Pattern Recognition Letters 192 (2025) 59–64

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

FoodMem: Near real-time and precise food video segmentation

Ahmad AlMughrabi [a,*], Adrián Galán [a], Ricardo Marques [b,1], Petia Radeva [a,c,1]

[a] Universitat de Barcelona, Gran Via de les Corts Catalanes, 585, Barcelona, 08007, Spain
[b] Interactive Technologies Group (GTI), Universitat Pompeu Fabra, Carrer de Tànger, 122-140, Barcelona, 08018, Spain
[c] Institut de Neurosciències, University of Barcelona, Passeig de la Vall d'Hebron, 171, Barcelona, 08035, Spain

ARTICLE INFO

Editor: Gangyi Jiang

Dataset link: http://doi.org/10.6084/m9.figshare.28638806

Keywords:
Video food segmentation
Fast segmentation
Food tracking
Segmentation transformer
Memory-based models
Near real-time segmentation

ABSTRACT

Food segmentation, including in videos, is vital for addressing real-world health, agriculture, and food biotechnology issues. Current limitations lead to inaccurate nutritional analysis, inefficient crop management, and suboptimal food processing, impacting food security and public health. Improving segmentation techniques can enhance dietary assessments, agricultural productivity, and the food production process. This study introduces the development of a robust framework for high-quality, near-real-time segmentation and tracking of food items in videos, using minimal hardware resources. We present FoodMem, a novel framework designed to segment food items from video sequences of 360-degree unbounded scenes. FoodMem can consistently generate masks of food portions in a video sequence, overcoming the limitations of existing semantic segmentation models, such as flickering and prohibitive inference speeds in video processing contexts. To address these issues, FoodMem leverages a two-phase solution: a transformer segmentation phase to create initial segmentation masks and a memory-based tracking phase to monitor food masks in complex scenes. Our framework outperforms current state-of-the-art food segmentation models, yielding superior performance across various conditions, such as camera angles, lighting, reflections, scene complexity, and food diversity.[2] This results in reduced segmentation noise, elimination of artifacts, and completion of missing segments. We also introduce a new annotated food dataset encompassing challenging scenarios absent in previous benchmarks. Extensive experiments conducted on MetaFood3D, Nutrition5k, and Vegetables & Fruits datasets demonstrate that FoodMem enhances the state-of-the-art by 2.5% mean average precision in food video segmentation and is 58× faster on average. The source code is available at: [3].

1. Introduction

Video Object Segmentation (VOS) [1] stands as a prevalent task within the domain of computer vision, finding applications in various fields, including object recognition, scene comprehension, medical imaging, and enhancing filter effects in video communication. While automated methodologies leveraging pretrained models for object segmentation are widely utilized, interactive user input is frequently incorporated to annotate novel training datasets or accomplish precise rotoscoping for intricate footage, particularly within visual effects contexts. This becomes particularly beneficial when videos present challenging lighting conditions and dynamic scenes or necessitate partial region segmentation. Although automatic VOS techniques strive to delineate entire objects with clearly defined semantic boundaries, Interactive Video Object Segmentation (IVOS) and Semi-supervised Video Object Segmentation (SSVOS) approaches [2,3] offer enhanced flexibility. These typically entail the utilization of scribble or contour drawing interfaces for manual refinement.

Cutting-edge IVOS and SSVOS methodologies rely on memory-based models [4,5] and have shown impressive segmentation outcomes in complex scenes through user-supplied mask annotations. However, these techniques are primarily designed to improve individual annotation performances [6–9], making them unsuitable for practical production scenarios. They tend to over-segment recognized semantic contours (such as individuals, hair, faces, and entire objects) while struggling to accurately delineate partial regions (such as parts of a person's face or a dog's tail), and face challenges with lighting and extreme object orientations. As a result, when given only one annotated frame, inconsistent segmentations arise due to the inherent ambiguity in the object's appearance, particularly when subjected to significant variations in viewing angles and complex lighting conditions. This

* Corresponding author.
E-mail addresses: ahmad.almughrabi@ub.edu (A. AlMughrabi), ricardo.marques@upf.edu (R. Marques), petia.ivanova@ub.edu (P. Radeva).
[1] Equal supervision.
[2] More details in the supplementary material.
[3] https://amughrabi.github.io/foodmem

https://doi.org/10.1016/j.patrec.2025.03.014
Received 16 July 2024; Received in revised form 29 January 2025; Accepted 7 March 2025
Available online 25 March 2025

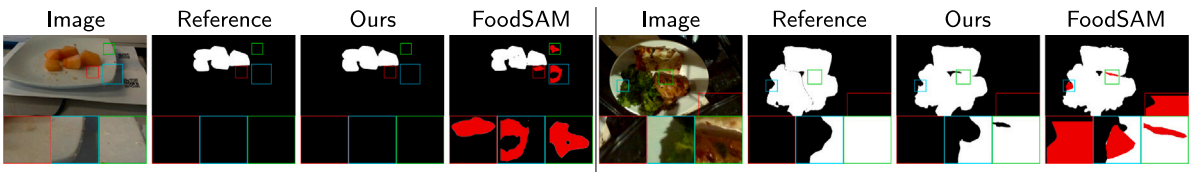

Fig. 1. Comparison among original images, ground truth masks, masks produced by FoodMem, and masks produced by FoodSAM using the Nutrition5k dataset. The color red region highlights any artifacts or missing segments. Our method demonstrates robust performance under complex food geometry and low-light conditions.

limitation limits the scalability of these techniques, as it is unclear which frame annotations are a priority, especially in extended sequences. A new SSVOS framework, XMem++ [4], stores annotated frames for reference and handles multiple frame annotations iteratively. XMem++ [4] adopted XMem [6] as a backbone. It improves object segmentation in complex scenes with fewer annotated frames without re-training. Using attention-based frame suggestion, it predicts the next annotations, supporting both sparse and dense annotations for better quality scaling. The system achieves real-time video segmentation with instant consideration of frame annotations.

In contrast, a novel zero-shot framework, FoodSAM [10], is designed to integrate the original semantic mask with category-agnostic masks generated by the Segment Anything (SAM) [11]. While SAM exhibits impressive food image segmentation capabilities, its lack of class-specific information limits its effectiveness [10]. Conversely, conventional segmentation methods maintain category information but often sacrifice segmentation quality [10]. FoodSAM proposes fusing the original segmentation output with SAM's generated masks to enhance semantic segmentation quality. Identifying the mask's category based on its predominant elements represents a novel and practical approach to improving the semantic segmentation process [10]. Still, FoodSAM fails to generalize the segmentation across multiple frames in a given video; for instance, the estimated mask IDs (i.e., mask colors) are assigned differently for the same food portion in different frames for the same scene. Moreover, FoodSAM fails to segment perfectly and generates masks from different camera views, such as missing food parts, and segments non-food objects, such as plates. Furthermore, FoodSAM is slow, which makes it an unsuitable scene segmentation framework for production scenarios.

To overcome these limitations, we propose FoodMem, a hybrid framework that uniquely leverages the semantic segmentation capabilities of SeTR with the memory-based tracking efficiency of XMem++. This fusion allows FoodMem to track and refine food object segmentations across video frames with minimal annotations while maintaining robustness under complex lighting and food geometries with varying viewing angles. Our framework generates one or a few annotated images as a first step and then tracks them throughout the scene. We further introduce a dataset for benchmarking purposes, which includes the new challenging scenes and food use cases, including: 1. **different capturing settings** such as 360° food scenes; high and low speed capturing; blurry images; complex and simple backgrounds; different illumination; 2. **food diversity and complexity** such as basic ingredients (e.g., apple) or complex ingredients (e.g., cuisine). 3. **bounded and unbounded scenes** such as scene has been captured by Mobile phone in a free movement; while some scene has been captured by Intel Realsense camera in a restricted movement. We extensively evaluate and compare our framework on our dataset, which is built from real-world and challenging datasets [12,13] that are publicly available to show how our framework outperforms the current state-of-the-art.

2. Related work

Various approaches have been proposed to tackle the food segmentation challenge. FoodSAM [10] has achieved a groundbreaking milestone as the first system to accomplish instance segmentation, panoptic segmentation, and promptable segmentation. FoodSAM utilizes the ViTh [14] as SAM's image encoder, employing the same hyperparameters outlined in the original paper. Object detection is performed using UniDet [15] with Unified learned COIM RS200. For semantic segmentation, SeTR [16] serves as FoodSAM baseline, incorporating ViT-16/B as the encoder and MLA as a decoder in FoodSeg103 [17], which SeTR focuses on extracting richer feature representations. Following a comprehensive evaluation of the FoodSeg103 [17] and UECFoodPix Complete [18] benchmarks, FoodSAM has outperformed the state-of-the-art methods on both datasets, demonstrating the exceptional potential of SAM as a powerful tool for food image segmentation. However, FoodSAM [10] exhibits considerable performance drawbacks (significant latency and high memory consumption), attributable to the integration of multiple models within its framework. Also, FoodSAM cannot track food items throughout a video, as it segments each image independently.

In contrast, the k-Means++ [19] technique is extensively utilized in various food segmentation tasks [13,19] due to its ease of implementation, high speed, and the fact that it is an analytical solution requiring no training, prior knowledge or GPU to run. This method is particularly effective when the camera is positioned above the food, and the table background is simple, containing only food, a fork, a spoon, and a plate. It primarily serves as an RGB image segmentation method employing K-means clustering to differentiate the object from the background, although additional processing, such as median filtering, is necessary to address issues like uneven illumination. Subsequent steps involve converting the clustered image to grayscale, applying median filtering, Otsu thresholding, and morphological operations (erosion, dilation, opening, and closing) to enhance segmentation quality and accuracy in volume prediction, particularly for brightly colored food objects.

For video segmentation, DEVA [20] is a decoupled video segmentation approach that uses task-specific image-level segmentation and class/task-agnostic bi-directional temporal propagation, allowing it to 'track anything' without needing video data training for each individual task. This method only requires training an image-level model for the target task and a universal temporal propagation model once, which generalizes across tasks. DEVA outperforms end-to-end approaches in data-scarce tasks, including large-vocabulary video panoptic segmentation, open-world video segmentation, referring video segmentation, and unsupervised video object segmentation. Using DEVA with the prompt "food" demonstrates good overall model performance and speed; however, it exhibits limitations in recognizing a diverse range of food items and consumes a high amount of memory.

To overcome these limitations, we propose FoodMem, a production-grade video food segmentation framework that accepts a video or a set of images and segments food in near-real-time performance and memory-friendly. In contrast to prior approaches, which process each frame independently, FoodMem utilizes stored annotations to ensure consistent tracking of food items across video sequences. The combination facilitates near-real-time segmentation performance and enhances memory efficiency, thereby improving the overall robustness and applicability of food segmentation in diverse production environments. Our main contributions are as follows:

1. We build a novel near-real-time food segmentation architecture for videos that combine Segmentation Transformer (SeTR) [16] and memory-based model [4] frameworks as a first exploration for video food segmentation.

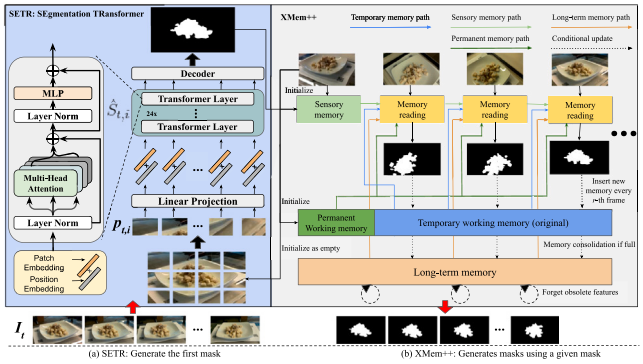

Fig. 2. FoodMem model architecture. We used a single image input for simplicity. Our two-stage framework (a) shows the SeTR framework, where it accepts an image and generates a mask, followed by (b) a memory-based framework, which accepts the mask and a set of images as a given input and produces masks for all frames.

2. We introduce a novel dataset tailored for food image segmentation tasks. Our dataset comprises a comprehensive selection of dishes sourced from the Nutrition5k [13] dataset, encompassing 31 distinct dishes with a total of 1356 annotated frames. Additionally, we include 11 dishes from the Vegetable and Fruits [19] dataset, augmented with 2308 annotated frames. Our dataset features 42 diverse dishes, accompanied by 3664 meticulously annotated frames. We believe this expansive dataset is a valuable resource for advancing research in video food segmentation and related computer vision tasks.
3. We conducted an extensive series of experiments on our dataset to assess the effectiveness and flexibility of our framework against the baselines.
4. Our framework outperforms the state-of-the-art performance in video food segmentation for 2.5% mean average precision with similar recall.
5. Our framework is 58 times faster than the baseline's [10] inference time.

The rest of the paper is structured as follows: We present the theoretical background in Section 3. We define our proposed methodology in Section 3.2. A thorough set of experimental results is presented in Section 4. Finally, we present our conclusions and future work in Section 5.

3. Proposed methodology

This section details the structure of our proposed method for automated food portion segmentation and tracking. Section 3.1 provides a high-level overview of the methodology, with reference to Fig. 2. Initially, we employ the SeTR model [16] to generate one or several food masks per scene. Subsequently, these masks are tracked within complex scenes using the XMem++ model [4]. The following sections delve into the specifics of each step in our approach.

3.1. Overview

Our framework leverages a two-phase framework, combining SeTR and memory-based frameworks to achieve efficient and rapid video segmentation, addressing the critical time complexity issue. This combination is inherently complex due to their distinct architectural designs and operational methodologies. Merging these models is non-trivial and involves significant challenges in aligning their functionalities and optimizing their interoperability. We have developed robust solutions to seamlessly integrate these components, ensuring precise segmentation and tracking while maintaining computational efficiency.

3.2. Our proposal: FoodMem

Our methodology starts with semantic segmentation with the pre-trained SeTR model for food segmentation in video data. Specifically, we utilized the Sequence-to-Sequence Transformer (SeTR) model [16], leveraging pre-trained weights to perform initial semantic segmentation on the first n frames $\{I_t\}_{t=1}^{n}$ of the video sequence. Here, I_t denotes the t-th frame of the video sequence. The video frames are partitioned into N non-overlapping patches $\{p_{t,i}\}_{i=1}^{N}$, each of size $P \times P \times C$. In this context, P represents the height and width of each square patch, and C denotes the number of color channels (e.g., $C = 3$ for RGB images). These patches are embedded into feature vectors $E_{t,i} \in \mathbb{R}^D$ using a linear projection, where D denotes the dimensionality of the embedded feature vector $E_{t,i} = W_p \cdot \text{flatten}(p_{t,i}) + b_p$, where W_p and b_p are the learnable parameters of the linear projection. These feature vectors are augmented with positional embeddings P_i to encode spatial relationships. The sequence of embedded patches $Z_{t,0} = \{E_{t,i} + P_i\}_{i=1}^{N}$ undergoes multiple layers of transformer encoding, leveraging pre-trained weights to enhance feature extraction capabilities. Here, $Z_{t,0}$ represents the initial sequence of embedded patches with positional embeddings. The output of the transformer encoder $Z_{t,L}$ is subsequently fed into the transformer decoder to refine feature representations. Segmentation predictions are generated by applying a linear projection followed by a softmax activation to produce the probability distribution [4,6]: $\hat{S}_{t,i} = \text{softmax}(W_s Z_{t,L}^i + b_s)$, for each patch i. In this case, $Z_{t,L}$, where L denotes the number of layers in the transformer encoder, is the output of the transformer encoder after L layers, and W_s and b_s are the learnable parameters for the linear projection in the decoder. The predicted segmentation probability distribution for patch i in frame t is denoted by $\hat{S}_{t,i}$, as shown in Fig. 2(a).

We leverage Long-Term Tracking with XMem++ [4,6] to maintain temporal coherence in segmentation across frames. We integrate the XMem model without training, focusing on memory mechanisms for tracking and refining segmentation masks. For each frame I_t, feature maps $F_t = \phi(I_t)$ are extracted and used to populate Short-Term Memory (STM) and Long-Term Memory (LTM). Here, ϕ denotes the feature extraction function, and F_t denotes the feature map extracted from frame I_t. STM retains recent feature maps [4,6]: $\text{STM}_t = \{F_{t-n}, \ldots, F_t\}$, while LTM accumulates feature maps selectively based on their significance [4,6]: $\text{LTM}_{t+1} = \text{LTM}_t \cup \{F_t\}$, To track segmentation masks over time, similarities $S_{t,i}$ between the current feature map F_t and memory features M_i are computed using dot products, normalized to obtain attention weights [4,6]: $\alpha_{t,i} = \frac{\exp(S_{t,i})}{\sum_j \exp(S_{t,j})}$, where $S_{t,i}$ represents the similarity between the current feature map F_t and memory feature M_i, and $\alpha_{t,i}$ denotes the attention weight for memory feature M_i. The memory read R_t is computed as a weighted sum of memory features [16]: $R_t = \sum_i \alpha_{t,i} M_i$. Fig. 2(b) illustrates Long-Term Tracking method, which takes the mask and a set of images as input and generates masks for all frames.

3.3. FoodMask: Our proposed dataset

We utilized the Nutrition5k [13] dataset, which comprises approximately 5000 dishes, offering a wide array of food items for analysis. For comprehensive testing, we carefully selected 31 plates from this dataset. Each plate was chosen to encompass diverse compositions of ingredients, ensuring a broad scope for evaluating our segmentation model's performance across various food categories. Each selected plate contains between 130 and 560 frames, totaling several hundred per plate. To streamline processing and reduce computational overhead, we employed an image processing pipeline to minimize the number of frames, specifically Imagededup [21]. This reduction consolidated the frames to approximately 20 to 65 keyframes per plate, preserving critical visual information essential for accurate segmentation analysis. Additionally, we incorporated the Vegetables & Fruits (V&F) dataset into our evaluation [19]. Like Nutrition5k, V&F is a substantial dataset

containing multiple dishes and frames. We strategically selected a subset of dishes from both datasets based on food variation, video complexity, and ingredient diversity to ensure efficient validation of our framework. Similarly, we employed a near-image similarity approach [21], eliminating highly overlapped images and optimizing dataset integrity. We manually annotated each selected dish frame using LabelMe [22]. This meticulous annotation resulted in a dataset comprising 31 dishes from Nutrition5k with 1356 annotated frames and 11 dishes from V&F with 2308 annotated frames, totaling 42 dishes with 3664 annotated frames. This curated dataset of masks from Nutrition5K and V & F we call *FoodMask*, forms a robust foundation for rigorous testing and validation of our segmentation framework, facilitating accurate assessment across diverse culinary compositions.

The MetaFood3D (MTF3D) dataset [23] includes twenty food scenes across three difficulty levels: easy (8 scenes, 200 images), medium (7 scenes, 30 images), and hard (1 image per scene). Each scene contains food masks, depth images, a reference board, QR codes, and metadata for detailed evaluation.[1]

Our methodology ensures robust food segmentation in video data by systematically integrating pre-trained SeTR for initial segmentation and memory-based XMem for tracking without retraining. Evaluated using mean average precision and recall, our approach provides quantitative measures of segmentation accuracy and temporal coherence, demonstrating its efficacy in real-world applications such as the FoodSeg103 dataset. This combination of transformer-based segmentation and memory-based tracking advances state-of-the-art video segmentation techniques, offering reliable food recognition and analysis tools.[1]

4. Experimental results

Our framework is evaluated on two public datasets: Nutrition5k dataset [13] and Vegetables and Fruits dataset [19]. We apply two quality metrics: mAP and recall metrics.

4.1. Implementation settings

We ran the experiments using an NVIDIA GeForce RTX 2080 Ti with 11 GB of VRAM. For SeTR, we set one-shot image segmentation; for keyframe selection, we set the hamming distance to 12.[1]

4.2. Quality metrics

We outline the quality metrics used to evaluate the performance and effectiveness of our framework. These metrics are crucial for assessing the model's accuracy, efficiency, and reliability. Mean Average Precision (mAP) [24] and recall [25] are the primary metrics for evaluating the FoodMem model's performance. By focusing on these metrics, we aim to ensure that our model is accurate and reliable for practical applications in food segmentation tasks. These metrics provide insights into the model's capability to detect and segment food items effectively, which is essential for dietary assessment and culinary automation applications.[1]

4.3. FoodMem results

We extensively evaluated our model using Nutrition 5k dataset, Vegetables & Fruits, and MetaFood3D datasets. For the experiments on the datasets, we compared it to FoodSAM [10]. We reproduced the DEVA, kMean++, and FoodSAM results using the same quality metrics. Table 1 presents mAP and recall for each dataset. Briefly, we calculated quality metrics for each image in each considered scene to ensure consistency with previous works. We then determine the quality metrics at the scene level by averaging the quality of all images of the same scene. Next, we average the quality metrics across all scenes to compute the final quality values per method. This process is repeated

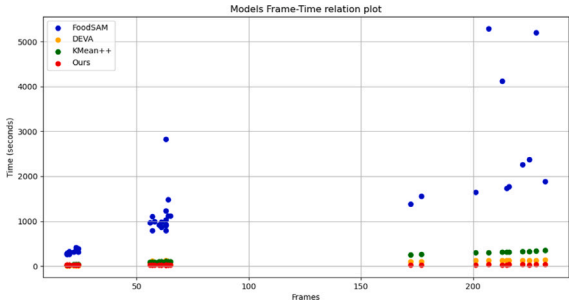

Fig. 3. The inference time for the baselines and our framework. Our framework shows a stable inference time across a different number of frames. Our inference time is 58 times faster than the baseline.

for the two datasets. Notably, our model requires only one day to complete training, which is 58 times faster than the baseline, as shown in Fig. 3. The qualitative results on the Nutrition 5k are shown in Fig. 1. Additionally, Fig. 4 shows the qualitative results on the V&F dataset. The figures show that our model excels in texture details, artifact correction, and missing data handling across different scene parts, surpassing other models. Table 1 shows the quantitative results of our comparison. The table shows that our model performs better in food segmentation on different datasets.[1]

4.4. Ablation study

To understand the factors contributing to FoodMem's success, we studied the impact of altering the masks generated by SeTR. Specifically, we examined the changes resulting from increasing the number of masks to 3, 6, or 9. We aimed to present both qualitative and quantitative outcomes of these modifications. For qualitative analysis, Fig. 5 compared the masks visually. For quantitative analysis, Table 2 assessed the execution time and evaluated performance using mean average precision and recall metrics.

4.5. Discussion

While our framework demonstrates robust performance in maintaining temporal coherence in segmentation tasks, our findings indicate that it generates artifacts in the resulting masks when multiple clean masks are provided. This issue arises from the model's reliance on feature similarity for memory read operations. When the features of multiple clean masks are highly similar, the dot product-based similarity measure used by XMem++ can lead to ambiguous attention weight assignments, resulting in incorrect weighting of memory features. Consequently, portions of the mask may be incorrectly assigned or blended, reducing segmentation accuracy and causing visual inconsistencies, as shown in Fig. 4 for Pear and Apple objects.[1]

These artifacts are particularly problematic in complex scenes where precise segmentation is critical. To mitigate these limitations, further research is needed to enhance the feature differentiation capabilities of XMem++. Potential approaches include improving the feature extraction process to generate more distinct feature vectors, incorporating advanced similarity measures to better handle subtle differences between features, or refining the memory update mechanism to selectively store and retrieve more discriminative features. Addressing this limitation is crucial for improving the reliability and accuracy of segmentation in complex video sequences.

Limitations. Similar to the baselines, several fundamental limitations of our framework are identified, which are pivotal for comprehending our findings and guiding future research directions.

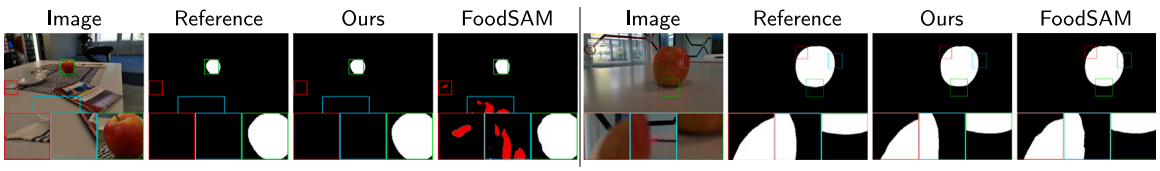

Fig. 4. Comparison between original images, ground truth masks, masks generated by FoodMem, and masks generated by FoodSAM using Vegetables & Fruits dataset. Our method demonstrates robust performance under blurry (caused by fast speed capturing) and far-near conditions.[1].

Table 1
Comparison of mAP and recall scores achieved by different models on two datasets: MetaFood3D, Nutrition5k, and V&F. The models evaluated include FoodSAM, DEVA, kMean++, and our framework. We also apply our method to each image individually for fairness*. All results are reproduced (best results in red, second-best in orange, and third-best in yellow) Time complexity, GPU memory usage (averaged over 67 images), and parameters for each evaluation model[a]. †This dataset forms part of the FoodMask collection, in which we manually annotate the ground truth masks.

Methods	mAP↑			Recall↑			Model(s) Size	Complexity	
	N5k†	V&F†	MTF3D	N5k†	V&F†	MTF3D		Speed↓	Memory↓
FoodSAM	0.9192	0.8914	0.9348	0.7752	0.9441	0.9973	636M	22 m33 s	12.5G
DEVA	0.8825	0.8548	0.2734	0.7301	0.9328	0.9511	241M	25 s	9.0G
kMean++	0.4232	0.4361	0.5572	0.6467	0.9245	0.9718	-	15 s	-
SeTR*	0.9170	0.8306	0.8946	0.9788	0.9906	0.9524	723M	14 m51 s	6.25G
Ours	0.9098	0.9499	0.9453	0.7708	0.9469	0.9941	785M	25s	6.25G

[a] More details in the supplementary material.

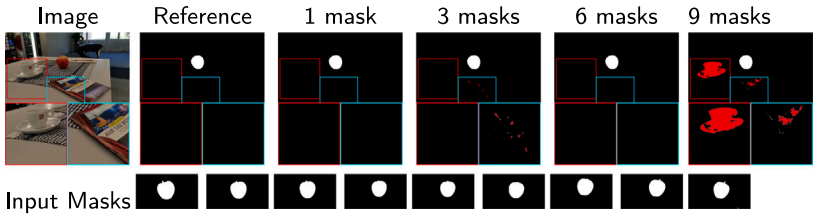

Fig. 5. Comparison of original images, ground truth masks, and masks generated by SeTR with configurations of 1, 3, 6, and 9 masks using Vegetables & Fruits dataset. We show the input masks for each scene.

Table 2
Comparison of recall and mAP scores obtained by various FoodMem configurations on two datasets, Nutrition5k, MetaFood3D, and V&F. The configurations include settings with 1 mask, 3 masks, 6 masks, and 9 masks. All results are reproduced (best results in red, second-best in orange, and third-best in yellow). †This dataset forms part of the FoodMask collection, in which we manually annotate the ground truth masks.

Masks	mAP↑			Recall↑		
	N5k†	V&F†	MTF3D	N5k†	V&F†	MTF3D
1 mask	0.9098	0.9499	0.9453	0.7708	0.9469	0.9941
3 masks	0.9025	0.9027	0.9781	0.7688	0.9419	0.9953
6 masks	0.9005	0.9124	0.9282	0.7663	0.9438	0.9923
9 masks	0.9082	0.9050	0.9213	0.7690	0.9430	0.9916

1. Leveraging XMem++ introduces inherited constraints related to its algorithmic design and computational requirements, which could impact scalability and efficiency when processing large datasets or high-resolution (e.g., 4k) videos.
2. Our framework exhibits sensitivity to lighting conditions, particularly in low-light environments or scenarios with significant shadows. This issue is most pronounced in the first frames processed by SeTR, where it can pose challenges in accurately detecting and segmenting food objects before the memory-based mechanism stabilizes the segmentation across subsequent frames.
3. Our framework can generate artifacts when processing multiple clean masks with similar features, resulting in ambiguous attention weights and decreased accuracy in scenes with visually similar food items, as shown in our Ablation study.

Addressing these limitations will enhance the framework's reliability and applicability across varied real-world conditions and datasets in future research endeavors.[1]

5. Conclusions and future work

Our proposed framework represents a significant advancement in video semantic segmentation of food items. We have introduced a novel architecture that effectively leverages the strengths of SeTR and memory-based models, leading to state-of-the-art performance in food video segmentation. Additionally, we have contributed a comprehensive dataset tailored for this task, providing a valuable resource for future research. Our experimental results demonstrate the framework's superior accuracy and efficiency, making it well-suited for practical food recognition and analysis applications. In future work, we aim to explore advanced preprocessing and feature extraction techniques and feature differentiation strategies to mitigate our limitations and improve the robustness of our framework in such scenarios.

CRediT authorship contribution statement

Ahmad AlMughrabi: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Adrián Galán:** Visualization, Validation, Methodology, Investigation. **Ricardo Marques:** Writing – review & editing, Supervision, Resources, Methodology, Formal analysis, Conceptualization. **Petia Radeva:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Ahmad AlMughrabi reports financial support was provided by University of Barcelona. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially funded by the EU project MUSAE (No. 01070421), 2021- SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220- VET0000894 34, Erasmus+ EU), DeepSense (ACE053/22/000029, ACCIÓ), and Grants PID2022141566NB-I00 (IDEATE), PDC2022-133642-I00 (DeepFoodVol), and CNS2022-135480 (A-BMC) funded by MICIU/AEI/ 10.13039/501100 011033, by FEDER (UE), and by European Union NextGenerationEU/ PRTR. A. AlMughrabi acknowledges the support of FPI Becas, MICINN, Spain.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patrec.2025.03.014.

Data availability

The data that support the findings of this study are openly available in Figshare at http://doi.org/10.6084/m9.figshare.28638806.

References

[1] H. Seong, J. Hyun, E. Kim, Kernelized memory network for video object segmentation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, Springer, 2020, pp. 629–645.
[2] V.K. Sharma, R.N. Mir, SSFNET-VOS: Semantic segmentation and fusion network for video object segmentation, Pattern Recognit. Lett. 140 (2020) 49–58.
[3] T. Zhou, F. Porikli, D.J. Crandall, L. Van Gool, W. Wang, A survey on deep learning technique for video segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (6) (2022) 7099–7122.
[4] M. Bekuzarov, A. Bermudez, J.-Y. Lee, H. Li, Xmem++: Production-level video segmentation from few annotated frames, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 635–644.
[5] J. Mei, Y. Yang, M. Wang, Z. Li, J. Ra, Y. Liu, LiDAR video object segmentation with dynamic kernel refinement, Pattern Recognit. Lett. 178 (2024) 21–27.
[6] H.K. Cheng, A.G. Schwing, Xmem: Long-term video object segmentation with an Atkinson-Shiffrin memory model, in: European Conference on Computer Vision, Springer, 2022, pp. 640–658.
[7] Q. Gao, W. Zhong, J. Li, T. Zhao, CLUE: Contrastive language-guided learning for referring video object segmentation, Pattern Recognit. Lett. 178 (2024) 115–121.
[8] S.W. Oh, J.-Y. Lee, N. Xu, S.J. Kim, Video object segmentation using space-time memory networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9226–9235.
[9] Z. Yang, Y. Yang, Decoupling features in hierarchical propagation for video object segmentation, Adv. Neural Inf. Process. Syst. 35 (2022) 36324–36336.
[10] X. Lan, J. Lyu, H. Jiang, K. Dong, Z. Niu, Y. Zhang, J. Xue, Foodsam: Any food segmentation, IEEE Trans. Multimed. (2023).
[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
[12] J. Steinbrener, V. Dimitrievska, F. Pittino, F. Starmans, R. Waldner, J. Holzbauer, T. Arnold, Learning metric volume estimation of fruits and vegetables from short monocular video sequences, Heliyon 9 (4) (2023).
[13] Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand, J. Sim, Nutrition5k: Towards automatic nutritional understanding of generic food, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8903–8911.
[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
[15] X. Zhou, V. Koltun, P. Krähenbühl, Simple multi-dataset detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7571–7580.
[16] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.
[17] X. Wu, X. Fu, Y. Liu, E.-P. Lim, S.C. Hoi, Q. Sun, A large-scale benchmark for food image segmentation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 506–515.
[18] K. Okamoto, K. Yanai, UEC-FoodPIX complete: A large-scale food image segmentation dataset, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V, Springer, 2021, pp. 647–659.
[19] Y.A. Sari, A. Gofuku, Measuring food volume from RGB-depth image with point cloud conversion method using geometrical approach and robust ellipsoid fitting algorithm, J. Food Eng. 358 (2023) 111656.
[20] H.K. Cheng, S.W. Oh, B. Price, A. Schwing, J.-Y. Lee, Tracking anything with decoupled video segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 1316–1326.
[21] T. Jain, C. Lennan, Z. John, D. Tran, Imagededup, 2019, https://github.com/idealo/imagededup.
[22] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, LabelMe: a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (2008) 157–173.
[23] Y. Chen, J. He, C. Czarnecki, G. Vinod, T.I. Mahmud, S. Raghavan, J. Ma, D. Mao, S. Nair, P. Xi, A. Wong, E. Delp, F. Zhu, MetaFood3D: Large 3D food object dataset with nutrition values, 2024, arXiv preprint arXiv:2409.01966.
[24] R. Padilla, S.L. Netto, E.A. Da Silva, A survey on performance metrics for object-detection algorithms, in: 2020 International Conference on Systems, Signals and Image Processing, IWSSIP, IEEE, 2020, pp. 237–242.
[25] K. Oksuz, B.C. Cam, E. Akbas, S. Kalkan, Localization recall precision (LRP): A new performance metric for object detection, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 504–519.