Deep learning-based rotoscoping: a systematic review of methods and applications

Deniz Yuce
Visual Effect Artist and Supervisor at Artworkslab, TX, USA
E-mail: *dnzyc87@gmail.com*

Received 6 May 2026; accepted 18 May 2026; published online 5 June 2026
DOI https://doi.org/10.21595/jmai.2026.26653

Abstract. This study systematically examines deep learning-based rotoscoping systems along the axes of video object and instance segmentation, memory-based models, optical flow integration, foundation and transformer-based video approaches, as well as matting and production-oriented evaluation metrics. The aim is to classify contemporary rotoscoping methods within a holistic framework, reveal their strengths and limitations under production conditions, and propose a production-oriented evaluation perspective for future research. Methodologically, prominent post 2015 approaches, including VOS/VSS models, memory-based and optical flow-based methods, prompt-driven foundation segmentation models, and transformer-based video systems, are analyzed comparatively with respect to datasets, training strategies, evaluation metrics, and application scenarios. The findings indicate that rotoscoping has evolved from frame-by-frame manual tools toward human-supervised hybrid automation systems based on memory-augmented video segmentation, optical flow-assisted propagation, prompt-based foundation models, and transformer-based video approaches. However, domain gap issues, computational costs in high-resolution sequences, limitations in fine-detail preservation and matting consistency, long-term temporal stability, and the lack of production-specific evaluation metrics remain significant challenges, rendering fully automated rotoscoping an unresolved problem under real-world production conditions. The study suggests that rotoscoping workflows will become highly automated in the near future, yet quality assurance and creative decision-making will continue to rely on human experts within human-in-the-loop hybrid architectures. Accordingly, future research should prioritize standardized evaluation protocols, methods tailored to high-resolution and long-duration video sequences, and hybrid system designs.

Keywords: video object segmentation, rotoscoping, video matting; video instance segmentation, optical flow, human-in-the-loop annotation, memory-based video segmentation, foundation model segmentation.

1. Introduction

Rotoscoping originated in the early 20th century as a technique for generating motion imagery through frame-by-frame drawn masks in animation production, and it has since evolved from physical manipulation of film strips to digital spline-based workflows [1]. Manual rotoscoping requires substantial time and labor for each object and each frame, creating a major bottleneck in the processing of large-scale video datasets and in real-time applications [1], [2]. Methods developed during the classical computer vision era, particularly those based on active contours and graph cuts, provided the first semi-automatic solutions for rotoscoping; however, they still required intensive user intervention in the presence of complex camera motion, long-term occlusions, and fine structural details [3].

Deep learning-based segmentation models (e.g. FCN, U-Net, DeepLab, Mask R-CNN) have largely automated per-frame mask generation by enabling high-accuracy semantic and instance segmentation on individual images [4], [5]. However, since these models process frames largely independently, they have remained limited regarding issues of temporal consistency (temporal drift, flickering) and re-identification (ReID) following object occlusions [6], [7].

In recent years, with the rise of Deep Convolutional Neural Networks (DCNNs) and, in particular, Fully Convolutional Networks (FCNs), rotoscoping technology has undergone a transformative shift [1], [8].Whereas traditional rotoscoping methods relied on hand-crafted features and extensive manual labor, deep learning-based algorithms now offer substantially higher accuracy and efficiency than systems built on hand-crafted representations [1], [6], [8]. Graph Convolutional Networks (GCN) is a current artificial neural network research topic. The GCN model is derived from graph theory and convolution theorems to apply machine learning to data represented by graphs [9]. Studies on GCN in the literature are generally based on object recognition and classification [10].

Video Object Segmentation (VOS) refers to the process of identifying target objects with specific properties in a video scene and partitioning video frames into multiple segments or objects [1], [6]. This technology plays a critical role across a wide spectrum of applications, ranging from enhancing visual effects in the film industry to scene understanding for autonomous driving, robotics applications, and virtual background generation in video conferencing systems [1], [11].

In the literature, depending on how outputs are defined, two core categories are typically distinguished: Video Object Segmentation (VOS), which separates foreground objects independently of their semantic category, and Video Semantic Segmentation (VSS), which assigns a semantic label to every pixel [6], [12]. These two tasks intersect in approaches that integrate mechanisms such as optical flow and memory-based models to promote temporal consistency and to move toward architectures capable of modeling long-term dependencies [6], [12].

Recent works on VOS/VSS has increasingly focused on memory-based architectures, such as Space-Time Memory (STM) networks, which leverage contextual information from past frames to enhance temporal coherence, achieving notable performance gains on benchmark datasets such as DAVIS and YouTube-VOS [6], [13], [14]. Optical flow-based mask propagation and "joint segmentation + flow" approaches (e.g., SegFlow type models) further improve performance by exploiting the bidirectional exchange of information between segmentation and flow estimation tasks [15], [16]. However, the direct applicability of these core models to long videos can be constrained by memory management, computational cost, and scalability, which has motivated research into concepts such as regional memory (e.g., RMNet) and regionally focused, flow-based methods that explicitly target the trade-off between cost and efficiency [13], [17]. The adoption of foundation models (e.g., SAM/SAM 2, Track Anything) further broadens the possibilities for interactive and real-time segmentation, reducing the degree of user intervention required in rotoscoping workflows and accelerating prototyping [12], [14], [18].

In recent days, generating temporally consistent, high-detail mask sets for target objects across an entire video sequence remains one of the most labor-intensive components of modern visual effects (VFX) pipelines [6], [11]. Despite significant technological advances and the growing body of published research, challenges such as high-resolution processing, object occlusions in complex scenes, and maintaining long-term temporal consistency persist [1], [6], [8].

This study systematically addresses the rotoscoping problem within the broader ecosystem of VOS/VSS, including memory-based and transformer-based video models, foundation approaches such as SAM/SAM 2/Track Anything, optical flow integration, matting, and production-oriented metrics. In this regard, it aims to classify existing methods, discuss their advantages and limitations, and provide an evaluation framework to guide future research. In this review article, deep learning-based rotoscoping methods are examined using a systematic methodology, and the employed techniques and current applications are presented from an academic perspective.

The overall structure of this survey is illustrated in Fig. 1. It outlines the progression from methodological foundations (Section 3) to applications (Section 4), followed by the discussion and conclusion sections in the subsequent parts of the paper.

2. Methodology

This review article is grounded in a systematic literature search conducted across major

academic databases, including Google Scholar, arXiv, the Computer Vision Foundation (CVF) open-access archive, IEEE Xplore, and the ACM Digital Library. The search strategy employed domain-specific keywords such as video object segmentation, rotoscoping, video matting, video instance segmentation, temporal consistency, optical flow segmentation, foundation model segmentation, memory-based video segmentation, human-in-the-loop annotation, and synthetic data domain adaptation. All queries were issued in English, and field-specific technical terminology was explicitly incorporated into the search expressions to maximize recall within the target domain.

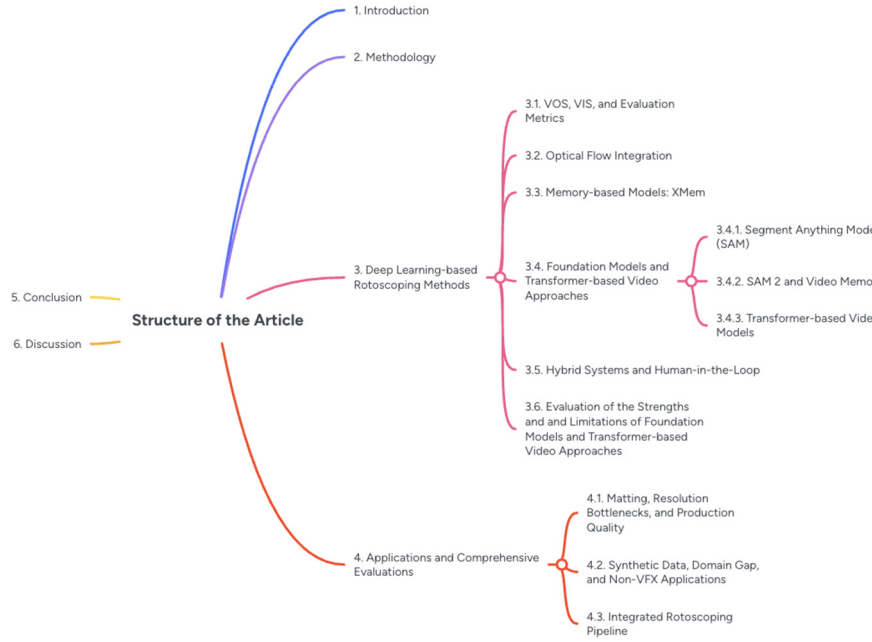

Fig. 1. Overall structure of this survey on deep learning-based rotoscoping

The temporal scope of the survey spans approximately 2015-2026, corresponding to the period in which deep learning-based segmentation models became the dominant paradigm in computer vision. The lower bound of 2015 is motivated by the introduction of Fully Convolutional Networks, which marked the beginning of the deep learning era in semantic segmentation and directly influenced subsequent work on video segmentation [5]. Earlier approaches are considered only as contextual background and are not included in the core systematic analysis.

Inclusion was restricted to studies that proposed deep learning-based methods, reported quantitative performance on established public benchmarks (e.g., DAVIS, YouTube-VOS, MOSE), and explicitly addressed problems within the rotoscoping domain [12], [19]. Studies without a deep learning component, purely theoretical analyses lacking experimental validation, proprietary industrial tools, and work focusing primarily on 3D or volumetric segmentation were excluded from the review.

The selected works were categorized according to their architectural characteristics (memory-based, transformer-based, optical-flow integrated, foundation model–based), the challenges they target (temporal consistency, occlusion handling, matting quality, domain adaptation), and the evaluation metrics they employ (e.g., J&F, LPIPS, boundary-focused measures). This structured classification framework serves two main purposes: it ensures systematic coverage of the literature and provides a principled basis for identifying research gaps and open problems, which are discussed in detail in Section 5.

3. Deep learning-based rotoscoping methods

Modern rotoscoping pipelines have evolved from fully manual drawings to automatic and interactive segmentation systems driven by deep learning architectures [1], [20]. In this section, we examine the core approaches and technological components proposed in the literature.

3.1. VOS, VIS, and evaluation metrics

In the literature, the most fundamental and widely adopted taxonomy for video segmentation is the distinction between Video Object Segmentation (VOS) and Video Semantic Segmentation (VSS) [1], [6]. These two branches are clearly differentiated by their objectives for pixel-level scene understanding and by the structure of their output spaces [1].

Video Object Segmentation (VOS): VOS corresponds to the traditional video segmentation setting and focuses on separating the dominant objects in a video into "foreground" and "background," independently of their semantic category [6], [12]. It is commonly used in applications such as editorial compositing, object removal, and virtual background generation, and it does not, by design, require knowledge of whether an object is, for example, a "car" or a "person" [1], [6]. By representing objects with pixel-accurate masks rather than coarse bounding boxes, VOS provides a much richer scene representation than basic object tracking and supports fine-grained control over edges and transparency, which is critical for production-quality rotoscoping [1], [21].

From the perspective of user involvement in the segmentation process, VOS is typically organized into three main categories:

Automatic VOS (unsupervised/single-object, AVOS): Automatically detects and segments foreground objects without any manual user input, usually by exploiting motion saliency to isolate the most prominent moving object in the scene [1], [11].

Semi-supervised VOS (SVOS): SVOS typically propagates an object mask (or bounding box) specified in the first frame to track the target throughout the video via mask propagation [1], [22]. In short, the model follows the target object across the sequence conditioned on the masks provided in the initial frames [11], [23].

Interactive VOS (IVOS). Incorporates user input such as scribbles, clicks, or bounding boxes in an iterative loop and refines the segmentation results interactively [2], [20]. Referring expression-based and language-guided interfaces can be seen as special cases in which the target object is specified via text or high level user intent [24].

Video Semantic Segmentation (VSS): VSS can be viewed as the temporal extension of image semantic segmentation, aiming to assign a predefined semantic category label to every pixel across all frames [1], [6], [12]. Tasks such as Video Instance Segmentation (VIS) and Video Panoptic Segmentation (VPS) are often treated as specialized subproblems under this broader semantic umbrella [1]. VPS unifies semantic and instance-level labeling in a single output space by jointly assigning semantic classes and consistent instance identities throughout the video, thus providing a comprehensive scene representation that combines "stuff" and "thing" categories [25]. VIS, in contrast, is tailored toward detecting, segmenting, and tracking individual object instances over time, with an explicit focus on instance identities and category labels for each object [1], [25].

VIS requires simultaneously segmenting objects, tracking their identities across frames, and assigning a semantic class label to each instance [19], [26], [27]. The key difference is that VIS is inherently category-aware, whereas VOS is typically category-agnostic and primarily driven by appearance and motion cues [1], [19]. VOS is thus formulated as a segmentation problem that relies heavily on motion and appearance consistency and must maintain coherent masks under deformations, occlusions, and challenging phenomena such as mirror reflections, all of which are critical failure modes in production rotoscoping [6], [19]. VSS, on the other hand, targets dense classification of all pixels in the scene; to maintain temporal coherence across frames, it often integrates mechanisms such as optical flow and memory-based architectures that explicitly model

temporal dependencies [6], [18].

As illustrated in Fig. 2, in the semi-supervised VOS (SVOS) case, the target athlete is specified by a user-provided mask in the first frame and then tracked across subsequent frames, whereas in automatic VOS (UVOS) the dominant moving object in the scene is segmented as foreground without any user input.

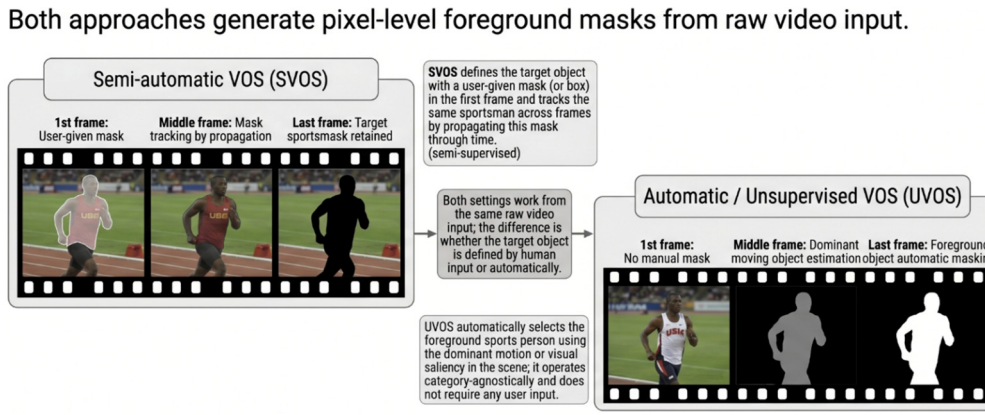

Fig. 2. Schematic illustration of two common VOS settings

Standard metrics for evaluating video segmentation performance have been defined by the DAVIS (Densely Annotated Video Segmentation) benchmark [22], [28], [29]. The Jaccard index (J), which measures region similarity, computes the intersection over union (IoU) between the predicted mask and the ground truth mask to determine what proportion of pixels are correctly classified [1], [11], [19]. The F-measure (F), which represents boundary accuracy, evaluates the local precision and recall of mask contours via a bipartite graph matching between predicted and ground-truth boundaries [6], [11], [19]. The arithmetic mean of these two scores, J&F, has been adopted as a unified ranking criterion that reflects the overall performance of a system [6], [19], [22], [29].

Different datasets focus on measuring specific capabilities of video segmentation models. For example, the YouTube-VOS dataset partitions object categories into “seen” and “unseen” classes to test the generalization ability of models beyond the categories encountered during training [6], [30]. The MOSE dataset, which contains more modern and complex scenes, stresses long-term re-identification and tracking performance in scenarios where objects undergo heavy occlusions, completely disappear and reappear, or exhibit significant changes in appearance [19].

From the perspective of professional rotoscoping and visual effects (VFX) applications, J&F scores alone do not constitute a sufficient indicator of quality [11]. For rotoscoping, not only pixel-wise accuracy but also the temporal consistency of masks across frames is critical, and even a model with high J&F scores cannot be used in production if it exhibits frame-to-frame flicker or edge chatter in successive frames [11]. For this reason, the Temporal Stability (T) metric has emerged as an important measure, as it computes the cost between consecutive segmentation boundaries and penalizes low-quality flickering while compensating for motion and deformation [1], [6], [11]. In addition, in order to better reflect the quality of the final composite, perceptual metrics that measure similarity in deep feature space, as well as temporal perceptual loss terms that capture temporal consistency, are increasingly incorporated into evaluation protocols [31].

Within VFX pipelines, production-oriented measures, rather than academic metrics alone, ultimately determine practical efficiency [2], [22]. These include parameters such as the frequency of flicker, the amount of manual correction time required per frame, the number of iterations needed for client or supervisor approval (iteration count per shot), and the artist disagreement rate, i.e., the inconsistency between masks produced by different artists on the same shot [20], [22]. This situation indicates that, when conducting academic evaluations, models should be assessed

not only in terms of pixel accuracy but also in terms of their ability to meet production requirements related to edge chatter, validation workflows, and iteration overhead [11], [32].

3.2. Optical flow integration

Optical flow estimates pixel-level motion between consecutive video frames and is one of the core carriers of temporal consistency in video sequences [6], [15]. In modern VOS and rotoscoping workflows, optical flow is heavily used in key components such as mask propagation, temporal supervision, and joint learning of segmentation and flow [6].

In the literature, segmentation and flow coupling (segmentation + flow coupling) is treated as a complementary formulation [15]. The SegFlow architecture, for example, introduces a two-branch model that jointly learns segmentation and optical flow in an end-to-end fashion, enabling bidirectional information exchange between the two tasks [15]. Multi-branch architectures such as SegFlow reinforce both flow and segmentation, thereby improving consistency over long video sequences [12], [16]. In this setting, the segmentation output guides optical flow to be sharper and more coherent along object boundaries, while the flow field in turn helps the segmentation produce smoother transitions across frames [15], [33].

Consequently, this relationship is reinforced through warp-based supervision strategies, which make it possible to align masks or feature maps from the previous frame to the current frame via optical flow vectors [1], [30], [34]. Early deep learning-based systems such as MaskTrack refined the previous-frame mask with optical flow and transferred it to the current frame [6]. Similarly, methods such as ObjectFlow (OFL) have aimed to minimize error by iteratively updating the object segmentation and flow models within a unified scheme.openaccess [33]. Flow information is essential for maintaining stable mask boundaries, particularly in scenes that exhibit fast motion and strong deformation, and this flow-based propagation provides a critical control mechanism for preventing error accumulation and temporal drift over time [1], [23].

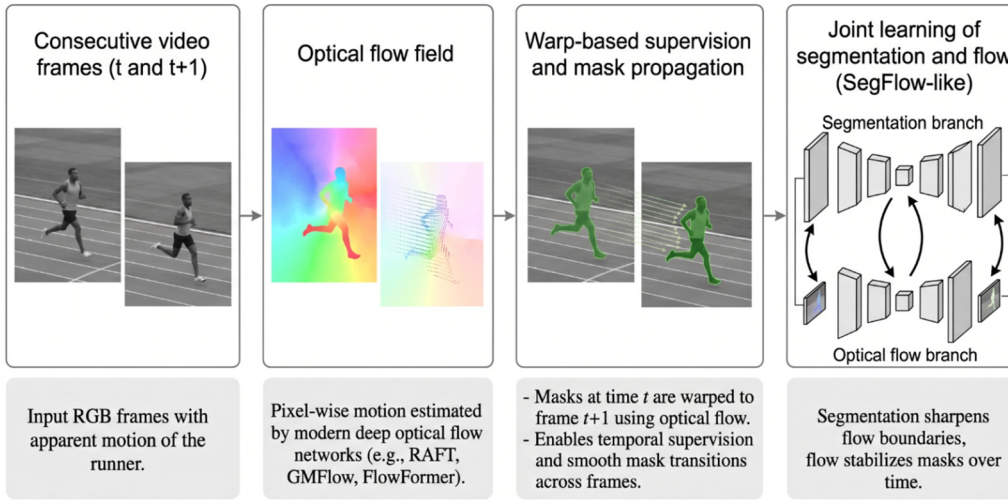

Fig. 3. Conceptual illustration of optical-flow-guided video object segmentation

As depicted in Fig. 3, optical flow provides pixel-wise motion fields that support warp-based supervision and mask propagation, while joint models (e.g., SegFlow) learn segmentation and flow in a unified framework to improve temporal consistency.

In contemporary rotoscoping pipelines, early deep optical flow methods (such as FlowNet) have been largely superseded by more advanced deep flow architectures [15]. This paradigm shift has been accelerated by RAFT, which models all-pairs correlations between pixels and employs iterative updates to refine the flow field [35]. More recently, models such as GMFlow, with its

global matching formulation, and FlowFormer, which adapts Transformer architectures to optical flow estimation, have established new standards for capturing complex motion and fine-grained detail [36]. The high-accuracy flow fields produced by these models enable rotoscoping systems to deliver professional-grade results with significantly reduced edge chatter [30], [35].

3.3. Memory-based models: XMem

In video object segmentation and rotoscoping pipelines, the use of feature memory is essential for propagating information from the first frame or from user interactions to subsequent frames [37]. However, conventional methods such as Space-Time Memory (STM) require an ever growing memory bank as the video length increases, which leads to memory explosion and performance degradation on consumer grade hardware [37]. To achieve high accuracy on long videos without incurring memory explosion, the XMem architecture introduces a multi-memory framework inspired by the Atkinson–Shiffrin model of human memory [37].

The XMem architecture consists of three independent memory components:

– Sensory Memory: This memory module, updated at every frame via a Gated Recurrent Unit (GRU), stores low level object localization cues and enforces temporal smoothness [37], [38]. However, sensory memory is susceptible to representation drift when used for long-term predictions [37]. As a short-lived, rapidly updated store for fast-changing high-level visual information, it captures instantaneous appearance and fast motion cues, providing the core representations for short-term decisions. XMem uses this memory pool to enable fast access and rapid updating of semantic and appearance representations [38].

– Working Memory: This memory aggregates high-resolution features from a subset of frames [37], [38]. It acts as a buffer, providing high-precision object tracking [37]. The working memory is designed to retain the desired level of detail while reusing contextual information derived from previous frames [38].

– Long-term Memory: This memory enhances robustness in ReID-like tasks on recurring object appearances and strengthens consistency over long video sequences [38]. Through a potential memory potentiation mechanism between memories, the most frequently used features are transferred to this store as compact prototypes when the working memory becomes saturated, thereby preserving visual information over the long term while keeping the overall memory footprint under control [37], [38].

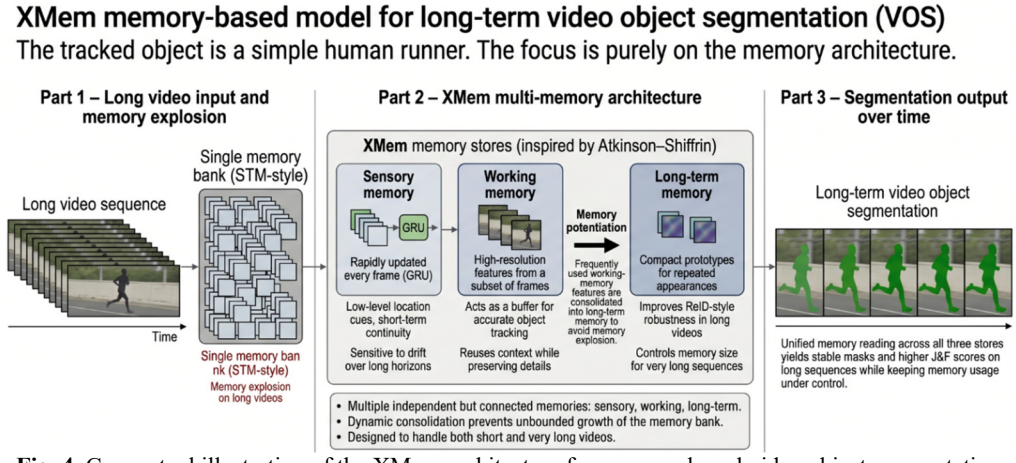

Fig. 4. Conceptual illustration of the XMem architecture for memory-based video object segmentation

Inspired by the Atkinson–Shiffrin model of human memory, XMem maintains sensory, working, and long-term feature memories, along with a memory potentiation mechanism, to mitigate memory explosion and preserve temporal consistency in long video sequences.

In XMem's memory reading mechanism, the integrated use of these three stores (sensory, working, and long-term) is designed to deliver strong performance not only on long videos but also on short clips, thereby balancing fast responsiveness for short sequences with sustained consistency over extended ones [38]. This design aims to maintain temporal coherence even on very long sequences while keeping memory consumption under control and has been shown to achieve higher segmentation quality (J&F scores) than prior memory-based methods [37], [38].

3.4. Foundation models and transformer-based video approaches

This section examines recent foundation models and transformer-based video approaches within the broader landscape of deep learning-based rotoscoping methods. Foundation models trained on large-scale datasets have established a new standard for modern VFX workflows, owing to their zero-shot capabilities that allow successful adaptation even to previously unseen object categories [22], [27].

3.4.1. Segment anything model (SAM)

The Segment Anything Model (SAM) is defined as a prompt-based foundation segmentation model that formulates image segmentation as a promptable task built on large-scale visual representation learning [24], [27]. Composed of a powerful image encoder, a flexible prompt encoder, and a lightweight mask decoder capable of real-time inference, SAM can segment arbitrary objects by responding instantaneously to user inputs such as points, boxes, or free-form strokes [24].

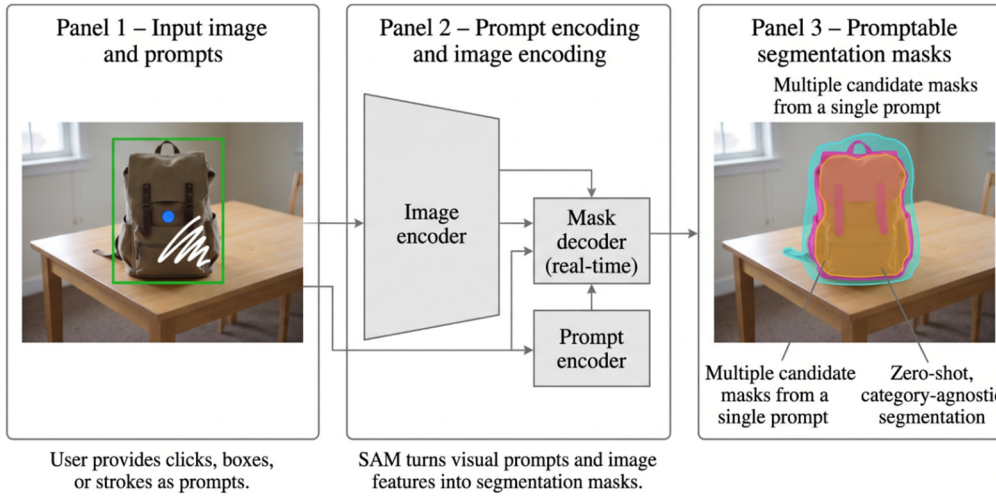

Fig. 5. Conceptual illustration of the segment anything model (SAM) as a promptable image segmentation foundation model

As illustrated in Fig. 5, SAM employs an image encoder, a prompt encoder, and a lightweight mask decoder to generate object masks from simple user prompts.

SAM's primary objective is to generate multiple valid masks from a single prompt and thereby provide a general-purpose segmentation engine with broad coverage [24]. SAM consists of three main modules that enable flexibility and interactive use: an image encoder that produces visual representations of the image, a prompt encoder that converts user prompts into numerical embeddings, and a mask decoder that produces mask predictions [24]. In the context of rotoscoping, SAM substantially reduces manual workload for generating the initial mask; however, enforcing temporal continuity across successive video frames introduces challenges such as mask erosion and issues related to temporal propagation [21], [39].

3.4.2. SAM 2 and video memory

Segment Anything Model 2 (SAM 2) formulates video segmentation as a temporally continuous Promptable Visual Segmentation (PVS) task [22]. SAM 2 is designed as an extension of SAM that brings its three core building blocks, namely the image encoder, the prompt encoder, and the mask decoder, into the video domain [17], [40]. It aims to generate a sequence of masks for a target object across an entire video by using inputs such as clicks, bounding boxes, or masks provided on any frame of the sequence [22].

The key innovation of SAM 2 is a streaming memory module that supports temporally consistent segmentation and object tracking throughout a video [17], [22]. This memory stores information from past frames in a memory bank organized as a set of FIFO (First-In-First-Out) queues and consults these stored features via a memory attention mechanism while processing the current frame [22]. This memory-based design enables SAM 2 to track an object throughout the entire video and to propagate user corrections from a single interaction across the full sequence [22]. It is explicitly designed to maintain prompt-based interaction over time and to enable object isolation in mosaicked or heavily edited scenes [17], [21].

In conventional approaches, a significant error often requires restarting the segmentation from scratch, whereas the memory in SAM 2 allows a new click to update the object context without discarding prior information [22]. However, in scenes with heavy occlusions, long sequences, and multiple visually similar objects, SAM 2 can still fail due to identity switching and memory capacity limitations [6], [14]. Considering challenges such as multi mask management, memory cost, and real-time performance, these issues may be mitigated through efficient memory module design, improved prompt denoising strategies, and the integration of domain adapted memory mechanisms [41]. In professional rotoscoping workflows and large-scale video dataset production, SAM 2 stands out in terms of video memory, interactivity, and continuity [22].

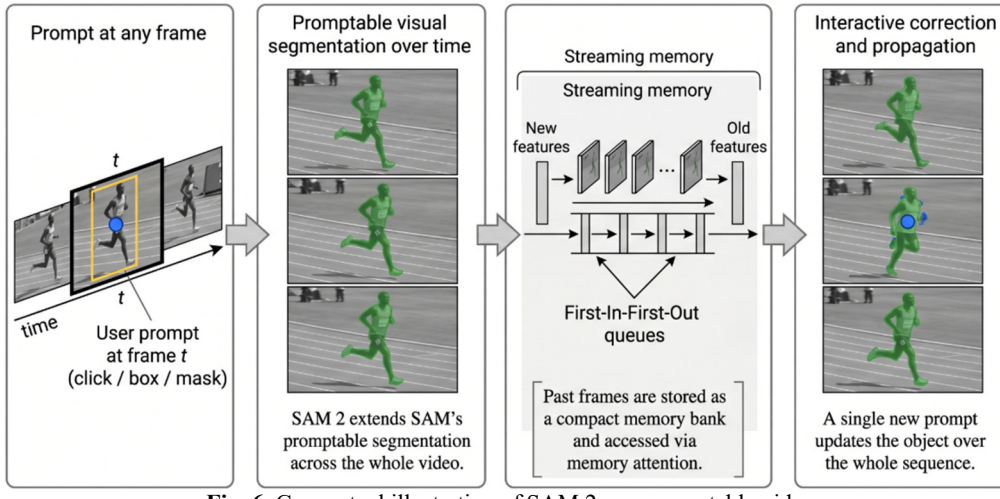

Fig. 6. Conceptual illustration of SAM 2 as a promptable video segmentation model with streaming memory

A user-provided prompt at an arbitrary frame is propagated to generate segmentation masks across the entire video sequence, while a FIFO-based memory bank and memory attention mechanisms support long-term tracking and facilitate interactive refinement.

3.4.3. Transformer-based video models

In recent years, a shift has been observed from memory bank-based foundation models toward

transformer-based architectures that model objects holistically across video sequences [1], [27], [42]. The transformer constitutes the underlying technical framework that defines how such models are architected [43], [44].

Models such as Vision Transformer (ViT), Swin Transformer, and segmentation-oriented approaches including Mask2Former and SeqFormer are built upon transformer architectures [17], [45]. Methods such as Mask2Former, when combined with ViT- or Swin Transformer–based backbones, are capable of addressing panoptic, instance, and semantic segmentation tasks within a unified mask representation, both at the intra-frame and inter-frame levels [43].

Video-oriented transformer models, including DEVA and IDOL, employ space-time attention mechanisms to capture broader contextual information compared to the capacity-limited memory banks used in earlier approaches, while also providing advantages in maintaining identity consistency over long sequences [27], [46]. In contrast, methods based on explicit key-value memory, such as STM and XMem, store past frames as feature banks and establish direct pixel-level correspondences, thereby achieving high segmentation accuracy [27], [37].

Transformer-based models aim to address challenges such as identity consistency and occlusion in long video sequences, while producing high-quality segmentation outputs [1], [27].

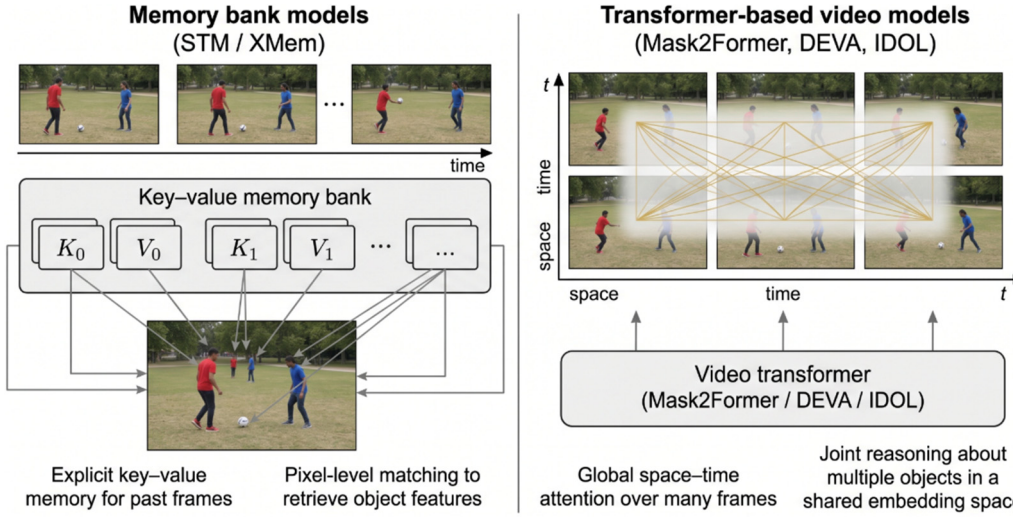

Fig. 7. Memory bank models store explicit features for selected frames

Memory bank models store explicit feature representations extracted from selected frames, enabling efficient retrieval of temporal information. In contrast, transformer-based models employ spatiotemporal attention mechanisms to reason over long video sequences and multiple objects; however, their effectiveness is often constrained by computational complexity as sequence length increases.

3.5. Hybrid systems and human-in-the-loop

Hybrid systems enable users to generate high-quality masks across entire video sequences by providing simple interactions (clicks, scribbles, bounding boxes) on selected key frames only [1], [20]. The Track Anything Model (TAM) presents a hybrid framework that combines SAM's robust image segmentation capability with XMem's long-term video memory, aiming to minimize human correction requirements for interactive object tracking in professional rotoscoping workflows [2], [21].

In such systems, user interaction is structured as a refinement loop that activates when necessary: when the model fails or loses track of the object in certain frames, the artist pauses the

process and provides additional positive or negative prompts to correct the mask [21]. These refined masks generated by SAM are then incorporated into XMem's temporal memory bank, enhancing the discriminability and segmentation quality of subsequent frames [21]. Thus, the system provides a human-supervised solution that offers high efficiency, balancing between the labor intensity of fully manual rotoscoping and the error margins of fully automatic methods [21], [20]. This approach enables state-of-the-art results with minimal human intervention, even in complex scenes [21], [37]. Contemporary hybrid systems have transformed the segmentation task from simple frame-based tracking into a continuous command-response interaction paradigm [22].

From a broader perspective, modern systems elevate the human-in-the-loop paradigm to a strategic level [1], [22]. In interactive systems, approaches such as EVA-VOS models iteratively predict which frame should be annotated and which interaction type (click, box, or full mask) will yield the highest quality improvement at the lowest cost through an agent-based mechanism [2]. Such integrated architectures shift rotoscoping processes from a fully manual task toward an artist-supervisor model that oversees and refines AI-generated suggestions [2], [3]. Validating the accuracy of masks produced by automated tools in rotoscoping workflows through human expertise meets reliability requirements while providing advantages in terms of computational cost and workflow efficiency [40].

In Fig. 8, masks are generated through automatic segmentation, while sparse user interactions correct errors and update the memory, ensuring the continuity of high-quality results with significantly less manual effort compared to fully hand-drawn rotoscoping. Hybrid systems, characterized by the interaction between automation and human expertise in rotoscoping, aim to enhance the efficiency of automatic segmentation, reduce the impact of human error, and establish user-supervised automated systems.

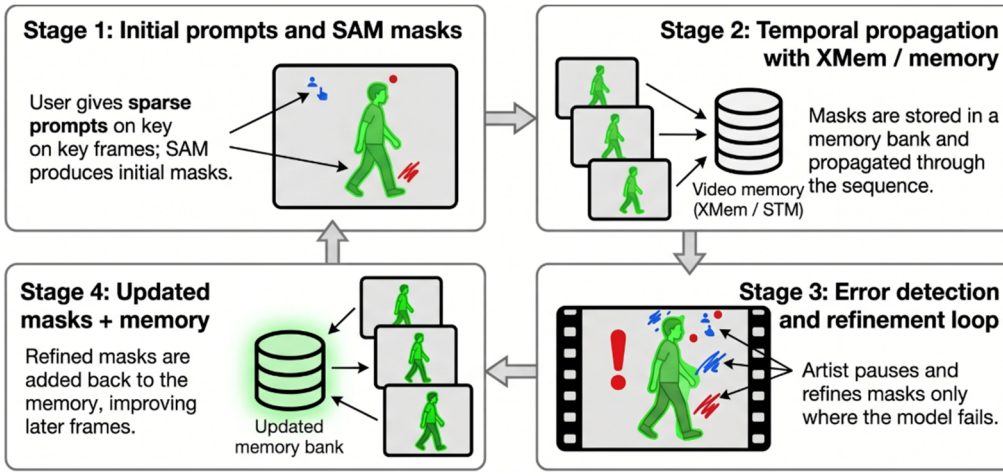

Fig. 8. Conceptual diagram of hybrid human-in-the-loop rotoscoping systems (e.g., Track Anything, EVA-VOS)

3.6. Evaluation of the strengths and limitations of foundation models and transformer-based video approaches

Foundation models and transformer-based video approaches have introduced significant innovations in rotoscoping and video segmentation, yet they also present certain weaknesses and limitations. Foundation models, trained on large-scale datasets, can successfully segment even object categories never encountered during training [1], [24]. However, they may struggle to capture very fine structures such as individual hair strands or thin cables and can introduce artifacts at mask boundaries [22], [47]. Consequently, they occasionally fail to meet the high precision critical for professional rotoscoping [47].

Transformer models produce more efficient outputs in long sequences through space-time attention mechanisms that capture long-range temporal dependencies across video frames [17], [42]. Transformer-based approaches can perform multi-object segmentation as efficiently as single-object segmentation by associating multiple objects within the same high-dimensional embedding space [42]. Nevertheless, due to memory constraints, they exhibit quadratic complexity as the number of pixels and frames increases [46], [49].

In models such as SAM 2, a promptable video segmentation system combined with a human-in-the-loop approach can accelerate data annotation speed by approximately 8.4 fold compared to traditional methods [2], [22]. While hybrid systems enable rapid correction in error cases, they require additional workload and user training [40].

The integration of foundation models and transformer-based approaches into the rotoscoping domain provides artists with significant advantages in terms of speed and automation [40], [48]. However, the need for manual intervention in complex scenes, rapid motions, and frames requiring high precision has not been entirely eliminated, necessitating the development of hybrid and user-controlled workflows [22], [47]. Moreover, the requirement for substantially larger training datasets for these models to learn effectively and integrate into the rotoscoping industry is observed as an academic challenge.

4. Applications and comprehensive evaluations

4.1. Matting, resolution bottlenecks, and production quality

In professional rotoscoping workflows within the visual effects (VFX) industry, the primary objective is not merely to generate a binary mask separating the object from the background, but rather to obtain a high-quality alpha matte that represents transitions and transparency at object boundaries [49], [50]. According to the image compositing equation;

$I = \alpha F + (1 - \alpha)B$, the alpha channel determines the blending ratio between foreground and background pixels [34], [50]. Particularly in scenes containing hair strands, smoke, glass, or motion blur, the hard edges produced by traditional segmentation are insufficient for production quality, rendering alpha matting techniques indispensable [49], [50].

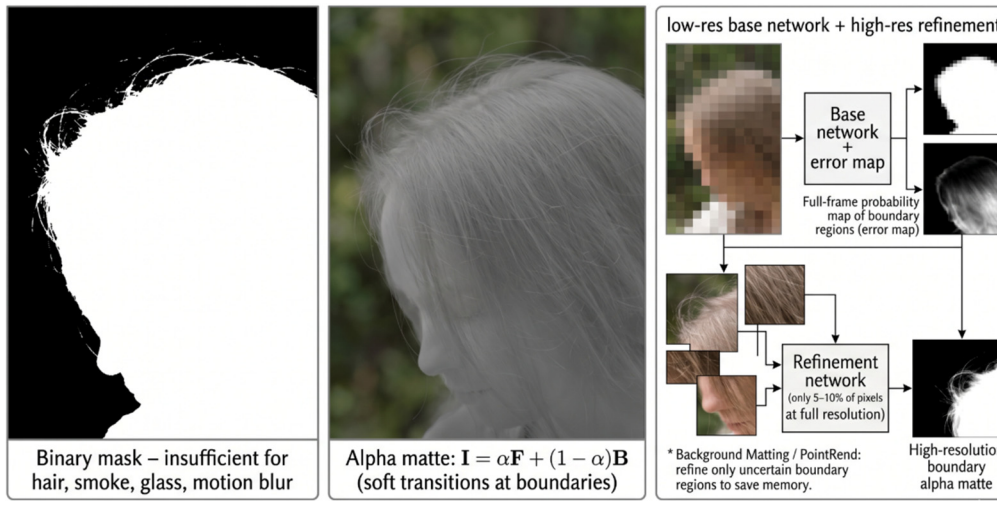

Fig. 9. High-quality alpha mattes are required for production compositing

When applying matting at high resolutions (HD, 4K, and beyond) in the modern visual effects industry, directly running deep neural networks at these resolutions creates a significant resolution bottleneck due to massive GPU memory requirements and low processing throughput [6], [49].

To address these challenges, two fundamental strategies have emerged in the literature: Modern architectures such as Background Matting employ a decoupled structure that, instead of processing the entire image at high resolution, uses a low-resolution base network to generate coarse results and an error prediction map; subsequently, a refinement network processes only selected patches with high error margins (typically 5-10 % of the image area, such as object boundaries) at the original resolution, thereby minimizing memory consumption [49]. The PointRend approach treats segmentation as a rendering problem and optimizes high-resolution details in terms of memory and computational cost by iteratively sampling uncertain points from boundary regions [34].

Modern methods overcome GPU memory limits by predicting coarse results at low resolution and refining only boundary regions at full resolution. In VFX production, the quality of a mask is measured not only by its accuracy in individual frames but also by the temporal consistency it exhibits throughout the video sequence [11], [50]. Models operating on independent frames cause micro-level mask variations (jitter/edge chatter) across consecutive frames, and recurrent structures based on ConvGRU or ConvLSTM have been developed to suppress these temporal fluctuations [34], [50].

4.2. Synthetic data, domain gap, and non-VFX applications

This section examines the use and training of synthetic data in deep learning-based rotoscoping models, the effects of domain gap, and application areas beyond VFX. The costly and time-consuming nature of producing high-quality masks has directed researchers toward synthetic datasets [6], [22]. Synthetic datasets are data collections generated in digital environments using computer algorithms, physics simulators, or game engines (e.g., Blender, GTA V) rather than capturing real-world scenes with cameras [6], [20], [51]. Since manual rotoscoping of objects at the pixel level in real videos requires extremely high labor per object per frame, synthetic datasets automatically generate masks with perfect accuracy from known object positions in computer environments [2], [6], [20]. These datasets are utilized to provide the massive amounts of data required for training deep learning models and their corresponding error-free annotations (masks, optical flow data, etc.) at low cost [39], [20]. Synthetic data offers significant advantages in rotoscoping due to its rapid generation, absence of ethical and copyright issues, and ability to simulate diverse lighting conditions, camera angles, and motion types [20].

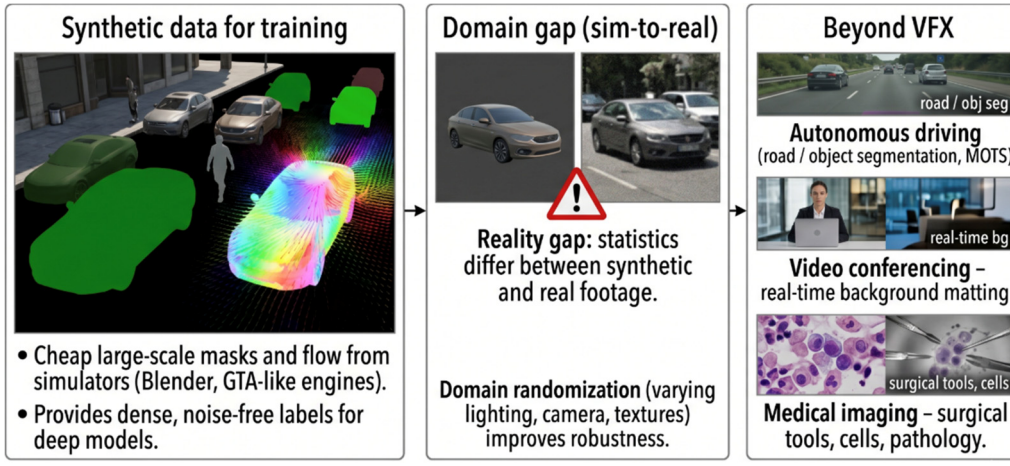

Fig. 10. Synthetic datasets provide cheap dense labels but introduce a sim-to-real domain gap. Rotoscoping-style video segmentation is also used in autonomous driving, telepresence, robotics, and medical imaging

Domain gap refers to the statistical distribution difference between the dataset on which a model is trained (source domain) and the dataset on which it is tested (target domain) [44], [51]. In rotoscoping, synthetic datasets are highly robust in terms of geometric accuracy and flawless annotation quality; however, physical and optical inconsistencies between simulation environments and the real world give rise to the "reality gap" or "sim-to-real domain shift" problem [51]. The "reality gap", particularly prevalent in robotics and simulation studies, occurs when physical modeling deficiencies in simulators (friction, latency, etc.) and low-fidelity sensor data (rendered images) differ from their real-world counterparts [51]. "Sim-to-real domain shift", which represents the transition from synthetic to real-world data, leads to performance degradation on noisy real-world data due to the model's overfitting to the perfection in synthetic data [34], [51]. To mitigate differences stemming from the lack of photorealism in synthetic images and the presence of camera noise and optical artifacts in real production footage, domain randomization techniques are employed; in this approach, camera angles, lighting conditions, object positions, and textures in the simulator are randomly varied to increase the model's robustness against real-world variations [51].

Rotoscoping and video segmentation technologies are utilized across numerous fields beyond the visual effects industry.

1) Object detection, road segmentation, and multi-object tracking and segmentation (MOTS) tasks play critical roles in decision support mechanisms for safe driving systems in autonomous vehicles [1], [30]. Datasets such as KITTI, Cityscapes, and BDD100K dominate research in this domain [1], [52].

2) In robotics, synthetic-to-real transfer methods are leveraged for tasks including object grasping and object localization in complex environments [51].

3) Modern video conferencing applications and remote work tools provide real-time alpha matting solutions for background replacement and privacy preservation [49].

4) In healthcare and medical imaging, particularly in areas such as digital pathology, surgical instrument segmentation (EndoVis), and microscopic cell tracking, models with zero-shot capabilities that operate with limited data are vital for supporting clinical datasets and monitoring treatment responses through temporal consistency [22], [53].

4.3. Integrated rotoscoping pipeline

Modern rotoscoping systems have evolved from independent segmentation models into integrated pipeline architectures that synthesize complementary systems with distinct areas of expertise [1], [48]. These workflows enhance production quality while simultaneously reducing manual rotoscoping costs and enabling the generation of outputs that meet professional standards [2].

The promptable segmentation integrated pipeline tracks objects using prompts from user inputs through foundation models such as Segment Anything Model (SAM) and generates high-quality silhouettes by refining them in each frame [21], [24]. Artists can distinguish complex objects, intervene in frames where the model fails, instantly update the memory structure, and ensure error correction [21], [22], [47].

The temporal propagation integrated pipeline ensures temporal consistency by storing the object's past appearance features through memory modules such as XMem and SAM 2 [22], [54]. In challenging scenes, recurrent structures supported by optical flow data minimize mask drift by tracking object motions [31], [55].

In Identity Tracking and Video Instance Segmentation (VIS), maintaining identity consistency across objects is critically important, particularly in scenes containing multiple objects [26], [46]. VIS components such as IDOL and GenVIS associate objects in a high-dimensional feature space using ReID (re-identification) embeddings and temporal contrastive learning strategies [44], [46]. This strategy reduces identity switching errors in crowded scenarios where objects occlude each other or where visually similar objects are present [26], [44], [46].

Unified rotoscoping pipelines combine promptable segmentation, temporal propagation, identity tracking, high-resolution matting, and human-in-the-loop refinement to reduce manual work while preserving pixel-level quality and temporal consistency.

The matting and high-resolution refinement integrated pipeline performs refinement by integrating modules such as HQ-SAM or Background Matting V2 to convert binary masks into alpha mattes suitable for compositing [21], [47], [49]. This layer transforms coarse segmentation results into high-quality silhouettes appropriate for professional compositing workflows [47], [49].

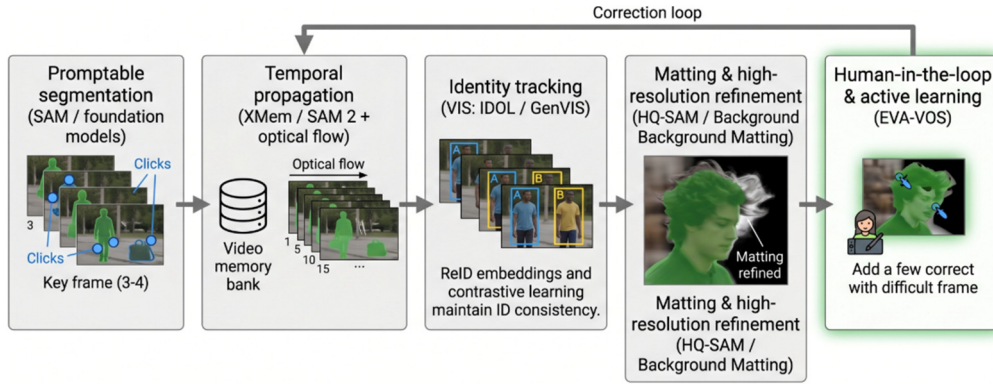

Fig. 11. Unified rotoscoping pipeline diagram for modern, modular rotoscoping systems

Human-in-the-loop and active learning integrated pipelines are systems that enhance efficiency and determine which frames require human intervention [1], [2], [20]. Systems operating on active learning principles, such as EVA-VOS, iteratively suggest through an agent which frame of the video should be annotated and by which method [2], [30]. Through this integrated architecture, modern unified rotoscoping pipelines minimize manual labor while guaranteeing the pixel-level precision and temporal stability required by professional productions [22], [39].

5. Discussion

This review demonstrates that rotoscoping has been redefined not merely as a production technique but as a complex problem representing one of the most comprehensive challenges in computer vision. Since approximately 2015, the literature has followed not a unidirectional but a multi-axial evolution, undergoing simultaneous transformations from frame-based segmentation to temporal modeling, from closed systems to the foundation model paradigm, and from fully automatic approaches to hybrid designs [22], [24], [28]. The first wave, initiated by FCN, U-Net, DeepLab, and Mask R-CNN, automated per-frame mask generation and reduced labor burden, while benchmarks such as DAVIS and YouTube-VOS established a standard evaluation framework around the J&F metric [4], [11], [29], [30]. Subsequent memory-based (STM, XMem, Cutie) and optical flow-based (SegFlow, RAFT) models significantly mitigated drift and occlusion problems in long sequences; SAM, SAM 2, SegGPT, and transformer-based video models have repositioned rotoscoping within the broader video segmentation paradigm and foundation model framework [46], [48], [56].

One of the most striking aspects in the literature is the persistence of practical deployment challenges in production despite these technical advances. Although performance metrics on DAVIS and YouTube-VOS benchmarks have nearly reached saturation, the same models still require substantial manual correction in real VFX pipelines; continuous increases in J&F scores have not eliminated usability issues in production environments [1], [6], [19]. Furthermore, the lack of systematic reporting of production metrics such as edge chatter frequency, correction time

per frame, and artist disagreement rate limits the industrial utility of academic progress and highlights the need for more realistic evaluation protocols [2], [11], [22]. The structural neglect of topics such as 4K/8K standards, hair strands and transparent details, and high-resolution matting reinforces the "resolution bottleneck and production quality" discussion addressed in Section 4.1. [49].

The momentum created by foundation models such as SAM and SAM 2 and transformer-based video approaches has brought with it a tendency to position these systems as the ultimate solution for rotoscoping; however, issues such as prompt sensitivity, hallucinated masks, domain gap, and long-term occlusion demonstrate that these models still have significant limitations [22], [24]. New benchmarks such as MOSE reveal that modern methods, including memory-based approaches, produce meaningful error rates when confronted with dense scene complexity; this confirms that the balance of "strengths and weaknesses" discussed in Section 3.6 has not yet been fully resolved in the production context [27]. These findings support the view that foundation models should be regarded as modular components supporting human-supervised hybrid systems rather than replacing existing rotoscoping pipelines.

Finally, the human-in-the-loop approach has been reevaluated in this review not as an indicator of deficiency but as a foundational architecture for quality assurance and creative control. MiVOS, Track Anything, and human-in-the-loop annotation systems demonstrate that user input is critical not only for error correction but also for uncertainty management and active learning [2], [21]. These findings indicate that the unified rotoscoping pipelines discussed in Sections 3.5 and 4.3 will be based on hybrid designs that position human expertise at the center. As of 2026, artificial intelligence appears not as an autonomous producer in rotoscoping tasks but as a high-capacity assistant requiring supervision; fully automatic rotoscoping under production constraints remains an unsolved problem [11], [19]. Therefore, the findings of this paper indicate that future research should move toward more holistic designs that reflect production realities in terms of both evaluation metrics and hybrid human-in-the-loop architectures. Moreover, these hybrid designs offer a critical framework that will facilitate the adoption of deep learning-based rotoscoping not only in VFX production but also in non-VFX application areas such as video editing, AR/VR, security, and medical imaging [11], [22].

6. Conclusions

This study has systematically examined developments in the field of deep learning-based rotoscoping within the context of the VOS/VSS ecosystem, memory-based and transformer-based video models, foundation segmentation approaches, optical flow integration, matting, and production metrics, classified existing methods to reveal gaps in the literature, and provided an evaluation framework for future research.

The review clearly demonstrates that rotoscoping has evolved in recent years from frame-based manual tools toward human-supervised hybrid automation systems based on memory-based video object segmentation, optical flow-assisted propagation, transformer-based video models, and foundation segmentation approaches [1], [6], [22].

Optical flow-based structures, widely used in early work to ensure temporal consistency, have been largely superseded by transformer-based self-attention mechanisms and hierarchical memory banks [6], [22], [37]. This technological transformation now defines rotoscoping not merely as a tracking problem but as an interactive sequence prediction task shaped by user-provided prompts such as points, boxes, or masks [22], [27].

On the other hand, while the use of synthetic datasets has largely addressed the data scarcity problem, the domain gap in the sim-to-real transition and computational costs in high-resolution processing remain central focal points for future research [19], [20], [49]. The common denominator in the literature is that rotoscoping is increasingly approaching professional production standards through multi-modular, interactive, and memory-centric unified pipelines [31], [37]. While hybrid human-in-the-loop designs enable the increasing adoption of automation,

artist feedback mechanisms in cases of erroneous automation, safety and ethical concerns, individual privacy, identity recognition, and data security emerge as important discussion topics [1], [2], [56].

In conclusion, fully automatic rotoscoping under production constraints remains an unsolved problem [2], [7], [24]. Human-supervised hybrid structures that render the rotoscoping process highly automated, further reduce human intervention in VFX and non-VFX domains, yet reserve quality assurance and creative decision-making roles for professional experts appear inevitable in the foreseeable future. This review anticipates that future rotoscoping research will concentrate on developing standardized evaluation protocols that reflect production realities, establishing research agendas specific to high-resolution and long-sequence scenarios, and maturing human-in-the-loop hybrid architectures.

Acknowledgements

The authors have not disclosed any funding.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] T. Zhou, F. Porikli, D. J. Crandall, L. van Gool, and W. Wang, "A survey on deep learning technique for video segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 6, pp. 7099–7122, Jun. 2023, https://doi.org/10.1109/tpami.2022.3225573

[2] T. Delatolas, V. Kalogeiton, and D. P. Papadopoulos, "Learning the what and how of annotation in video object segmentation," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6936–6946, Jan. 2024, https://doi.org/10.1109/wacv57701.2024.00680

[3] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in *Proceedings 8th IEEE International Conference on Computer Vision*, pp. 105–112, 2001, https://doi.org/10.1109/iccv.2001.937505

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with Atrous separable convolution for semantic image segmentation," in *Lecture Notes in Computer Science*, Vol. 11211, Cham: Springer International Publishing, 2018, pp. 833–851, https://doi.org/10.1007/978-3-030-01234-2_49

[5] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 4, pp. 640–651, 2017, https://doi.org/10.1109/tpami.2016.2572683

[6] M. Gao, F. Zheng, J. J. Q. Yu, C. Shan, G. Ding, and J. Han, "Deep learning for video object segmentation: a review," *Artificial Intelligence Review*, Vol. 56, No. 1, pp. 457–531, Jan. 2022, https://doi.org/10.1007/s10462-022-10176-7

[7] H. Hu, K. Ying, and H. Ding, "Segment anything across shots: a method and benchmark," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40, No. 6, pp. 4825–4833, Mar. 2026, https://doi.org/10.1609/aaai.v40i6.42485

[8] 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 4, pp. 834–848, Apr. 2018, https://doi.org/10.1109/tpami.2017.2699184

[9] E. Şahin and H. Yüce, "Prediction of water leakage in pipeline networks using graph convolutional network method," *Applied Sciences*, Vol. 13, No. 13, p. 7427, Jan. 2023, https://doi.org/10.3390/app13137427

[10] E. Şahin and H. Yüce, "Fault detection in pipelines using the graph convolutional network (GCN) method," (in Turkish), *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, Vol. 40, No. 1, pp. 673–684, Aug. 2024, https://doi.org/10.17341/gazimmfd.1306916

[11] F. Perazzi, J. Pont-Tuset, B. Mcwilliams, L. van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 724–732, Jun. 2016, https://doi.org/10.1109/cvpr.2016.85

[12] N. Xu et al., "Youtube-vos: sequence-to-sequence video object segmentation," in *Computer Vision*, pp. 603–619, 2018, https://doi.org/10.1007/978-3-030-01228-1_36

[13] H. Xie, H. Yao, S. Zhou, S. Zhang, and W. Sun, "Efficient regional memory network for video object segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1286–1295, Jan. 2021, https://doi.org/10.1109/cvpr46437.2021.00134

[14] K. Xu and A. Yao, "Accelerating video object segmentation with compressed video," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1332–1341, Jun. 2022, https://doi.org/10.1109/cvpr52688.2022.00140

[15] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: joint learning for video object segmentation and optical flow," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 686–695, Oct. 2017, https://doi.org/10.1109/iccv.2017.81

[16] J. Gong, F. C. Holsinger, and S. Yeung, "FlowVOS: weakly-supervised visual warping for detail-preserving and temporally consistent single-shot video object segmentation," in *Proceedings of the British Machine Vision Conference 2021*, Jan. 2021, https://doi.org/10.5244/c.35.103

[17] Z. Liu et al., "Swin transformer: hierarchical vision transformer using shifted windows," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, Oct. 2021, https://doi.org/10.1109/iccv48922.2021.00986

[18] M. Lan, J. Zhang, F. He, and L. Zhang, "Siamese network with interactive transformer for video object segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, No. 2, pp. 1228–1236, Jun. 2022, https://doi.org/10.1609/aaai.v36i2.20009

[19] H. Ding, C. Liu, S. He, X. Jiang, P. H. S. Torr, and S. Bai, "Mose: a new dataset for video object segmentation in complex scenes," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20167–20177, Oct. 2023, https://doi.org/10.1109/iccv51070.2023.01850

[20] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Modular interactive video object segmentation: interaction-to-mask, propagation and difference-aware fusion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5555–5564, Jun. 2021, https://doi.org/10.1109/cvpr46437.2021.00551

[21] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track anything: segment anything meets videos," *arXiv:2304.11968*, Jan. 2023, https://doi.org/10.48550/arxiv.2304.11968

[22] N. Ravi et al., "Sam 2: segment anything in images and videos," *arXiv:2408.00714*, Jan. 2024, https://doi.org/10.48550/arxiv.2408.00714

[23] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixe, D. Cremers, and L. van Gool, "One-shot video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5320–5329, Jul. 2017, https://doi.org/10.1109/cvpr.2017.565

[24] A. Kirillov et al., "Segment Anything," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, Oct. 2023, https://doi.org/10.1109/iccv51070.2023.00371

[25] D. Kim et al., "TubeFormer-deeplab: video mask transformer," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jan. 2022, https://doi.org/10.1109/cvpr52688.2022.01354

[26] J. Qi et al., "Occluded video instance segmentation: A benchmark," *International Journal of Computer Vision*, Vol. 130, No. 8, pp. 2022–2039, Aug. 2022, https://doi.org/10.1007/s11263-022-01629-1

[27] J. Wu, Y. Jiang, S. Bai, W. Zhang, and X. Bai, "SeqFormer: sequential transformer for video instance segmentation," in *Computer Vision*, pp. 553–569, Jan. 2022, https://doi.org/10.1007/978-3-031-19815-1_32

[28] S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. V. Gool, "The 2019 davis challenge on vos: unsupervised multi-object segmentation," in *CVPR Workshop/Challenge*, 2019.

[29] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbelaez, A. Sorkine-Hornung, and L. V. Gool, "The 2017 davis challenge on video object segmentation," in *CVPR Workshop/Challenge*, 2017.

[30] N. Xu et al., "Youtube-Vos: a large-scale video object segmentation benchmark," *arXiv:1809.03327*, 2018.

[31] T.-C. Wang et al., "Video-to-Video synthesis," in *Conference on Neural Information Processing Systems*, Jan. 2018.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, Jun. 2018, https://doi.org/10.1109/cvpr.2018.00068

[33] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3899–3908, Jun. 2016, https://doi.org/10.1109/cvpr.2016.423

[34] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, "Robust high-resolution video matting with temporal guidance," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3132–3141, Jan. 2022, https://doi.org/10.1109/wacv51458.2022.00319

[35] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Lecture Notes in Computer Science*, Vol. 12347, Cham: Springer International Publishing, 2020, pp. 402–419, https://doi.org/10.1007/978-3-030-58536-5_24

[36] X. Shi et al., "Flowformer++: masked cost volume autoencoding for pretraining optical flow estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1599–1610, Jun. 2023, https://doi.org/10.1109/cvpr52729.2023.00160

[37] H. K. Cheng and A. G. Schwing, "XMem: long-term video object segmentation with an atkinson-shiffrin memory model," in *Computer Vision*, pp. 640–658, Jan. 2022, https://doi.org/10.1007/978-3-031-19815-1_37

[38] S. Shaviro, "Emotion capture: affect in digital film," *Projections*, Vol. 1, No. 2, Jan. 2007, https://doi.org/10.3167/proj.2007.010204

[39] H. K. Cheng, S. Wug Oh, B. Price, A. Schwing, and J.-Y. Lee, "Tracking anything with decoupled video segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1316–1326, Oct. 2023, https://doi.org/10.1109/iccv51070.2023.00127

[40] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: evolution of siamese visual tracking with very deep networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4277–4286, Jun. 2019, https://doi.org/10.1109/cvpr.2019.00441

[41] X. Liu, Y. Luo, and W. Sun, "ASDeM: augmenting SAM with decoupled memory for video object segmentation," *IEEE Access*, Vol. 12, pp. 73218–73227, Jan. 2024, https://doi.org/10.1109/access.2024.3404463

[42] X. Yang, H. Wang, Xie, C. Deng, and D. Tao, "Associating objects with transformers for video object segmentation," *IEEE Transactions on Image Processing*, Vol. 31, pp. 2839–2849, 2022, https://doi.org/10.1109/tip.2022.3161832

[43] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289, Jun. 2022.

[44] J. Wu, Q. Liu, Y. Jiang, S. Bai, A. Yuille, and X. Bai, "In defense of online models for video instance segmentation," in *Computer Vision*, pp. 588–605, Jan. 2022, https://doi.org/10.1007/978-3-031-19815-1_34

[45] A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," in *International Conference on Learning Representations*, Jan. 2020.

[46] M. Heo et al., "A generalized framework for video instance segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14623–14632, Jun. 2023, https://doi.org/10.1109/cvpr52729.2023.01405

[47] M. Danelljan et al., "Segment anything in high quality," in *Advances in Neural Information Processing Systems 36*, pp. 29914–29934, 2023, https://doi.org/10.52202/075280-1303

[48] H. K. Cheng, S. W. Oh, B. Price, J.-Y. Lee, and A. Schwing, "Putting the object back into video object segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3151–3161, Jun. 2024, https://doi.org/10.1109/cvpr52733.2024.00304

[49] S. Lin, A. Ryabtsev, S. Sengupta, B. Curless, S. Seitz, and I. Kemelmacher-Shlizerman, "Real-time high-resolution background matting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8758–8767, Jun. 2021, https://doi.org/10.1109/cvpr46437.2021.00865

[50] Y. Sun, G. Wang, Q. Gu, C.-K. Tang, and Y.-W. Tai, "Deep video matting via spatio-temporal alignment and aggregation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6971–6980, Jun. 2021, https://doi.org/10.1109/cvpr46437.2021.00690

[51] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, Sep. 2017, https://doi.org/10.1109/iros.2017.8202133

[52] J. Xu, R. Ranftl, and V. Koltun, "Accurate optical flow via direct cost volume processing," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5807–5815, Jul. 2017, https://doi.org/10.1109/cvpr.2017.615

[53] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, Vol. 9351, pp. 234–241, 2015, https://doi.org/10.1007/978-3-319-24574-4_28

[54] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9225–9234, Oct. 2019, https://doi.org/10.1109/iccv.2019.00932

[55] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Rethinking space-time networks with improved memory coverage for efficient video object segmentation," in *Proceedings of the 35th International Conference on Neural Information Processing System*, 2021.

[56] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "SegGPT: segmenting everything in context," *arXiv:2304.03284*, Jan. 2023, https://doi.org/10.48550/arxiv.2304.03284

Deniz Yuce received a degree in Radio, Television, and Cinema from Kocaeli University, Kocaeli, Turkey, in 2011. He has been a Visual Effects Artist and Supervisor with more than 13 years of experience in high-end visual effects and post-production. He is currently a Flame Artist and Supervisor at Artworkslab, Texas, USA. His work specializes in compositing, beauty work, cleanup, and commercial finishing for international advertising, film, and episodic projects. His research interests include video segmentation, rotoscoping, video matting, and deep learning-based computer vision applications in visual effects production.