

Experiences Memory Network for Video Semantic Segmentation

Shalaka Deore¹, Alfiya Shahabad², Shobha Raskar³, Akanksha Pathak⁴, Atharva Deore⁵ and Harsh Dalvi⁶

¹⁻⁶Department of Computer Engineering, Modern Education Society's Wadia College of Engineering, Pune 411001, India
Email: shalaka.deore@mescoepune.org

Abstract—Human memory, while remarkable, struggles to retain and recall intricate details from various experiences. This abstract proposes a novel “Experience Memory Network” to address this limitation in the context of video semantic segmentation. The network functions as a fully connected architecture, comprising information storage units that retain data on individual video frames and relations that capture the connections between them. To facilitate efficient information retrieval and updates, the network incorporates specialized read and write modules. This architecture empowers the system to leverage past experiences (stored in the memory network) to perform semantic segmentation on new video frames. Essentially, the network learns from each frame and utilizes this knowledge to understand and segment subsequent frames, leading to improved overall performance. The segmentation evaluation of Experience Memory Network yields two crucial metrics: mean region similarity (80.2%), indicating the overall overlap between predicted and true regions, and mean contour accuracy (84.2%), which specifically assesses how well predicted contours align with actual object boundaries. Analysing both metrics together provides a holistic understanding of segmentation effectiveness, encompassing both the extent of overlap and the precision of boundary delineation.

Index Terms— Video Semantic segmentation, Experience memory network, Dynamic mask utilization, Key frame identification.

I. INTRODUCTION

While the past decade has witnessed a meteoric rise in the efficacy of deep learning, largely fueled by breakthroughs in supervised learning paradigms, its continued progress hinges on the crucial, yet often arduous, task of acquiring vast quantities of meticulously annotated training data. Imagine dissecting an image down to its tiniest details, assigning each pixel a label based on its “meaning” - that’s essentially the essence of semantic segmentation. This ability to grasp the scene like never before has opened doors in various fields, from self-driving cars navigating their surroundings to environmental monitoring [3], [4] and even medical imaging analysis [1], [2]. However, similar to a student needing comprehensive study material, these cutting-edge approaches rely heavily on vast amounts of labelled data, where each pixel is meticulously tagged with its specific category. This data collection process can be quite cumbersome, but it forms the foundation for these powerful applications. Recognizing the data labelling bottleneck in semantic segmentation, researchers have pivoted towards self-supervised learning, a novel approach that unveils the inherent semantic richness within images without relying on human-generated labels. This exciting shift has been propelled by advancements in

many self-supervised representation learning algorithms [5]–[7]. These techniques empower the extraction of meaningful features from data without the need for explicit, pre-defined labels, paving the way for a more efficient and scalable approach to training semantic segmentation models. Amongst the many innovative approaches in this field, the DINO pretraining strategy stands out. When paired with the powerful Visual Transformer (ViT) architecture, DINO can analyse an image and understand its deeper meaning without any human-provided labelling. We’ll refer to this powerful combination as a “DINO backbone” and freeze this backbone just for the purpose of feature extraction. Building on the DINO backbone, we are using the approach [11] which refines how the model understands images. It does this by focusing on connections between different parts of the image without needing any human labelled examples. This approach, called “self-supervised correlation loss,” is clever because it improves the model’s understanding without requiring additional computing power. The results are impressive – Unsupervised Semantic Mask Generator (USMG) model performs remarkably well in the challenging task of understanding the meaning of images entirely on its own, achieving top results in unsupervised semantic segmentation tasks. This semantic segmentation mask serves as a potent tool for various applications demanding an intricate understanding of image content. While existing image segmentation models like FCN [8] and DeepLab [9], paper [10] have shown promise, video introduces unique challenges: capturing the intricate interactions between frames (temporal information) and the sheer amount of data compared to images, which strains computational resources. Existing approaches often analyse individual frames in isolation, neglecting valuable temporal information. To address this, we propose a Networked memory of experiences, a novel architecture that allows dynamic adjustment within a single pass, eliminating the need for multiple iterations. This memory empowers the model to learn and adapt dynamically throughout the video analysis process. This is achieved by incorporating valuable information from the very first frame of the video. This initial frame is semantically segmented using the unsupervised approach [11], which essentially categorizes each pixel in the image without the need for manual labelling. By leveraging this information, the model can begin to understand the scene and adjust its predictions as it analyses subsequent frames. This real-time learning capability enhances the model’s performance and allows it to adapt to the specific content of each video, so it doesn’t get stuck in a rut! i.e. it constantly updates its knowledge throughout the video analysis.

II. RELATED WORK

A. Video Semantic Segmentation

Recently, there’s been a surge in interest in video semantic segmentation! Well, it’s all about the ability to comprehend what’s unfolding in videos, analysing each frame step by step. This capability holds immense practical value for various real-world applications. Unlike its image counterpart, video segmentation goes beyond labelling objects in a static picture. Building on the success of image semantic segmentation techniques, researchers have explored various approaches to tackle the challenges of video segmentation. A significant portion of this research focuses on exploiting the temporal information inherent in videos. Some methods leverage techniques like recurrent neural networks [12], or attention modules [14] to analyse features from multiple frames simultaneously. This allows them to “remember” what they’ve seen in previous frames and make more consistent and accurate predictions throughout the video. Our paper proposes a novel approach to video semantic segmentation by leveraging the power of “experience memory networks.” This network, the heart of our system, relies on strategically chosen key frames extracted from the video sequence. These key frames act as crucial building blocks, providing the network with a rich tapestry of information. By meticulously analysing and storing the features extracted from these key frames, the experience memory network builds a comprehensive understanding of the video’s content.

B. Self-Supervised Semantic Segmentation

Traditional semantic segmentation relies on vast amounts of meticulously labelled data, where each pixel is assigned a specific category. However, acquiring such data can be a laborious and expensive process. This is where self-supervised learning steps in, offering a revolutionary approach to semantic segmentation. The realm of self-supervised semantic segmentation is witnessing a surge in exciting research [15]. Researchers are exploring methods that exploit the inherent relationships between pixels in an image, allowing the model to discern the spatial structure and relationships between objects. These advancements highlight the vibrant landscape of self-supervised semantic segmentation and its potential to revolutionize how we train models for understanding the visual world [16]. Our work delves into the intriguing realm of self-supervised semantic segmentation, aiming to unlock the semantic meaning hidden within individual video frames. By harnessing this

technique, we generate semantically segmented masks, acting as comprehensive maps that reveal the semantic content of each frame. These masks subsequently serve as crucial elements within our experience memory network, a cornerstone element of our video semantic segmentation system. This network meticulously stores and utilizes these masks, empowering it to dynamically adjust and refine its understanding throughout the video analysis process [11].

C. Space Time Memory

Understanding the ever-changing nature of videos poses a unique challenge for semantic segmentation. Unlike static images, videos require models to not only decipher the content within each frame but also grasp the intricate relationships between them. Here's where space-time memory networks enter the scene. The concept of space-time memory networks has ignited significant research interest in the field of video segmentation [17,18]. Additionally, researchers are investigating methods for efficiently storing and accessing information within the memory, ensuring optimal performance during segmentation tasks. These advancements highlight the space-time memory network's potential as a powerful tool for unlocking the hidden meaning within video data [19]. In our paper, we embark on a journey to achieve video semantic segmentation. However, our primary focus lies in establishing a novel "experience memory network." This network acts as a repository, meticulously storing spatiotemporal features, which are critical for deciphering the meaning of each pixel within a video sequence.

III. METHODOLOGY

A. Data Pre-processing

The raw data for this research comprises video frames captured from real-world environments, encompassing both indoor and outdoor scenes. These video frames serve as the input to the model. Here are the key details regarding the data preprocessing step: Source: The video frames are extracted from real-time video recordings. Scene Diversity: The videos capture a variety of environments, including both indoor and outdoor locations.

B. Semantic Segmented Mask

The Unsupervised Semantic Mask Generator (USMG) model for creating semantic segmentation output involves a series of steps to analyse and exploit the relationships between unlabelled video frames. Firstly, this method initiates by preparing the data. This includes generating KNN indices, which are essentially reference points used to identify similar features within the dataset. Next, the actual implementation of this method begins. This involves running a script that utilizes a shallow segmentation network, built upon a powerful image processing backbone like the DINO ViT-base architecture. This network learns by comparing connections between different parts of the same image and similar images within the dataset. This comparison is facilitated by three contrastive terms, which essentially guide the network to "distill" the correspondences between features in the images, allowing it to understand how objects relate to each other. By leveraging contrastive loss and carefully analyse the feature correspondences, USMG aims to achieve unsupervised semantic segmentation, meaning it can categorize elements within video frames without the need for manual labelling.

C. Experience Memory Network

This experiences memory network combines a special networked memory that stores information in a connected format with intelligent modules that can access and process that information. The memory acts like a temporary holding ground for newly learned information, and its interconnected design helps us understand the bigger picture. The modules interact with the networked memory using read and write operations. It can also learn and remember things over time by gradually adjusting its internal connections. The system, through its module mechanisms, figures out the best way to store information in its networked memory and how to use that information later to make segmentation decisions. The model built around a memory network with connections, analyses information in stages to effectively extract relevant patterns and focus on details crucial to the task at hand. Specifically, the networked memory is limited in capacity, fully connected memory network $N = (I, E)$, where node $im \in I$ denotes m th Information storage unit, and edge $em, n = (im, in) \in E$ indicates the relation between cell im and in . The video sequence consists of multiple image frames which contain the duplicate information. Such image frames are of no use. So we identify the key image frames based on similarity scores. First we compute the similarity score between the first and the rest of the frames until we get a difference between the similarities to go beyond a predefined threshold. Such frames are identified as key frames. The key image frames are taken as reference information. Given an image frame to be segmented, the Reference information is supposed to be consisting of the first labelled image frames and key image frames along with image frames that

are already segmented. The networked memory is Prepared from $C(= |I|)$ frames, taken from Reference information. The mask for the first frame (GROUND TRUTH LABELS) are generated by USMG. Each information storage unit Starting representation $i_0 \in \mathbb{R}^{W \times H \times C}$ is obtained from a fully convolutional network memory encoder from the Reference information image frame to grasp its visual features and identify different classes within it.

D. Information Retrieval from the Memory Network

A fully convolutional image frame encoder is also applied to the current image frame under segmentation to extract the visual feature $i \in \mathbb{R}^{W \times H \times C}$. A Trainable reading module takes i as input and generates its Baseline understanding the reading module interacts with the network memory by examining the details. At each iteration, the read module interacts with the experience graph memory by extracting the contents Using the key-value retrieval mechanism in [20][22], we first compute the similarity between the current image frame and each Information storage unit im . Afterwards we calculate the read weight $w_{k,m}$ by a softmax normalization function. Then the memory summarization i_k is calculated based on their importance scores of the Information storage unit. Then, read module refines its understanding of the situation

E. Updating information in Memory Network

After considering the summarized memory, there is a need to update the experience memory network. In each k th iteration, a trainable memory write module updates each Information storage unit im by considering its previous state m_{k-1} , current content from the read module, the states from other Information storage unit $\{i_{k-1} \}_{m=n}$. We first assess the relationship $(r_{k,m,n})$ between two pieces of information (i_n and im) by comparing their core attributes using a special scoring method like inner-product similarity. We then create a summary $(s_{k,m})$ for each piece of information (im) by considering related information from other parts of the system. The importance of each related piece of information is determined by how similar it is to the original information (im), considering a special scoring method. After incorporating insights from neighbours, the memory write module, refines the internal representation of im . We also made it possible to combine multiple video sequences and use the mask of each video sequence automatically to perform segmentation based on the similarity computation between consecutive frame. Then the difference between similarities is computed and compared with a threshold. When the difference goes beyond the threshold the mask of the corresponding video sequence is used for further segmentation task. As a result, we can dynamically use the video sequence specific mask for segmentation.

F. Segmentation Prediction

After K iterations, we predict the segmentation mask for the current image frame under segmentation.

IV. ARCHITECTURE

A. For Unsupervised Mask Generation

USMG, an intricate system proficient in deciphering unmarked images, employs a meticulously designed architecture to realize its objective. Below is a detailed exposition of its fundamental constituents: Vision Transformer (ViT): At the onset, this stage dissects the image into 8×8 squares, denoted as patches. Each patch undergoes a transformation into a rudimentary code that encapsulates its intrinsic characteristics. To comprehend the spatial arrangement, learnable positional information is infused into these codes. Finally, the processed outputs are reorganized into a feature map, essentially reconstructing the original image while assimilating the newfound comprehension of its contents. It is noteworthy that the ViTbase backbone employed in this process extracts features with a dimensionality of 768. Segmentation Head: This specialized component functions akin to a decoder, elucidating the significance embedded within the image. It operates in two stages: Individual Feature Analysis: Scrutinizing each element (feature) within the image independently, irrespective of potential influences from neighbouring features. This allows the system to concentrate on the specific characteristics of each pixel. Code Simplification: Condensing the intricate code generated earlier (FeViT) into a more manageable form (FeUSMG), with a diminished dimensionality of 80 for streamlined calculations. K-Means Clustering: USMG employs the k-means clustering algorithm to efficiently organize the unmarked data. This algorithm aggregates similar data points (image regions in this context) based on their features, with the objective of minimizing the distance between each point and its assigned cluster centre. Conditional Random Field (CRF) (Optional): The initial cluster formation can undergo further refinement through the application of a CRF. Visualize the initial clusters as an initial attempt to group similar data points. The CRF functions as a meticulous editor, scrutinizing

this preliminary assumption and making adjustments to enhance overall accuracy by considering not only individual points but also the relationships between them. This optional step can refine the final clusters, culminating in a more resilient and precise segmentation output. Fig.1 depicts the architecture of USMG.

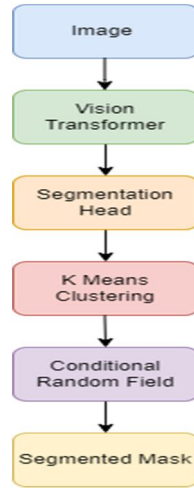


Fig 1. Working of Unsupervised Semantic Mask Generator

B. Experiences Memory Network

Full Network Architecture Our system operates entirely using convolution layers to analyse information from start to finish. The memory and image frame encoders have identical architecture, with the only difference being the inputs. The image frame encoder takes an RGB image frame as input, while, for the memory encoder, input is an RGB support frame concatenated along with a semantic mask. The read module and write module make use of ConvGRU [23], with 1×1 convolutional kernels. The decoder consists of four blocks with skip connections to the corresponding ResNet50 [24] blocks. The kernel size of each convolution layer in the decoder is set as 3×3 , except the last 1×1 convolution layer. The resulting 2-channel segmentation prediction is achieved by softmax operation. Image frame and memory encoders are carried out using four convolution blocks of ResNet50. Fig.2 depicts the architecture of the experience memory network model.

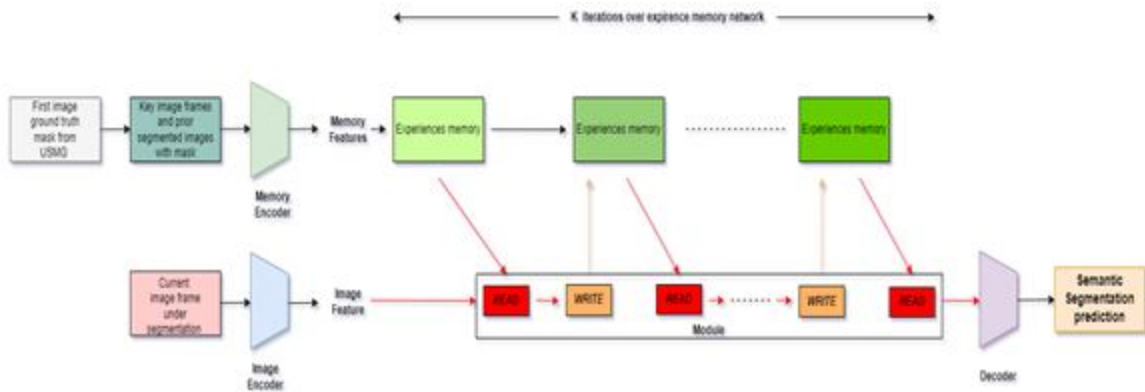


Fig 2. Working of experience memory network

V. IMPLEMENTATION

A. USMG

The self-supervised system's image understanding process starts by dividing the image into smaller squares, like a mosaic. Each square is then transformed into a basic code, capturing its essence. To understand the image's layout, the system adds location information to these codes. These enriched codes are then passed through multiple processing layers, allowing the system to identify connections between different parts of the image. The processed outputs are then reassembled into a map, similar to the original image, but now enriched with a deeper understanding of its content. A dedicated component then analyses each pixel individually and simplifies the

complex code. Finally, these simplified codes are used to group similar image regions based on their features using clustering algorithm. This initial grouping can be further refined for improved accuracy using conditional random field.

B. Experiences Memory Network

For pre-training, we crop 384×384 Region samples from static images, and generate a video clip of length $N+1=4$. For training, 384×640 patches are cropped from real training videos. To create training clips, we randomly pick four frames in order from the same video, making sure each frame isn't more than 25 frames away from the one before it. We use tricks like rotating, flipping, and adjusting the colour vibrancy of the data to create more variety in our training information. We use a batch size which is set to 16. The loss function is optimized with Adam optimizer using a "poly" learning schedule, with the base learning rate of $1e-5$ and power of 1.0.

VI. RESULTS

To evaluate the effectiveness of the proposed experience memory network model, we conducted experiments on the DAVIS dataset, containing 60 videos for training and 30 videos for validation. Each video sequence consists of multiple frames. The dataset encompasses diverse indoor and outdoor scenes. We were able to segment multiple video sequences and use the mask of each respective video sequence to perform segmentation. Based on the key frames identified, the experience memory network is able to use the spatio-temporal features stored in the memory to segment the video frames and accurate segmentation outputs are obtained. Fig.3 Depicts the qualitative results obtained. We evaluated our model using the same methods as a widely accepted benchmark. This evaluation considers two key aspects: how well the system identifies similar regions in each frame (mean region similarity J) and how accurately it outlines the objects (contour accuracy F). Our experience memory network achieves good scores of mean J & F (82.2%), mean J (80.2%) and mean F (84.2%). The model scores significantly better in both how well it identifies similar areas within each video frame (region similarity) and how accurately it outlines the objects themselves (contour accuracy).



Fig 3. Qualitative Results

VII. CONCLUSION

This research involves a novel experience memory network mechanism that allows segmentation network to handle videos without needing extensive training or adjustments. By iterating over experience memory network, the model can create a concise overview specific to each video. This summary significantly improves the final segmentation prediction greatly. In the meantime, trainable modules can perform Dynamic model adjustment. Hence we conclude that Our Experience Memory Network can efficiently segment the video frames with a mean contour accuracy of 84.2% and mean region similarity of 80.2%.

While our current research demonstrates the effectiveness of the proposed method, we acknowledge several areas for future exploration. Our current implementation does not employ a colour map to differentiate objects using distinct colours based on their class labels. In future work, we aim to establish a well-defined colour map scheme that assigns unique colours to different object classes, enhancing the visual interpretability of the segmentation results. This additional layer of information will provide a more intuitive understanding of the segmented image content.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to publish this paper.

REFERENCES

- [1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. UNet: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention, 2015

- [2] Monjoy Saha and Chandan Chakraborty. Her2Net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. *IEEE Transactions on Image Processing*, 2018
- [3] Tianrui Guan, Divya Kothandaraman, Rohan Chandra, Adarsh Jagan Sathyamoorthy, Kasun Weerakoon, and Dinesh Manocha. GA-Nav: Efficient terrain segmentation for robot navigation in unstructured outdoor environments. *IEEE Robotics and Automation Letters*, 2022.
- [4] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Conference on Computer Vision and Pattern Recognition*, 2021.
- [5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.
- [7] Jean-Bastien Grill, Florian Strub, Florent Altche, Corentin ´ Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi ´ Munos, and Michal Valko. Bootstrap your own latent – a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.
- [8] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015
- [9] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei. Deep feature flow for video recognition. In *CVPR*, pages 2349–2358, 2017.
- [10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017.
- [11] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snaveley, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2022
- [12] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi. Temporally distributed networks for fast video semantic segmentation. In *CVPR*, pages 8818–8827, 2020
- [13] P. Lei and S. Todorovic. Recurrent temporal deep field for semantic video labeling. In *ECCV*, pages 302–317. Springer, 2016.
- [14] D. Nilsson and C. Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *CVPR*, pages 6819–6828, 2018.
- [15] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised ~ image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9865–9874, 2019.
- [16] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *arxiv preprint arxiv:2102.06191*, 2021.
- [17] Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: *ICCV 2019*
- [18] Bao, L., Wu, B., Liu, W.: Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In: *CVPR 2018*
- [19] Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: *CVPR 2013*
- [20] Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R.: Ask me anything: Dynamic memory networks for natural language processing. In: *ICML 2016*
- [21] Miller, A., Fisch, A., Dodge, J., Karimi, A.H., Bordes, A., Weston, J.: Key-value memory networks for directly reading documents. In: *EMNLP 2016*
- [22] Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R.: End-to-end memory networks. In: *NIPS 2015*
- [23] Ballas, N., Yao, L., Pal, C., Courville, A.: Delving deeper into convolutional networks for learning video representations. In: *ICLR 2016*
- [24] He, K., Zhang, X., Ren, S., Sun, J: Deep residual learning for image recognition. In : *CVPR 2016*.