

# MAPS: A Morphology-Aware PPE Segmentation Framework for Healthcare Settings

Wanzhao Yang<sup>1</sup> Syed Anwar<sup>2</sup> Beomseok Park<sup>1</sup> Sifan Yuan<sup>1</sup>  
Aleksandra Sarcevic<sup>3</sup> Marius G. Linguraru<sup>2</sup> Randall S. Burd<sup>2</sup> Ivan Marsic<sup>1</sup>  
<sup>1</sup>Rutgers University <sup>2</sup>Children’s National Hospital <sup>3</sup>Drexel University

wanzhao.yang@rutgers.edu, sanwar@childrensnational.org, {bp593, sy609}@rutgers.edu  
as3653@drexel.edu, {mlingura, rburd}@childrensnational.org, marsic@rutgers.edu

## Abstract

*Monitoring adherence to personal protective equipment (PPE) guidelines is critical for infection control in clinical environments. Automated methods for monitoring require precise localization of PPE in complex real-world videos. While recent video segmentation models like SAM2 have shown strong performance, they underperform in healthcare settings due to cluttered backgrounds and frequent occlusions of small PPE items such as masks and gloves. We have identified two core limitations of SAM2 in this context: (1) difficulty in distinguishing PPE objects from complex backgrounds, and (2) tracking drift during occlusion. To address these issues, we propose MAPS: Morphology Aware PPE Segmentation, a training-free extension of SAM2 that incorporates two novel components: (1) a morphology-aware memory module that leverages shape descriptors to selectively retain reliable memory features and (2) a person-aware filtering module that removes predictions that do not align with detected person regions. MAPS achieves consistent improvements across multiple SAM2 model scales and outperforms recent SAM2-based extensions on a newly introduced PPE object tracking dataset. The code and the new dataset are available at <https://github.com/yangwanzhao/MAPS>.*

## 1. Introduction

Recent advancements in vision foundation models, such as Segment Anything Model 2 (SAM2) [21], have significantly enhanced image analysis performance in computer vision tasks, including zero-shot and few-shot segmentation. The effectiveness of SAM2 has not yet been evaluated in complex real-world settings, particularly in healthcare scenarios like personal protective equipment (PPE) detection and tracking. Monitoring PPE adherence is critical in healthcare environments for infection

control and medical personnel safety [1, 2, 16]. For automated monitoring, PPE items such as masks and gloves frequently become partially or fully occluded due to provider movements, patient interactions, and procedures, presenting unique challenges that standard segmentation methods have yet to address.

Healthcare environments are inherently complex, featuring diverse and cluttered backgrounds that can distract segmentation models, causing mistaken inclusion of unrelated background pixels as objects of interest. During occlusions, models like SAM2 may incorrectly shift their focus to unrelated objects, resulting in segmentation drift and incorrect object re-identification (Figure 1). To study these challenges, we develop a new PPE object tracking dataset derived from clinical simulation videos recorded in a resuscitation room. Guided by the observation that PPE objects of interest are usually shape-consistent, we propose two modules to address these challenges: the Morphology-Aware Memory module and the Person-Aware Filtering module. The Morphology-Aware Memory module leverages shape descriptors to maintain focus on correct PPE objects, ensuring shape consistency across frames. The Person-Aware Filtering module provides an effective strategy to reduce false positives by ensuring PPE regions remain spatially consistent with the corresponding person region.

Our main contributions are as follows:

- We developed a specialized PPE visual object tracking dataset explicitly capturing realistic clinical scenarios involving occlusions and background complexity.
- We conducted the first comprehensive evaluation and analysis of SAM2 performance for PPE detection and tracking, highlighting its limitations.
- We introduced MAPS, a novel Morphology-Aware PPE Segmentation framework incorporating the Morphology-Aware Memory module and

the Person-Aware Filtering module, that significantly improve segmentation accuracy and robustness in complex healthcare scenarios.

## 2. Related Work

### 2.1. PPE Detection

Automated PPE detection has emerged as an essential component in various industries to protect individuals from workplace hazards. In healthcare facilities, an increased emphasis on PPE adherence has emerged for infection control and healthcare personnel safety in the wake of the COVID-19 pandemic. Effective monitoring of PPE adherence helps ensure compliance, reduces the risk of healthcare-associated infections, and mitigates manual monitoring workloads. PPE detection using computer vision has relied on manual supervision or simplistic methods such as color thresholding or hand-crafted feature extraction [18, 19]. These methods were vulnerable to real-world variations including lighting conditions, camera angles, and partial occlusions.

Deep learning-based object detection methods, such as YOLO [22], Faster R-CNN [23], and Mask R-CNN [11], improved the robustness and accuracy of PPE detection by handling these environmental variations using extensive annotated datasets [3, 6, 10, 17, 31]. The inherently data-intensive nature of these fully supervised models limits their scalability and flexibility. Recent research has shifted towards few-shot and zero-shot detection strategies [20, 21, 25], that aim to address data limitations resulting from ground truth and training dataset size by learning generalized representations from generic datasets. These generalized methods often fail to adequately handle the unique and diverse complexities inherent to healthcare environments, such as varying object scales, cluttered backgrounds, and dynamic interactions among healthcare personnel. Our paper aims to bridge this gap by evaluating and enhancing advanced segmentation models tailored for real-world healthcare scenarios. In this paper, we propose the Morphology-Aware Memory and Person-Aware Filtering mechanisms to improve PPE detection and tracking accuracy.

### 2.2. Segment Anything Model 2

Segment Anything Model 2 [21] builds upon the original SAM model by incorporating interactive prompting mechanisms, refined segmentation capabilities, and advanced temporal tracking using a memory bank system. The memory bank in SAM2 stores segmentation data from preceding frames, improving prediction consistency and stability over video sequences. Several variants of SAM2 have been proposed to address specific segmentation and tracking challenges. SAMU-

RAI [29] primarily integrates a Kalman Filter-based tracking strategy to enhance temporal predictions, effectively addressing object drift and improving tracking accuracy. SAM2.1++ [27] uses a distractor-aware memory update module designed to manage and mitigate distractions, significantly enhancing object tracking stability. SAM2Long [7] introduces a constrained memory tree structure that optimizes the retrieval of historical segmentation predictions, maintaining robustness and accuracy over extended temporal sequences. Despite these advancements, current SAM2 variants primarily target generic object tracking scenarios and do not address the challenges posed by complicated environments like healthcare settings. These methods are not tailored to manage the complexities arising from cluttered backgrounds, frequent occlusions, and small object sizes commonly encountered in PPE tracking tasks. Our work addresses this gap by leveraging the consistent morphological characteristics of PPE items and introducing morphology-aware and spatial consistency constraints, enhancing segmentation performance in complex healthcare scenarios.

### 2.3. Morphology in Computer Vision

Morphological techniques have long served as a foundation in computer vision, offering fundamental operations such as erosion, dilation, opening, and closing for manipulating shapes in binary and grayscale images. These operations support tasks such as noise removal, segmentation, and boundary extraction, and are commonly used as pre-processing steps in vision pipelines [15, 24]. Building on these basic operations, shape descriptors like Hu moments [12] provide invariant features that are robust to translation, scale, and rotation. Hu moments enable quantitative shape comparison and have been widely applied in object recognition and pattern matching across domains such as industrial inspection and biomedical imaging [26]. More recently, the integration of morphological operations with deep learning frameworks has received growing attention. A meta-learning-based method was proposed to embed morphological operators into deep neural networks, showing that these nonlinear, shape-preserving transformations can enhance feature representations and improve performance in tasks such as image classification and edge detection [13]. Architectures that combine morphological and convolutional layers have achieved improved results in medical image classification, showing the synergy between classical morphology and contemporary learning-based methods [4].

Despite these advances, challenges remain in achieving invariance and robustness under real-world variability. The application of morphological methods to PPE segmentation and tracking, especially in dynamic and

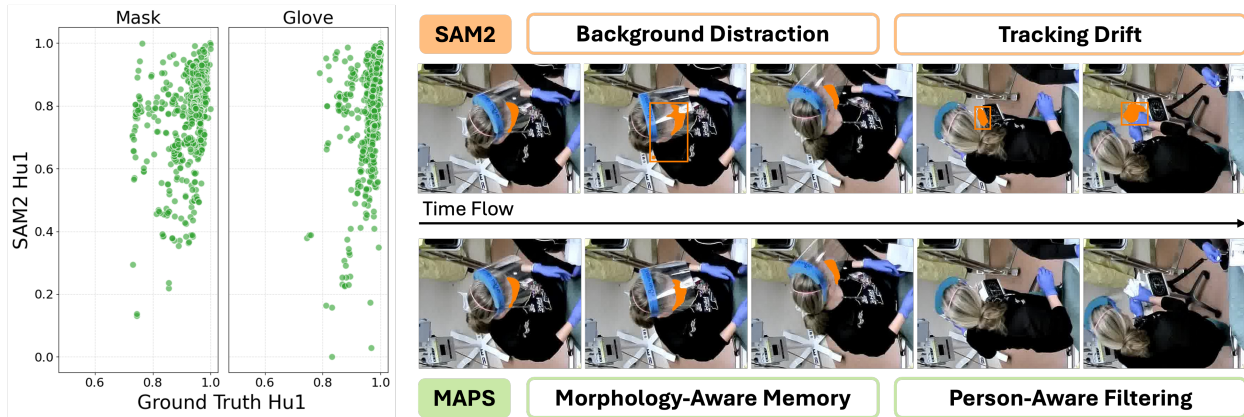


Figure 1. Overview of SAM2 limitations and MAPS improvements. Left: Scatter plots compare the normalized first Hu Moment (Hu1) for each frame between the ground truth and SAM2 predictions across two representative videos of PPE types, mask and glove. Ground truth values cluster tightly, reflecting consistent shape over time, while SAM2 values are scattered, indicating shape inconsistency. Right: Example frames show SAM2 predictions (top) include unrelated background pixels or drift off-target, highlighted with orange bounding boxes, while MAPS predictions (bottom) maintain focus by incorporating morphology-aware memory and person-aware filtering.

visually complex healthcare environments, has not been investigated. To address this gap, our study introduces an adaptive, Hu Moments-based filtering approach designed to maintain segmentation consistency and reliability in PPE tracking.

### 3. Exploring SAM2 for PPE detection

#### 3.1. PPE Tracking Dataset

We curate a specialized PPE object tracking dataset based on the R2PPE dataset [30], which was originally developed for PPE adherence monitoring in healthcare. The R2PPE dataset consists of images and videos recorded from medical simulations conducted in a resuscitation room. It captures various challenging scenarios, including small PPE objects, frequent occlusions due to personnel movements and interactions, and visually similar PPE adherence states. Original R2PPE annotations are bounding boxes for masks, gloves, gowns, and eyewear provided at 1 frame per second (fps). In this study, we focused on gloves and masks due to their role in infection control and tracking challenges.

To construct our visual object tracking dataset, we reviewed each video in R2PPE. When we identify a frame with annotated bounding boxes of targeted objects (masks or gloves), we extracted a one-minute video clip starting from that frame at 15 fps, yielding 900 frames per clip. SAM2 was then applied to each identified object in the initial frame using its bounding box as the prompt, propagating segmentation outputs across all frames in the one-minute clip. We repeated this process for subsequent frames in the original R2PPE videos to generate tracks. The SAM2-generated segmenta-

tion outputs served as guides to align with the original bounding box annotations available at 1 fps. By comparing object class IDs and spatial locations, we tracked each PPE object throughout the one-minute clips, creating consistent bounding box trajectories. Finally, a human-in-the-loop refinement step ensured the precision and correctness of the generated tracks. Human annotators reviewed and refined bounding box coordinates and corrected any inaccuracies or drift in tracking caused by SAM2. This refinement guarantees high-quality bounding box trajectories suitable for robust evaluation.

Our final PPE object tracking dataset comprises 160 one-minute video clips, including 112 glove and 48 mask clips, representing a resource for evaluating tracking performance in realistic clinical scenarios. This dataset is designed for evaluation purposes and is comparable in scale to widely used public object tracking datasets such as OTB100 [28], LaSOT [8], LaSOT<sub>ext</sub>[9], and GOT-10k[14].

#### 3.2. SAM2 Evaluation

We evaluated the performance of SAM2 for tracking PPE items within our PPE object tracking dataset. A key advantage of SAM2 is its interactive prompting capability, allowing manual correction of segmentation results at selected frames. To investigate this capability, we applied SAM2 to our dataset using different numbers of prompts. We first used only the bounding box from the first frame as a single prompt. We then increased the number of prompts, ensuring new prompts were evenly distributed throughout the video clips to maximize effectiveness.

The tracking performance of SAM2 on our dataset

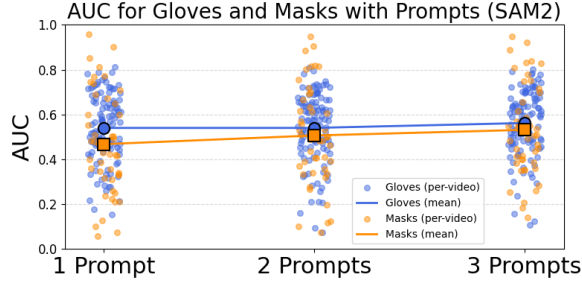


Figure 2. AUC of SAM2 on the PPE dataset for gloves and masks, using 1, 2, and 3 prompts. Each scatter point represents the AUC for an individual video, while the solid lines and highlighted markers indicate the mean AUC for each setting.

is shown using one, two, and three prompts (Figure 2). The scatter plot shows per-video Area Under Curve (AUC) values along with their means for both gloves and masks. Despite the increased number of prompts, the overall improvement in segmentation performance remains marginal. The per-video AUCs exhibit considerable variability, and the mean AUCs show limited change across different prompt counts. These results suggest that SAM2 frequently loses track of PPE items and fails to maintain consistent segmentation over time, even with additional user guidance.

In addition to the prompt-based evaluation, we further explored SAM2’s segmentation quality through visual and statistical analysis (Figure 1). This figure displays the distribution of the first Hu Moment values—a shape descriptor—for both SAM2 predictions and ground truth annotations, along with qualitative examples from representative video clips. The distributions reveal that SAM2 produces more variable shape representations compared to the consistent shapes in the ground truth, indicating the difficulty in preserving object morphology. The accompanying visual examples highlight cases where SAM2 outputs deviate from object boundaries or include unrelated background content, reinforcing the segmentation challenges in real-world clinical environments.

### 3.3. SAM2 Limitations on PPE Datasets

The results from prompt-based experiments, along with the insights from the first Hu Moment analysis, reveal limitations in SAM2’s ability to maintain consistent and accurate PPE segmentation. These analyses identify two primary failure modes of SAM2 in PPE tracking tasks:

- **Background Distraction:** SAM2 struggles to maintain accurate segmentation when operating in cluttered and dynamic clinical environments. It often includes additional pixels from the surrounding background, lead-

ing to incorrect object boundaries and noisy predictions. This limitation is exacerbated in healthcare settings where various instruments, uniforms, and overlapping body parts create visual distractions.

- **Tracking Drift:** When PPE items become partially occluded or change shape due to movement, SAM2 often fails to adapt, resulting in segmentation outputs drifting toward nearby regions. This drifting reduces tracking reliability and may lead to loss of object identity across frames.

These critical limitations observed in our dataset motivate our proposed approach, aiming to enhance segmentation robustness and accuracy by addressing both background distractions and tracking drift. The targeted morphological and spatial consistency modules introduced in subsequent sections mitigate these challenges.

## 4. Methodology

To address the limitations encountered by SAM2 during PPE segmentation and tracking tasks, we introduce the MAPS model (Figure 3). In this section, we first introduce the preliminaries of Hu Moments in section 4.1. The proposed Morphology-Aware Memory module and Person-Aware Filtering module in MAPS are introduced in section 4.2 and 4.3, respectively.

### 4.1. Hu Moments

Morphological descriptors are valuable in PPE segmentation tasks due to the consistent and distinctive shapes of PPE items such as masks and gloves. Among these descriptors, Hu Moments are particularly useful because they are invariant to translation, rotation, and scaling. This property makes them effective for assessing shape consistency across frames, especially in dynamic settings.

Given a binary segmentation output  $I(x, y)$ , central moments  $\mu_{pq}$  are computed as:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q I(x, y), \quad (1)$$

where  $p$  and  $q$  are non-negative integers that specify the order of the moment with respect to the  $x$  and  $y$  coordinates, and  $\bar{x}$  and  $\bar{y}$  are the centroid coordinates:

$$\bar{x} = \frac{\sum_x \sum_y x I(x, y)}{\sum_x \sum_y I(x, y)}, \quad \bar{y} = \frac{\sum_x \sum_y y I(x, y)}{\sum_x \sum_y I(x, y)}. \quad (2)$$

These central moments are normalized to form scale-invariant moments  $\eta_{pq}$ :

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(1+(p+q)/2)}}. \quad (3)$$

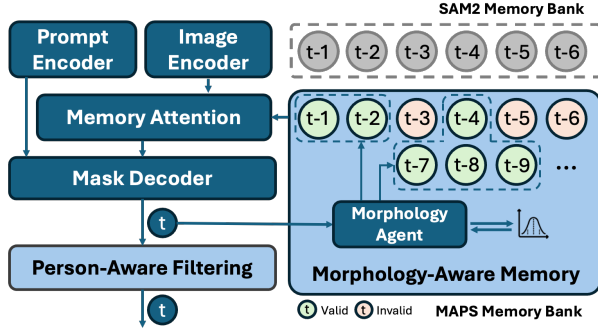


Figure 3. Overview of the proposed MAPS model. The Morphology-Aware Memory retains only shape-consistent masks for memory attention, unlike SAM2 which uses recent frames without selection. The Person-Aware Filtering discards PPE masks that lack sufficient overlap with person regions.

From these normalized central moments, Hu defined seven moment invariants ( $h_1, h_2, \dots, h_7$ ) through non-linear combinations. These moments capture different shape characteristics:  $h_1$  quantifies overall compactness,  $h_2$  measures symmetry, and  $h_3$  through  $h_7$  represent higher-order shape variations such as elongation, skewness, and spatial complexity.

These invariants allow comparison of object shapes regardless of their position, orientation, or scale. In this study, we focus on the first Hu Moment  $h_1 = \eta_{20} + \eta_{02}$  due to its ability to reflect the general compactness and outline of PPE shapes, which is particularly relevant for evaluating segmentation quality.

As described in Section 3.2, we compare the first Hu Moment distributions between ground truth and SAM2 predictions to assess shape consistency (Figure 1). We used two representative videos—one for gloves and one for masks—and compute  $h_1$  for each frame. Each point in the scatter plot corresponds to the  $h_1$  value of a ground truth PPE region (x-axis) and a SAM2 prediction (y-axis). Ground truth values cluster tightly, indicating consistent shapes, while SAM2 values are more dispersed, reflecting greater variability.

These findings highlight SAM2’s limitations in preserving shape consistency across frames, reinforcing the need for shape-aware filtering strategies.

## 4.2. Morphology-Aware Memory Module

Hu Moments are invariant under translation, rotation, and scaling, making them suitable for comparing shapes of PPE masks and gloves, which typically exhibit consistent morphological characteristics across frames. Building on this property, we introduce the Morphology-Aware Memory mechanism to identify and retain temporally consistent predictions during PPE tracking.

To quantify the shape consistency of segmentation predictions over time, we used an Exponential Moving Average (EMA) strategy to maintain running estimates of both the mean and variance of the first Hu Moment values. The design of our filtering module is inspired by Gaussian-based confidence bounds. We assume that consistent Hu Moments values over time should follow a distribution centered around a dynamic mean with bounded variance. We treat the first Hu Moment values as samples from a distribution and define validity based on how many standard deviations a sample deviates from the current mean. A prediction is considered valid if:

$$|x_t - \mu_t| \leq k_t \cdot \sqrt{v_t} \quad (4)$$

where  $x_t$  is the first Hu Moment value calculated from the segmentation prediction at time  $t$ ,  $\mu_t$  is the EMA mean,  $v_t$  is the EMA variance, and  $k_t$  is the confidence interval coefficient that controls the strictness of the Hu-based memory selection boundary.

We initialize the EMA from the first Hu Moment value  $x_0$  from the first frame, which is used as a prompt and is trusted. The mean is updated in two steps. First, an intermediate estimate is computed as

$$\hat{\mu}_{t+1} = \alpha_t x_t + (1 - \alpha_t) \mu_t, \quad (5)$$

allowing recent predictions to influence the estimate. Then, this estimate is anchored to the initial value with

$$\mu_{t+1} = \beta_t x_0 + (1 - \beta_t) \hat{\mu}_{t+1}, \quad (6)$$

preserving alignment with the prompt while allowing adaptation to the evolving statistics. The variance is updated with the standard EMA formula:

$$v_{t+1} = \alpha_t (x_t - \mu_t)^2 + (1 - \alpha_t) v_t. \quad (7)$$

This formulation enables the filter to adaptively reflect the distribution of consistent Hu values over time.

To progressively shift the filtering behavior from lenient to strict, we schedule the parameters  $\alpha_t$ ,  $\beta_t$ , and  $k_t$  based on the normalized frame index  $r_t = t/T$ .  $\alpha_t$  increases over time to give more weight to recent predictions, making the filter more adaptive as confidence builds.  $\beta_t$  decreases over time to reduce the influence of the initial prompt frame, allowing the filter to evolve based on observed consistency.  $k_t$  decreases over time to tighten the acceptance range, enforcing stricter shape consistency as more data accumulates. These scheduling functions are defined as:

$$\alpha_t = \alpha_{\text{start}} + (\alpha_{\text{end}} - \alpha_{\text{start}})(1 - e^{-\gamma r_t}), \quad (8)$$

$$\beta_t = \beta_{\text{end}} + (\beta_{\text{start}} - \beta_{\text{end}})e^{-\gamma r_t}, \quad (9)$$

$$k_t = k_{\text{end}} + \frac{k_{\text{start}} - k_{\text{end}}}{1 + e^{\lambda(r_t - 0.5)}} \quad (10)$$

where  $\alpha_{\text{start}}$ ,  $\alpha_{\text{end}}$ ,  $\beta_{\text{start}}$ ,  $\beta_{\text{end}}$ ,  $k_{\text{start}}$ , and  $k_{\text{end}}$  are the start and end values for each parameter, and  $\gamma$  and  $\lambda$  control the decay and sharpness of the scheduling transitions.

To integrate this module into the SAM2 framework, we maintain a separate EMA agent instance for each object being tracked. At each new frame, only predictions from prior frames whose Hu values are deemed valid by the EMA agent are included in the memory bank. Frames with no prediction or an invalidated shape are excluded. This selective memory mechanism ensures that only morphologically consistent predictions contribute to subsequent segmentation, improving the temporal stability and robustness of PPE tracking in complex environments. Because this approach operates entirely during inference, it does not require any model training.

### 4.3. Person-Aware Filtering Module

To further enhance segmentation reliability, we introduce the Person-Aware Filtering module, leveraging the spatial relationship between PPE objects and their corresponding persons. This approach is based on the assumption that PPE objects such as masks and gloves are worn on a person’s body and must therefore spatially overlap with a person’s segmentation region. Given the initial bounding box prompt at the first frame, we applied YOLO-World [5], a flexible zero-shot object detection model, to detect persons by setting person as the target category. Among the detected bounding boxes, we select the one containing the provided prompt bounding box, robustly associating the PPE object with its corresponding person. Users can also directly provide a bounding box around the person along with the prompt point, enabling straightforward initialization.

With the selected person bounding box, we run a parallel SAM2 instance dedicated to segmenting and tracking the person across subsequent frames. Person detection and tracking using SAM2 are generally reliable due to the large size and well-defined boundaries of human figures, making this process stable even across long temporal sequences.

The resulting person segmentation regions are then used as spatial constraints for PPE segmentation predictions. For each frame, we compute the overlap ratio between the predicted PPE segmentation region  $M_{\text{ppe}}$  and the corresponding person segmentation region  $M_{\text{person}}$  as:

$$\text{Overlap}(M_{\text{ppe}}, M_{\text{person}}) = \frac{|M_{\text{ppe}} \cap M_{\text{person}}|}{|M_{\text{ppe}}|}. \quad (11)$$

This overlap ratio quantifies how much of the PPE prediction lies within the detected person region. A low ratio suggests that the PPE prediction has likely drifted

to the background or another object, indicating a false positive. In these cases, the prediction is rejected.

## 5. Experiments

We evaluated the performance of MAPS by comparing it with SAM2 and other state-of-the-art segmentation models on the PPE tracking dataset. Following the original SAM2 paper [21] and recent extensions [7, 27, 29], we adopted two groups of evaluation metrics. Center-based metrics include precision (P@20) and normalized precision ( $P_{\text{norm}}$ ). IoU-based metrics include area under the curve (AUC), average overlap (AO), and overlap precision at different IoU thresholds (OP@0.5, OP@0.75). For our Morphology-Aware Memory module, we empirically set the scheduling ranges for the parameters as follows:  $\alpha$  from 0.2 to 0.8,  $\beta$  from 1.0 to 0.1, and  $k$  from 6 to 2. The decay coefficients  $\lambda$  and  $\gamma$  are set to 10 and 5, respectively. We used the updated SAM2.1 model suite for all SAM2-based experiments. Unless otherwise noted, all baselines and variants, including MAPS, were initialized from the same pretrained SAM2.1 checkpoints and inherit the original configurations.

### 5.1. Comparison with SAM2

To show the effectiveness of the proposed morphology-aware method, we conducted a detailed comparison of SAM2 and MAPS on our PPE tracking dataset across four model sizes (Table 1). MAPS consistently outperforms SAM2 in all evaluation metrics and improvements are observed across all scales. The performance gain is most noticeable in center-based metrics such as P@20 and  $P_{\text{norm}}$ , suggesting enhanced spatial accuracy of the segmentation predictions. Gains in IoU-based metrics indicate that MAPS provides more reliable segmentation boundaries over time. These improvements are particularly meaningful for PPE items like masks and gloves, where maintaining tight alignment with the actual object shape is crucial for adherence tracking.

### 5.2. Comparison with SOTA

To further validate the robustness of MAPS, we compared it with recent extensions of the SAM2 framework, including SAMURAI [29], SAM2.1++ [27], and SAM2Long [7]. We report results separately for two critical PPE types—mask and glove—using AUC,  $P_{\text{norm}}$ , and OP@0.75 as evaluation metrics (Table 2). These three metrics are selected because AUC captures overall segmentation quality, while  $P_{\text{norm}}$  and OP@0.75 represent center-based and IoU-based criteria, respectively.

For mask tracking, MAPS achieves the highest scores across all metrics, surpassing other methods by a notable margin. This result highlights the effectiveness of the proposed modules in handling subtle facial PPE

Table 1. PPE object tracking performance comparison between MAPS and SAM2 across model sizes.

Model	AUC(%)	P@20(%)	P <sub>norm</sub> (%)	AO(%)	OP@0.5(%)	OP@0.75(%)
SAM2.1-tiny	50.38	72.21	78.26	50.30	56.60	22.70
MAPS-tiny	<b>51.52</b> (1.14↑)	<b>79.14</b> (6.93↑)	<b>82.37</b> (4.11↑)	<b>51.42</b> (1.12↑)	<b>59.52</b> (2.92↑)	<b>24.14</b> (1.44↑)
SAM2.1-small	52.13	77.58	81.48	52.05	59.64	23.28
MAPS-small	<b>53.29</b> (1.16↑)	<b>82.17</b> (4.59↑)	<b>83.84</b> (2.36↑)	<b>53.21</b> (1.16↑)	<b>61.44</b> (1.80↑)	<b>25.34</b> (2.06↑)
SAM2.1-base	51.82	77.60	80.54	51.73	58.41	23.67
MAPS-base	<b>53.51</b> (1.69↑)	<b>83.57</b> (5.97↑)	<b>84.94</b> (4.40↑)	<b>53.42</b> (1.69↑)	<b>60.53</b> (2.12↑)	<b>25.21</b> (1.54↑)
SAM2.1-large	53.00	81.94	83.88	52.93	59.91	23.32
MAPS-large	<b>54.74</b> (1.74↑)	<b>86.64</b> (4.70↑)	<b>87.02</b> (3.14↑)	<b>54.67</b> (1.74↑)	<b>62.40</b> (2.49↑)	<b>25.98</b> (2.66↑)

Table 2. Comparison of MAPS with SOTA SAM2-based models on PPE object tracking. Results are reported separately for PPE type of mask and glove.

Mask			
Method	AUC(%)	P <sub>norm</sub> (%)	OP@0.75(%)
SAMURAI	33.23	66.03	13.33
SAM2.1++	42.18	72.13	22.43
SAM2Long	45.04	73.72	24.90
MAPS	<b>49.84</b>	<b>78.79</b>	<b>28.78</b>
Glove			
Method	AUC(%)	P <sub>norm</sub> (%)	OP@0.75(%)
SAMURAI	52.80	84.07	20.36
SAM2.1++	53.19	83.74	22.05
SAM2Long	<b>55.39</b>	85.93	22.56
MAPS	55.08	<b>87.57</b>	<b>23.68</b>

with precise spatial constraints. For glove tracking, MAPS maintains competitive performance. SAM2Long slightly outperforms MAPS in AUC for gloves, which may be attributed to its use of multiple segmentation pathways. These results indicate that MAPS improves robustness across PPE types while retaining the overall strengths compared with existing SAM2 extensions.

### 5.3. Ablation Studies

**Effect of the first Hu Moment constraint.** To further show the morphological consistency improvements brought by MAPS, we analyzed the distribution of Hu1 values from segmentation predictions. Compared with SAM2, the standard deviation of Hu1 decreases by 9%, from 0.33 to 0.30, indicating more stable shape predictions over time. This improvement is achieved even though the Hu1 constraint is not directly applied to the predicted regions. It is instead used as a soft filtering

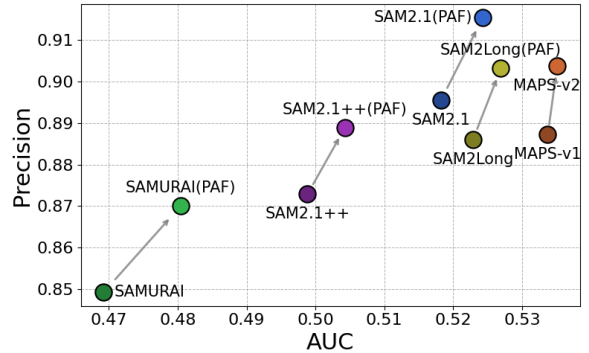


Figure 4. AUC and precision for SAM2.1, its extensions, and MAPS models with and without the Person-Aware Filtering module (PAF). MAPS-v1 includes only the Morphology-Aware module while MAPS-v2 additionally applies PAF. PAF-enhanced variants consistently improve precision and AUC across all models.

mechanism during memory selection. Despite this indirect influence, the constraint still enhances the overall consistency. In addition, we compared the AUC scores of the proposed model when only the Morphology-Aware Memory module is applied (MAPS-v1) (Figure 4). This variant already outperforms SAM2.1 and all other recent extensions of SAM2 by a clear margin, showing the effectiveness of incorporating morphological filtering.

**Effect of the Person-Aware Filtering module.** The Person-Aware Filtering module is designed to remove predicted PPE regions that do not sufficiently overlap with person regions, which often occurs due to tracking drift caused by PPE occlusions (Figure 1). This mechanism effectively reduces false positives that appear outside of reasonable human body regions. The module features a plug-and-play design that enables it to be seamlessly integrated into any person-related object tracking

Table 3. Ablation on the sensitivity of the confidence interval coefficient  $k$ .

$k$	AUC(%)	$P_{\text{norm}}$ (%)	OP@0.75(%)
1	53.47	84.68	<b>25.58</b>
2	<b>53.51</b>	<b>84.94</b>	25.21
3	53.34	84.80	25.23

framework without requiring retraining. To evaluate the effectiveness of the Person-Aware Filtering module, we analyzed changes in precision, which reflects how well false positives are suppressed. We applied the filtering module to several baselines, including SAM2.1 and its extensions. We compared each model’s AUC and precision (Figure 4). MAPS-v1 and MAPS-v2 represent the proposed model without and with the Person-Aware Filtering module, respectively. Applying the module consistently improved precision by approximately 2% and leads to corresponding increases in AUC. These findings confirm the module’s effectiveness in enhancing prediction reliability and segmentation quality.

**Effect of the confidence interval coefficient  $k$ .** The confidence interval coefficient  $k$  plays a critical role in determining whether a segmentation prediction’s first Hu Moment is accepted into the memory bank. This effect is especially important in later stages of prediction when the mean and variance estimates are more stable, making the threshold sensitive to new observations. To examine the effect of  $k$ , we evaluate three fixed values (Table 3).  $k = 2$  offers the best balance across AUC,  $P_{\text{norm}}$ , and OP@0.75, providing a reasonable filtering threshold without being overly restrictive.

## 6. Conclusion

We identified two key limitations of SAM2 in PPE object tracking in clinical settings: its tendency to include irrelevant background regions and frequent tracking drift due to occlusions. To address these challenges, we proposed MAPS, which introduces two targeted modules: a Morphology-Aware Memory module that filters out shape-inconsistent predictions based on shape descriptors, and a Person-Aware Filtering module that removes off-target predictions by enforcing spatial alignment with person regions. Our experiments show that MAPS consistently outperforms SAM2 and its state-of-the-art extensions across various metrics and model sizes, confirming the effectiveness of our approach for tracking PPE items such as masks and gloves.

While MAPS achieves strong results, it still has limitations. It underperforms when compared to SAM2Long in glove tracking AUC, likely due to the inherently vari-

able morphology of gloves caused by finger movements. MAPS relies on several scheduled parameters, which increases design complexity. In future work, we aim to explore higher-order Hu moments to better represent deformable PPE shapes, develop parameter-free alternatives to simplify deployment, and extend our system to support a broader set of PPE categories to enhance broad infection control coverage.

## References

- [1] Emily C Alberto, Kathleen H McCarthy, Colleen A Hamilton, Jacob Shalkevich, Zachary P Milestone, Rima Izem, Jennifer L Fritzeen, Ivan Marsic, Aleksandra Sarcevic, Karen J O’Connell, et al. Personal protective equipment adherence of pediatric resuscitation team members during the covid-19 pandemic. *Annals of emergency medicine*, 78(5):619–627, 2021. 1
- [2] Onga Bangani, René English, and Angela Dramowski. Intensive care unit nurses’ knowledge, attitudes and practices of covid-19 infection prevention and control. *Southern African Journal of Infectious Diseases*, 38(1):478, 2023. 1
- [3] Ng Wei Bhing and Patrick Sebastian. Personal protective equipment detection with live camera. In *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 221–226. IEEE, 2021. 2
- [4] Martha Rebeca Canales-Fiscal and José Gerardo Tamez-Peña. Hybrid morphological-convolutional neural networks for computer-aided diagnosis. *Frontiers in Artificial Intelligence*, 6:1253183, 2023. 2
- [5] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xingang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. *arXiv preprint arXiv:2401.17270*, 2024. 6
- [6] Rishit Dagli and Ali Mustafa Shaikh. Cppe-5: Medical personal protective equipment dataset. *SN Computer Science*, 4(3):263, 2023. 2
- [7] Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. *arXiv preprint arXiv:2410.16268*, 2024. 2, 6
- [8] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 3
- [9] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 129:439–461, 2021. 3
- [10] Weili Fang, Lieyun Ding, Hanbin Luo, and Peter ED Love. Falls from heights: A computer vision-based approach for safety harness detection. *Automation in construction*, 91:53–61, 2018. 2

- [11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2
- [12] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8(2): 179–187, 1962. 2
- [13] Yufei Hu, Nacim Belkhir, Jesus Angulo, Angela Yao, and Gianni Franchi. Learning deep morphological networks with neural architecture search. *Pattern Recognition*, 131:108893, 2022. 2
- [14] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019. 3
- [15] Y. Kong et al. A review of mathematical morphology applications in image analysis. *Journal of Imaging*, 8(10):272, 2022. 2
- [16] Tomer Lamhoo, Noa Ben Shoshan, Hagit Eisenberg, Gilad Fainberg, Mansour Mhiliya, Neta Cohen, Orly Bisker-Kassif, Orly Barak, Carolyn Weiniger, and Tali Capua. Emergency department impaired adherence to personal protective equipment donning and doffing protocols during the covid-19 pandemic. *Israel Journal of Health Policy Research*, 10:1–6, 2021. 1
- [17] Jye-Hwang Lo, Lee-Kuo Lin, and Chu-Chun Hung. Real-time personal protective equipment compliance detection based on deep learning algorithm. *Sustainability*, 15(1):391, 2022. 2
- [18] Bahaa Eddine Mneymneh, Mohamad Abbas, and Hiam Khoury. Vision-based framework for intelligent monitoring of hardhat wearing on construction sites. *Journal of Computing in Civil Engineering*, 33(2):04018066, 2019. 2
- [19] Man-Woo Park, Nehad Elsafty, and Zhenhua Zhu. Hardhat-wearing detection for enhancing on-site safety of construction workers. *Journal of Construction Engineering and Management*, 141(9):04015024, 2015. 2
- [20] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4051–4070, 2022. 2
- [21] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 2, 6
- [22] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2
- [23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 2
- [24] Jean Serra. *Image Analysis and Mathematical Morphology*. Academic Press, 1982. 2
- [25] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 55(13s):1–40, 2023. 2
- [26] Usman Tariq et al. Shape-based image retrieval using hu moments. *Multimedia Tools and Applications*, 79: 32491–32508, 2020. 2
- [27] Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with sam2. *arXiv preprint arXiv:2411.17576*, 2024. 2, 6
- [28] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013. 3
- [29] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024. 2, 6
- [30] Wanzhao Yang, Mary S Kim, Genevieve J Sippel, Aaron H Mun, Kathleen H McCarthy, Beomseok Park, Aleksandra Sarcevic, Marius George Linguraru, Ivan Marsic, and Randall S Burd. An image dataset for surveillance of personal protective equipment adherence in healthcare. *Scientific Data*, 12(1):96, 2025. 3
- [31] Qiang Zhang, Ziyu Pei, Rong Guo, Haojun Zhang, Wanru Kong, Jie Lu, and Xueyan Liu. An automated detection approach of protective equipment donning for medical staff under covid-19 using deep learning. *CMES-Computer Modeling in Engineering & Sciences*, 132(3), 2022. 2