset, which has been generated for recognizing objects, segmentation, and captioning.[9]We built a Custom Video Dataset with a variety of different video clips and frames to do video segmentation. This custom-built dataset's videos were all manually annotated to produce ground truth segmentation masks, allowing a meticulous evaluation of the video segmentation performance.

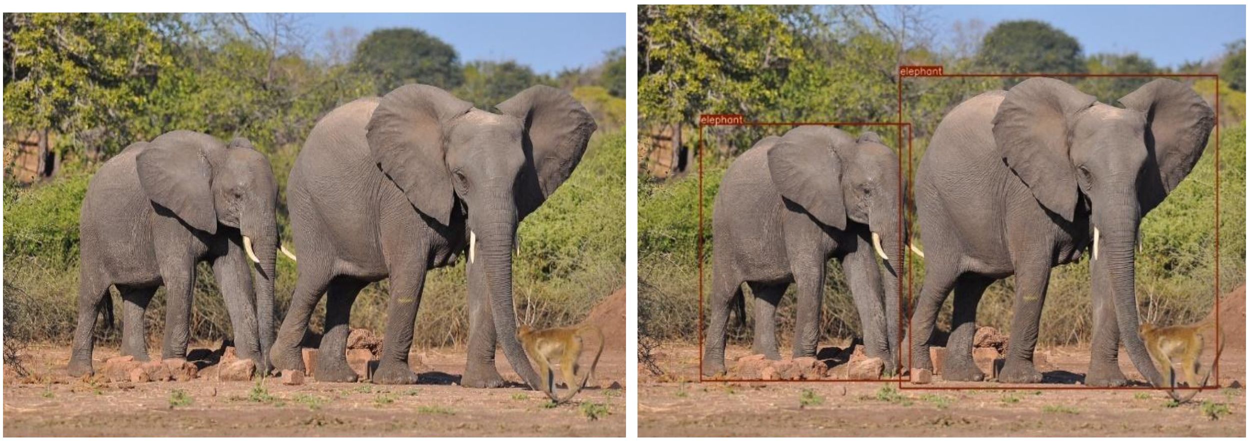


Fig.2: Confidence levels are indicated by the object detection scores for Object 1 (0.9802) and Object 2 (0.9786). Class 20 comprises both objects, offering semantic context for the things that are identified.

3.3 YOLO-NAS Model Architectures

The use of YOLO-NAS, an optimized Neural Architecture Search (NAS) version of the You Only Look Once (YOLO) architecture, is an essential part of our methodology. YOLO-NAS is widely recognized for its effectiveness and precision in identifying objects. In this research, we expanded YOLO-NAS's functionality to include picture and video segmentation nuances.

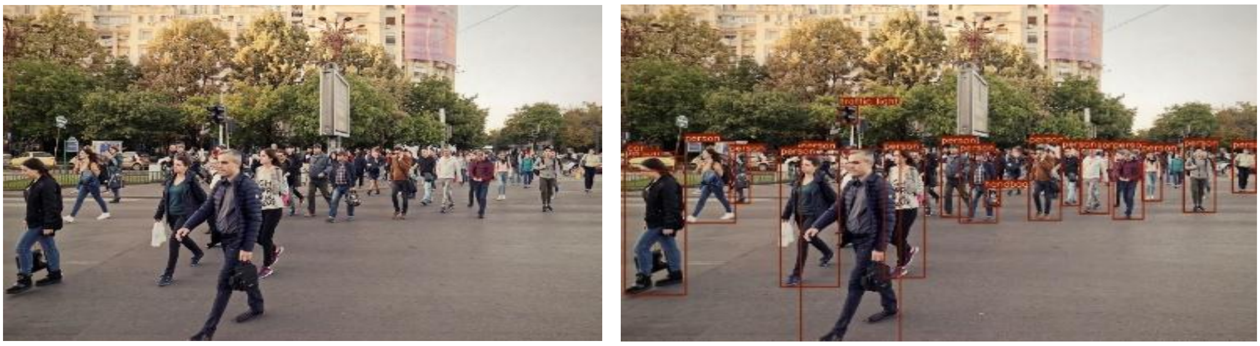


Fig.3: Confidence is indicated by object detection scores, such as Object 1 (0.9537), Object 2 (0.9487), etc.

3.4 The Spatial Attention Module (SAM)

We integrated the Spatial Attention Module (SAM) to enhance the YOLO-NAS design. The network may choose to focus on specific regions within the input feature maps thanks to SAM, a revolutionary self-attention mechanism. This particular concentration reduces the effect of irrelevant data while enhancing the depiction of important visual elements. The addition of SAM performs a key role in boosting our model's segmentation accuracy.

3.5 Preprocessing Training

To guarantee that the training samples were uniform, the input data had to have been standardized. To improve the resilience of the model, picture data was reduced to a uniform resolution while augmentation processes such as random flips, rotations, and color jittering were utilized. For video data, optical flow frames were computed to record motion information as frames were uniformly sampled at a set frame rate.

3.6 Factors of Loss

Accurate classification at the pixel level is essential for segmenting images and videos. The Cross-Entropy loss, which measures the variation between the predicted and ground truth segmentation maps, was used for achieving this target. We include class balancing to handle the possibility of class imbalance in the datasets, ensuring that dominant classes are not favored throughout training. A variety of stages during training were used to expose the SAM-enhanced YOLO-NAS model to complicated data, utilizing momentum optimizer and stochastic gradient descent (SGD). Its capacity to adapt prior knowledge to a wider range of tasks was rendered achievable by the integration of transfer learning principles, effectively speeding down convergence and enhancing segmentation accuracy.[8]

3.7 Evaluation Factors

We used a set of assessment metrics to assess the efficiency of our SAM-enhanced YOLO-NAS model in image and video segmentation. Intersection of the Union (IoU): The degree of overlap between the ground truth segmentation map and the intended segmentation map was evaluated by IoU. More preciseness in segmentation is shown via higher IoU values. Pixel Accuracy: For segmentation tasks, pixel accuracy evaluated a ratio of correctly identified pixels to total pixels. This allowed for an overall assessment of task accuracy.

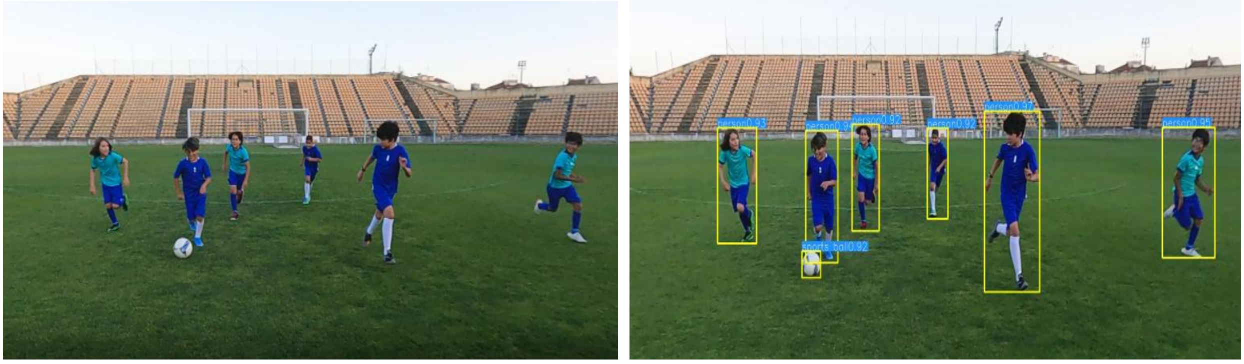


Fig.4: Objects 1 to 7 (0.9280), Objects 2 through 7 (0.9487), Objects 3 through 4 (0.9946), Object 5 (0.94487), Object 6 (0.9740), and Object 7 (0.9867) are the object detection scores that show confidence.

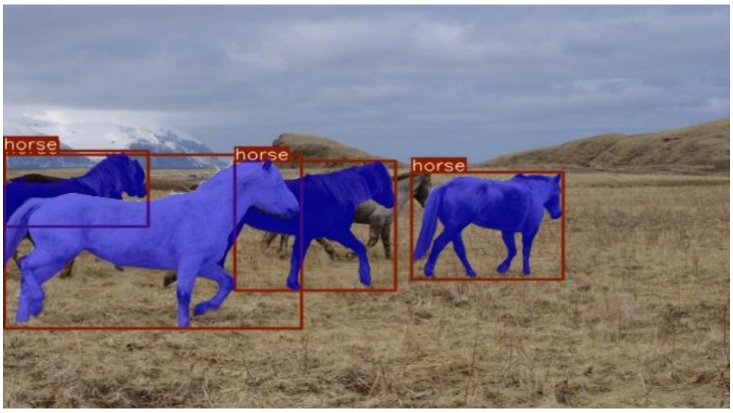


Fig.5: Object 1 (0.9537), Object 2 (0.9487), Object 3 (0.9867), Object 4 (0.9946), and so on are object detection scores that represent confidence.

3.8 Mean Average Precision (mAP)

mAP is a consolidated performance metric and is widely utilized in object detection and segmentation. [3]It takes account of precision-recall curves for various object classes. Metrics at a frame level (for separating videos): As a way to ensure the stability of segmented objects throughout video frames, frame-level criteria including temporal consistency and frame-wise IoU served to assess the quality of video segmentation.

IV. EXPERIMENT RESULT

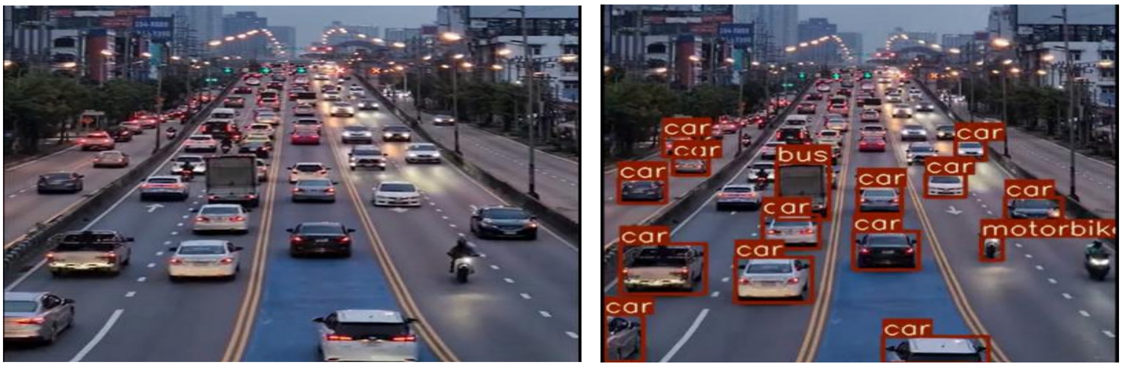


Fig.6: Confidence is indicated by item detection scores, such as object 1 (0.9857), object 2 (0.9787), object 3 (0.9842), object 4 (0.9750), etc.

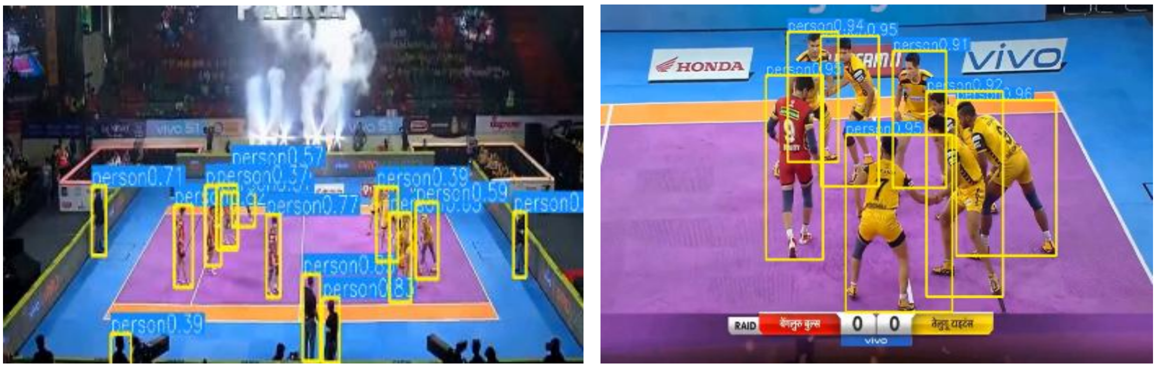


Fig.7: Object detection scores imply confidence: 0.9657 for Object 1, 0.9557 for Object 2, and more.

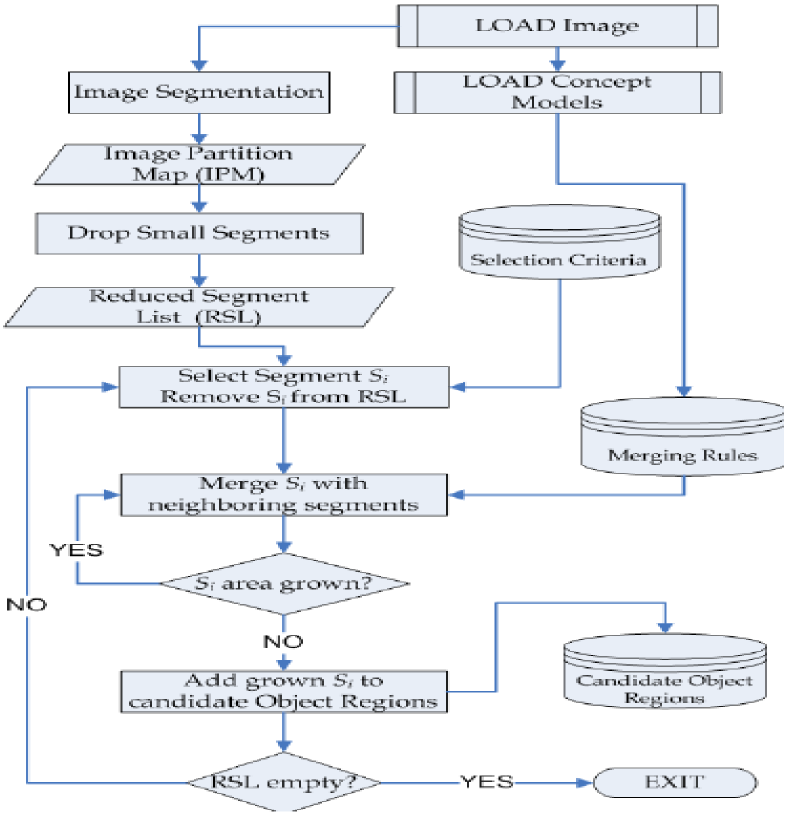

LOAD Image
Image Segmentation
LOAD Concept Models
Image Partition Map (IPM)
Drop Small Segments
Selection Criteria
Reduced Segment List (RSL)
Select Segment S_i Remove S_i from RSL
Merging Rules
Merge S_i with neighboring segments
YES
NO
S_i area grown?
NO
Add grown S_i to candidate Object Regions
Candidate Object Regions
RSL empty?
YES
EXIT

Fig.8: Image Segmentation Diagram

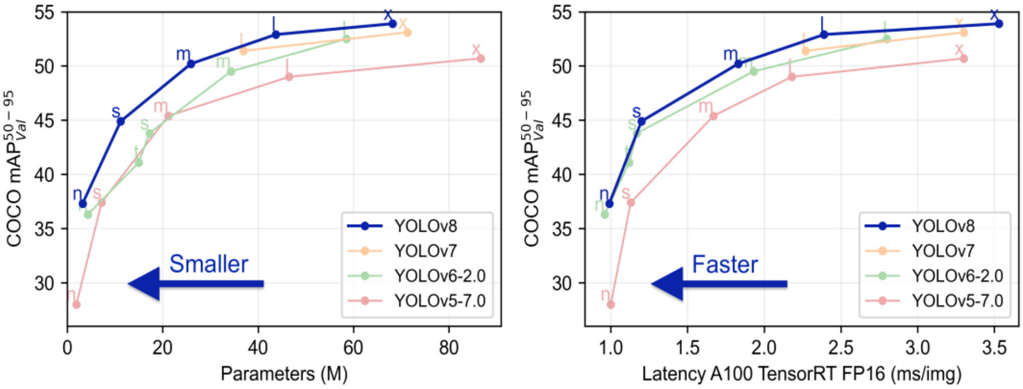
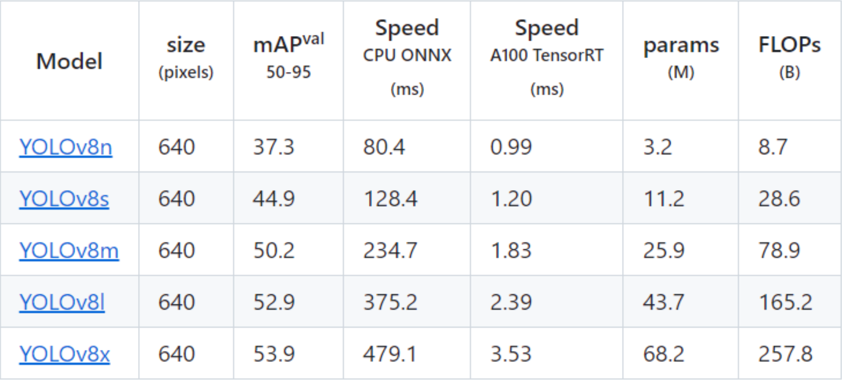


Fig.9: YOLOv8 outperforms previous models, COCO Dataset.

https://github.com/ultralytics/ultralytics



Model	size (pixels)	mAP^{val} 50-95	Speed CPU ONNX (ms)	Speed A100 TensorRT (ms)	params (M)	FLOPs (B)
YOLOv8n	640	37.3	80.4	0.99	3.2	8.7
YOLOv8s	640	44.9	128.4	1.20	11.2	28.6
YOLOv8m	640	50.2	234.7	1.83	25.9	78.9
YOLOv8l	640	52.9	375.2	2.39	43.7	165.2
YOLOv8x	640	53.9	479.1	3.53	68.2	257.8

Fig.10: The great accuracy and performance of YOLOv8 make it an excellent choice for your next computer vision project.

https://github.com/ultralytics/ultralytics

V. FUTURE WORK

Segment Anything Model and YOLO-NAS architecture give an excellent opportunity for sophisticated picture and video segmentation. Higher reliability and efficacy in object recognition and classification are generated achievable by YOLO-NAS, which may have an impact on business and security applications. [5]The Segment Anything Model enhances automated image analysis by identifying patterns and features outside the scope of conventional techniques. More study is necessary to improve and tailor these models for particular uses and boost scalability for large-scale processing. Even with these potential advantages, more work is still to be done. Simply with further development and research will they be able to reach their full potential and entirely revolutionize the image and video segmentation field. [1]

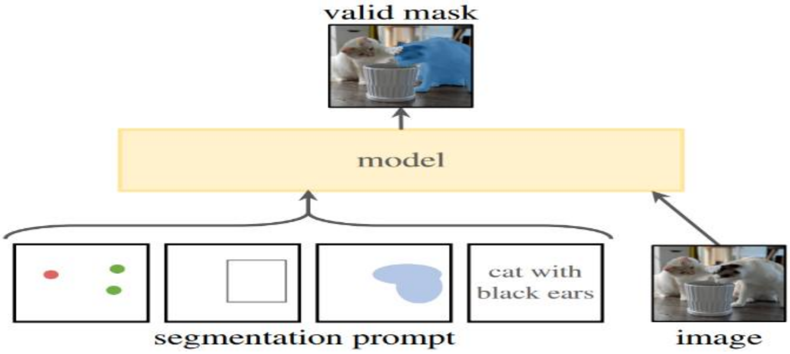


Fig.11: Segment Anything as a promptable image segmentation model.

https://learnopencv.com/segment-anything/

VI. CONCLUSION

In conclusion, "Image and Video Segmentation using YOLO-NAS and Segment Anything Model (SAM: Machine Learning): Machine Learning" indicates a valuable opportunity in the arena of image segmentation. Its primary contributions include the large coco dataset segmentation and the novel SAM model. This branch of research is entirely devoted to pushing boundaries, as seen by the publication of over 1 billion masks and the creation of commandment segmentation. These achievements are expected to inspire more research and imagination and lay the foundations for future attempts. Future research will focus on SAM's ongoing development and improvement to increase its capabilities, including better real-time performance and more accurate object delineation. The combination of SAM and YOLO-NAS into an integrated framework for inclusive image and video evaluation reveals considerable capabilities. Privacy, security, and fairness in segmentation applications will all be assured by ethical considerations surrounding picture segmentation and the responsible use of this technology. The route ahead carries an opportunity for research, imaginative thinking, and moral contemplation, influencing how we work with visual data in the years to come.

VII. REFERENCES

[1] A. Kirillov et al., “Segment Anything,” 2023, doi: 10.48550/ARXIV.2304.02643.

[2] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, “Track Anything: Segment Anything Meets Videos,” 2023, doi: 10.48550/ARXIV.2304.11968.

[3] S. He et al., “Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets,” 2023, doi: 10.48550/ARXIV.2304.09324.

[4] Matheus Henrique Fonseca Afonso, E. H. Teixeira, M. Cruz, G. P. Aquino, and E. C. V. Boas, “Vehicle and Plate Detection for Intelligent Transport Systems: Performance Evaluation of Models YOLOv5 and YOLOv8,” 2023, doi: 10.13140/RG.2.2.11022.95042.

[5] Y. Cheng et al., “Segment and Track Anything,” 2023, doi: 10.48550/ARXIV.2305.06558.

[6] J. Cao, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, “SipMask: Spatial Information Preservation for Fast Image and Video Instance Segmentation,” 2020, doi: 10.48550/ARXIV.2007.14772.

[7] K. Psychogyios, H. C. Leligou, F. Melissari, S. Bourou, Z. Anastasakis, and T. Zahariadis, “SAMStyler: Enhancing Visual Creativity With Neural Style Transfer and Segment Anything Model (SAM),” IEEE Access, vol. 11, pp. 100256–100267, 2023, doi: 10.1109/ACCESS.2023.3315235.

[8] A. Athar, J. Luiten, A. Hermans, D. Ramanan, and B. Leibe, “HODOR: High-level Object Descriptors for Object Re-segmentation in Video Learned from Static Images,” 2021, doi: 10.48550/ARXIV.2112.09131.

[9] Y. Zhang and R. Jiao, “Towards Segment Anything Model (SAM) for Medical Image Segmentation: A Survey,” 2023, doi: 10.48550/ARXIV.2305.03678.

[10] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, “Segment anything model for medical image analysis: An experimental study,” Med. Image Anal., vol. 89, p. 102918, Oct. 2023, doi: 10.1016/j.media.2023.102918.