Fig. 1. (a) Multi-View Inconsistency Example (b) Video Sequence Case: A simpler form of multi-view inconsistency appears in video sequences, where adjacent frames exhibit small camera pose changes. (c) Results of the original SAM: The figure shows SAM's outputs on the two frames. Same-colored regions share the same mask ID. SAM fails to consistently match objects across views, splitting or merging them differently in each frame. (d) Results of MastSAM: Our MastSAM method assigns the same mask ID (color) to the same object across frames, effectively correlating masks across different views.

***Abstract*—**

Abstract

With the emerging importance of understanding 3D environments, such as spatial intelligence and 3D foundation models, researchers have sought to distill knowledge from off-the-shelf 2D foundation models such as CLIP and SAM. However, these 2D foundation models often produce inconsistent information across different views. To tackle this issue, we present MastSAM. This method leverages Mast3R's ability to map 2D pixel coordinates from image pairs into a shared 3D space. By doing so, MastSAM enables consistent tracking of corresponding points across multiple views so that 2D foundation models such as SAM can output multi-view-consistent segmentation. Our main contributions are as follows: 1) Clearly defining the multi-view inconsistency problem in 2D foundation models. 2) Proposing a novel solution to minimize the multi-view inconsistency problem using MastSAM.

***Index Terms*—component, formatting, style, styling, insert**

Introduction

Recent advances in 2D foundation models such as CLIP and SAM have greatly simplified traditional 2D vision tasks. Meanwhile, 3D awareness, perception, and understanding have emerged as central areas of focus in computer vision. Several works have tried to combine 3D information with 2D embedding to attain more precise information about 2D images. [**?**]However, several challenges arise in 3D, with data scarcity and diversity being at the forefront. Compared with 2D images, 3D data are more difficult to collect. Although large-scale outdoor scene datasets [**?**], indoor scene datasets [**?**], [**?**], object-level datasets [**?**], [**?**], and part-level datasets [**?**] suffice for demonstration-level experiments, they still lack generality for open-vocabulary settings. Furthermore, the diversity of 3D representations such as meshes, point clouds, and occupancy grids makes it difficult to devise a universal model architecture.

A common approach is to distill information from 2D foundation models (e.g., CLIP or SAM) into 3D. Early work by [**?**] uses a radiance-field representation to learn language embeddings for each patch, while [**?**] distills knowledge from [**?**] to achieve 3D-aware segmentation. However, due to limitations of radiance fields' implicit representations, direct 3D manipulation is not possible. Instead, one must instead render images back to 2D to produce language embeddings and segmentation masks. With the emergence of Gaussian Splatting [**?**]—an explicit and compact representation—several works [**?**], [**?**], [**?**] have demonstrated that lifting 2D knowledge into 3D can significantly improve 3D perception.

Despite these developments, multi-view inconsistency remains a major challenge. As shown in Fig.1, such inconsistencies occur in both object-level and scene-level segmentation, ultimately causing semantic corruption in 3D. For example in Fig.1 (a), the bear's nose, eyes, and feet are segmented separately in one view, but merged into a single mask in another. This issue is also evident in simpler video segmentation tasks where SAM often fails to maintain consistent masks across adjacent frames, leading to increasing errors with larger viewpoint changes. Fig.1 shows two adjacent frames from the Davis dataset [**?**]—used here as inputs for both our MastSAM model and the original SAM model [**?**].

To address this, we formally define the multi-view inconsistency problem and propose a baseline solution. Our method, MastSAM, leverages Mast3R [**?**] to establish point-to-point correspondences. We initialize masks using SAM's Auto-Mask-Generator by selecting representative points from each mask. We then track these points in subsequent frames according to Mast3R's guidance. During this procedure, we iteratively introduce new masks in previously unmasked regions. As shown in Fig.1, MastSAM effectively mitigates multi-view inconsistency. Our key contributions are threefold:

- **Formal Definition:** We formally define multi-view inconsistency in 3D segmentation.
- **New Metric:** We propose a novel metric to quantitatively measure multi-view consistency.

- **MastSAM Algorithm:** We present MastSAM and demonstrate its upper-bound performance on the proposed metric.

In the following sections, we will first define the problem and introduce related work. Then we will introduce our method, show metrics, and evaluate both qualitative and quantitative experiment results.

II. Multi-view Inconsistency

In this section, we formally define the multi-view inconsistency problem. We start with an intuitive explanation, and then a mathematical definition. As the name suggests, there are several 2D images where the information corresponding to the same 3D location are inconsistent among all 2D views. As shown in the Fig.**??**.

Formally, we have an image set $\mathscr{I}$, where $\mathscr{I} = I_1, I_2, ..., I_n$. We also define a pixel space $\mathbb{P}$ for each image i as shown in Equ.1. where W_i and H_i is the width and height of image i.

$$\mathbb{P}_i = \{x, y | x \in 1, 2, \ldots, W_i, y \in 1, 2, \ldots, H_i\} \tag{1}$$

We then define a function $\mathscr{F}$ that generates a functionf_i that maps pixel space to consistent coordinates space, i.e. $\mathbb{P} \to \mathbb{R}^3$ as shown in Equ.3

$$f_i : \mathbb{P}_i \to \mathbb{R}^3, f_i x, y = x', y', z' \tag{2}$$

$$\mathscr{F} : I \times \mathscr{I} \to \mathbb{P}_i \to \mathbb{R}^3, Fi, \mathscr{I} = f_i \tag{3}$$

Given the above equation, we define the multi-view consistency function h_w as a mapping from a pixel space to a score, i.e. $\mathbb{P}_w \to \mathbb{R}$. Its specific definition is shown in Equ.6. Notice that the multi-view consistency is used to measure the power of a model instead of measuring the dataset we use. Therefore, we currently regard function $\mathscr{F}$ as a perfect function that outputs accurate 3D coordinates.

$$\varepsilon_w x, y = \{i, j, k \in \mathscr{I} \times \mathbb{P}_i \mid \|f_w x, y - f_i j, k\| \leq \varepsilon\} \tag{4}$$

$$h_w x, y, M = \frac{\sum_{i,j,k \in \varepsilon_w x,y} \frac{M_i j,k \cdot M_w x,y}{\|M_i j,k\|_2 \cdot \|M_w x,y\|_2}}{|\varepsilon_w x, y|} \tag{5}$$

$$H\mathscr{I}, \mathscr{M} = \frac{1}{|\{I, \mathbb{P}_i\}|} \sum_{w,x,y \in \{I, \mathbb{P}_i\}} h_w x, y, M \tag{6}$$

As demonstrated in the above equation Equ.6, we first find all pixels in the different images within a reasonable range to the input pixel on the corresponding 3D space. This reasonable range means that those pixels point at approximately the same location in 3D space. For each pixel, we then calculate the cosine similarity with the input pixel. Finally, the weighted average is the result of the consistency score of that pixel using a current 2D foundation model M.

Although the above metrics would be reliable and accurate to measure multi-view inconsistency, it would be practically impossible to calculate the metrics due to computational complexity. Therefore, an easier approximation may vary from a different down-stream task. The aforementioned multi-view inconsistency measurement metrics poses several issues, especially when using the metrics as a loss function. This is known as the degeneration problem. When all outputs of a model become the same, the inconsistency becomes zero, but the original function of the model is lost. Therefore, it would be unreasonable to use these metrics. They should be used in conjunction with other downstream tasks, such as segmentation accuracy. In the experiment section, we will explain how an implementation of an approximation of the above metrics to down-stream tasks such as SAM and CLIP.

III. Related work

Related work

A. Image Matching

Image matching aims to establish correspondences between pixels across different images of the same scene, capturing global spatial consistency. Previous works have utilized keypoint-based matching, connecting sparse, locally invariant features between images. Traditional methods, such as those based on epipolar geometry [**?**], [**?**], [**?**], enforce spatial constraints by leveraging geometric relationships between camera views. Handcrafted approaches like SIFT [**?**], [**?**] achieve image matching through robust local feature extraction. Modern methods like SuperGlue [**?**] enhance image matching using graph-based attention, improving robustness under challenging conditions. However, their reliance on keypoint-based matching limits their effectiveness in handling extreme viewpoint changes.

In contrast, recent advances such as DUSt3R [**?**] and MASt3R [**?**] use dense matching, establishing correspondences for all pixels. In addition, they treat image matching as a 3D problem, centered around camera pose and scene geometry. DUSt3R redefines pairwise reconstruction as the regression of 3D pointmaps [**?**], achieving robustness to viewpoint and illumination changes without relying on explicit matching supervision. Building on DUSt3R, MASt3R significantly improves the accuracy of pairwise matches and the reciprocal matching speed by adding a feature matching module that aligns dense local features in 3D space [**?**]. By using dense matching and grounding spatial consistency in a 3D framework, MASt3R addresses key limitations of traditional and modern methods, making it a robust and efficient solution for image-matching tasks.

B. Image segmentation

Segment anything (SAM) [**?**] is a zero shot image segmentation model that inputs points or bounding boxes and outputs a corresponding segmentation mask. Built on the Vision Transformer (ViT) architecture [**?**] and trained on the large-scale SA-1B dataset [**?**], SAM demonstrates remarkable segmentation performance for single-image tasks.

Extending SAM's [**?**] capabilities to video segmentation, recent approaches have emerged. SAM2 [**?**] introduces frame-by-frame segmentation using a sparse attention mechanism to efficiently track temporal changes. While this approach adapts well to frame variations, it relies heavily on the quality of

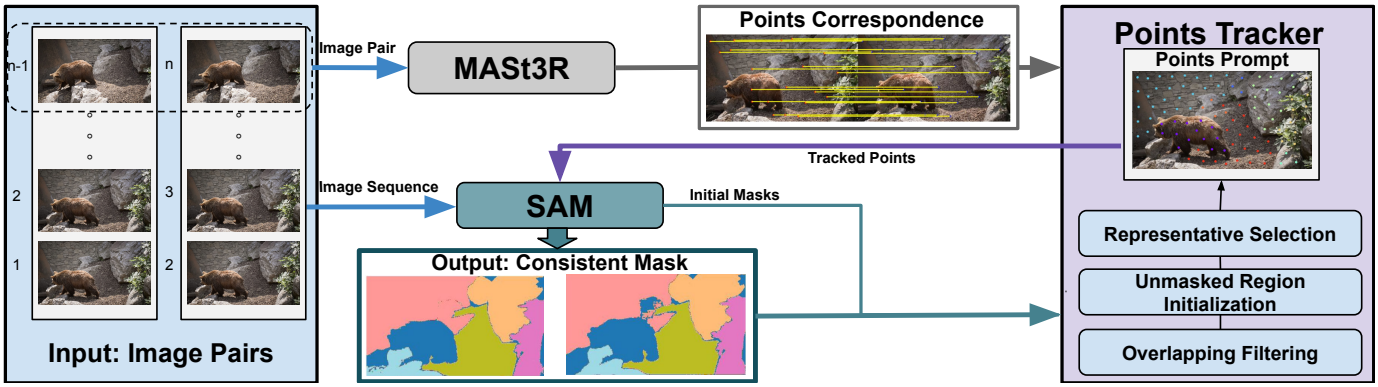

Fig. 2. Overview of our approach. Given a pair of image sequences, one sequence is input to SAM, and an image pair is input to MASt3R. SAM produces a set of image masks, which the Prompt Points Tracker processes. The output of MASt3R is a set of correspondence points, which is also input to the Prompt Points Tracker. The Prompt Points Tracker applies overlapping filtering, decreasing the number of masks. The Point Tracker then initializes the unmasked regions, increasing the number of masks back to dynamic equilibrium from the previous step, and selects representative points to ensure multi-view consistency. Each group of points is then input back into SAM to produce the final segmented image, aiming to address the multi-view consistency problem. The next frame of masks is sent to the Points Tracker to generate consistent masks for subsequent frames.

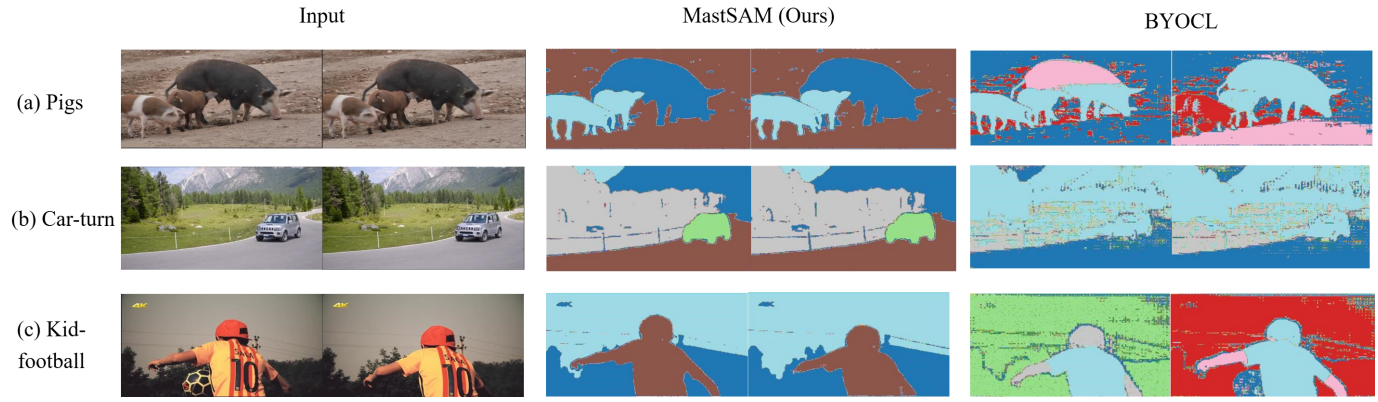

Fig. 3. This figure compares Mast-SAM and BYOCL in different data sequences. For each data sequence, the left-most image is the original image, the middle image is the output of Mast-SAM, and the right-most image is the output of BYOCL. As one can see, the MastSAM can correlate previous images and current images together. This is shown by the same color on the correlated mask. Although the BYOCL method can sometimes correlate the consistency well, it loses its ability to correctly segment the original mask.

initial prompts and lacks robust mechanisms for continuity tracking across frames. Similarly, the Track Anything Model (TAM) [**?**] addresses the need for consistent segmentation by combining SAM's high-quality segmentation with XMem's [**?**] memory mechanism. TAM integrates SAM [**?**] for precise mask initialization and refines with XMem for semi-supervised video object segmentation. On the other hand, the matching Anything by Segment Anything (MASA) [**?**] model focuses on tracking multiple moving objects through bounding boxes, utilizing SAM and various data transformers [**?**]. Despite these advancements, current video segmentation models face several limitations. Their performance tends to decline over time, especially during long video sequences, as segmentation accuracy heavily depends on the quality of the initialization. Poor-quality initial masks can significantly impair the results. Most critically, these works treat each frame as an independent unit, providing individually optimized outputs with a limited understanding of overall continuity. This frame-centric approach often fails to capture global spatial relationships, making them susceptible to challenges like object occlusion, rotation, or viewpoint transformation.

To address these limitations, our work combines SAM [**?**] with MASt3R [**?**], leveraging MASt3R's ability to robustly capture spatial continuity across frames. By grounding segmentation in a 3D perspective, our approach ensures robust performance even when the camera poses changes, bridging the gap between frame-by-frame optimization and spatially consistent video segmentation.

IV. Methods

Methods

Our method consists of three main components: MASt3R, SAM, and the Points Tracker. These modules interact with each other in a sequence designed to ensure multi-view consistency and produce accurate, dynamically adjusted segmentation masks across all image frames.

A. Input and Pre-processing

The input comprises of sequential image pairs drawn from a multi-frame dataset, such as [**?**] [**?**], which is representative of typical video. This approach can be easily adapted to [**?**] and other 3D scan datasets. Each pair is structured for simultaneous processing by two pipelines. One pipeline is aimed at correspondence detection by MASt3R and the other pipeline is focused on mask consistency produced by SAM. It is significant to note that the dataset is processed for uniform resolution and aligned using intrinsic and extrinsic camera parameters for efficient downstream processing.

B. MASt3R for Points Correspondence

MASt3R takes each input image pair and identifies sparse, but accurate correspondences across frames. It leverages feature extraction through transformer-based models and cross-image attention mechanisms. Features between images are aligned, matching scores are computed, and correspondences are filtered based on confidence thresholds. These correspondence points are one of the two essential inputs for initializing the Points Tracker.

MASt3R outputs correspondence maps, which are then further refined by combining reciprocal matches, ensuring point consistency.

C. SAM for Mask Consistency

Simultaneously, the SAM module processes each image frame in the image sequence to generate segmentation masks. This allows SAM to adapt the segmentation task to produce consistent masks for regions of interest in the image sequence. SAM accounts for the dynamic nature of the sequence by integrating initial masks and point maps provided by the Points Tracker in subsequent iterations. This feedback loop ensures that the masks are constantly evolving with the updated tracked points.

D. Points Tracker for Multi-view Refinement

The Points Tracker serves as the critical intermediary for ensuring multi-view consistency and accurate propagation of masks across the sequence. Its core components include:

- **Overlapping Filtering:** The Points Tracker module applies a post-processing step to remove overlapping masks using geometric constraints and consistency checks, ensuring each segment remains well-defined and unique.
- **Representative Selection:** Correspondence points identified by MASt3R are refined through clustering (K-Means) and confidence filtering to select representative points for multi-view alignment.
- **Unmasked Region Initialization:** Regions outside the initial mask coverage are dynamically identified and initialized, ensuring that new regions entering the view are incorporated. This step is crucial to ensure we have an equilibrium in the number of input and output masks, especially after the Overlapping Filtering step reducing the number of masks in the module.

These steps optimize camera parameters, 3D depth maps, and point alignments iteratively to ensure robust multi-view consistency.

E. Pipeline Integration

The outputs from the MASt3R and SAM modules feed into the Points Tracker, which generate refined masks which are then re-inputted into SAM. This iterative process allows for feedback-driven mask updates, with MASt3R ensuring that tracked points maintain temporally and spatially consistent. Leveraging the synergy between segmentation and correspondence tracking, this pipeline effectively addresses challenges in dynamic multi-view sequences.

The final output is a set of consistently aligned segmentation masks and a reconstructed 3D point cloud for the scene. These outputs enable robust segmentation and analysis of multi-view dynamic datasets.

V. Experiment and Results

Experiment and results

A. Introduction to Experiment

To extensively evaluate MastSAM, we performed experiments on two open-source datasets: the Davis benchmark [**?**] and Mose.

The ground truth annotation in both datasets only refer to one mask of the major object, whereas MastSAM automatically outputs an image, possessing masks of every object. This discrepancy made direct comparisons between MastSAM and the ground truth annotations impractical. To address this, we developed the following evaluation protocol to ensure a comprehensive and fair assessment of MastSAM's performance.

- **Ground Truth Replication:** We replicated the ground truth (GT) annotation for each annotation image as many times as the number of masks generated by MastSAM.
- **Best Mask Selection:** For each predicted mask, we calculated the Intersection over Union (IoU) with the corresponding GT annotations. The mask with the highest IoU value among all predicted masks for each image was selected.
- **Metrics Computation:** Using the best mask for each image, we computed the following evaluation metrics: IoU, F1 score, precision, and recall. These metrics collectively offer a comprehensive evaluation of the regional and boundary precision of MastSAM. Here, we list the formula used to calculate the IoU, F1 score, precision and recall:

$$\text{IoU} = \frac{\text{Area}\big(\text{Predicted} \cap \text{Ground Truth}\big)}{\text{Area}\big(\text{Predicted} \cup \text{Ground Truth}\big)}$$

$$\text{F}_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision Recall}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP FP}}$$

Fig. 4. Here, we propose more comparisons regarding the segmentation result instead of the complexity result. However, our major focus in the current paper is on how to solve the consistency problem, and the downstream task result should be the secondary consideration

$$\text{Recall} = \frac{\text{TP}}{\text{TP FN}}$$

Furthermore, we adopted the parallel evaluation method to facilitate running speed and GPU space. Four GPUs were used following the 'queue' logic. Once a GPU was set free, a new unprocessed sequence would be sent to this GPU. With this logic, MastSAM was able to segment 100 images consistently within 15 minutes.

B. Experiment Results

1) Quantitative Results: Compared to BYOCL, our method demonstrates significant improvements across all metrics, highlighting its superior segmentation performance. Although the IoU and F1 scores of our method are slightly lower than those of SAM1, our approach achieves higher temporal consistency between consecutive frames, as evidenced by a more balanced Precision and Recall. The specific quantitative results are shown in the tableI below:

TABLE I
PERFORMANCE METRICS OF DIFFERENT METHODS ON THE DAVIS DATASET.

Method	IoU	F1	Precision	Recall	Consistency
BYOCL	0.3906	0.527	0.4621	0.6671	78.73
Ours	0.4553	0.5228	0.5226	0.5988	92.32
SAM	0.6787	0.786	0.9891	0.6965	NA

We did not compare our method with SAM2 because these two methods were different in segmenting videos. MastSAM segmented videos of one scene combining 3D knowledge. However, SAM2 simply segments and tracks a specified object using correspondence between frames. Although the results looked similar, it was unfair to compare these two methods on a dataset mostly containing one, monotonous view.

These results underscore the effectiveness of our method in addressing segmentation consistency across frames. By leveraging MastSAM's ability to ensure multi-view consistency, our approach achieves a robust trade-off between accuracy and temporal stability, which is essential for video segmentation tasks.

2) Visualization Results: As shown in Fig.3 Fig.4, our method is qualitatively much better compared to BYOCL. MastSAM is capable of segmenting consistent and complete masks while BYOCL suffers from mask confusion and fails to get satisfactory segmentation results.

C. Ablation

TABLE II
ABLATION BETWEEN NO UNMASKED REGION INITIALIZATION AND FULL CAPACITY

Method	IoU	F1	Precision	Recall
Ours w/o initialization	0.3678	0.4211	0.5682	0.3780
Ours	0.4553	0.5228	0.5226	0.5988

In this part, we show the essence of our design logic. Due to our method's nature of expanding masks, it will lose its ability to segment a meaningful result. This is due to the fact all masks will aggregate into one result. To prevent this degenerate situation, we import a counter measure by iteratively segmenting unsegmented areas using the original SAM model. This way, we stabilize the number of segmentation masks. Our ablation study shown in Tab.II displays our assumption clearly.

VI. FUTURE WORK

FUTURE WORK

We proposed an innovative method, MastSAM, to address the multi-view inconsistency problem inherent in 2D foundation models. Our experimental results demonstrate promising performance on benchmark video segmentation datasets like the DAVIS dataset. However, future work should expand the evaluation of MastSAM across diverse datasets. This includes not only dynamic video sequences, but also static multi-view scenarios that cover both indoor and outdoor scenes (such as those provided by ScanNet [**?**], ScanNet++ [**?**], or other 3D indoor datasets. This expanded evaluation will not only validate MastSAM's versatility but may reveal further areas for development.

MastSAM's approach unlocks numerous promising research directions and practical applications. More concretely, the conceptual framework that forms the basis of MastSAM can be adapted to additional 2D and 3D foundation models beyond just SAM [**?**], such as DINOv2 [**?**] and Segment Anything Model 2 [**?**]. By combining MastSAM's ability to correlate multi-view representations with recent advancements in foundation models, future research can achieve greater accuracy, efficiency, and versatility.

Extending to potential applications of MastSAM, future work can explore how to deploy our method in real-time and on mobile devices. With the growing demand for efficient and compact segmentation models in fields such as augmented reality, virtual reality, robotics, and mobile computing, real-time multi-view segmentation consistency would be revolutionary. The current limitation of MastSAM is its computational complexity, especially its reliance on foundation models. As foundational models develop, perhaps the core principle of MastSAM can be adapted to achieve real-time multi-view consistent segmentation. Simplifying MastSAM's architecture and developing a lightweight variation along with developments in foundation models would bring us closer to this goal.

Another promising direction for the future of MastSAM is incorporating a human-in-the-loop approach [**?**]. Having this approach could improve segmentation accuracy and flexibility. Applications in professional fields, such as medical imaging or industrial inspections, that often require a high degree of precision may especially benefit from having a human-in-the-loop. Moreover, this may enable adaptive learning within MastSAM, which could progressively improve performance based on human experience and feedback. This approach may bridge the gap between fully automated segmentation with discrepancies and enhance the day-to-day tasks of many industry professionals.

Moreover, further research can explore integrating the MastSAM framework with cross-modal approaches. For example, employing segmentation outputs with depth maps or LiDAR data could facilitate consistency and robustness in 3D reconstruction. Cross-modal fusion would not only reinforce MastSAM's segmentation accuracy but also enable richer 3D scene understanding.

By pursuing these avenues of future research, MastSAM can be enhanced significantly in terms of practicality, efficiency, accuracy, and adaptability.

VII. Conclusion

Conclusion

In this work, we formalize the concept of multi-view inconsistency in 3D segmentation, addressing a significant challenge that current 2D foundation models face when applied across multiple views. To quantify this inconsistency, we propose the first dedicated metric explicitly designed to evaluate consistency across views.

We introduced MastSAM, a novel algorithm tailored to mitigate multi-view inconsistency, achieving theoretical upper-bound performance on our proposed metric. Our experimental evaluations on established video object segmentation benchmarks demonstrates MastSAM's capability for robust multi-view analysis. These results reinforce MastSAM's value as a powerful tool for improving consistent segmentation, suggesting promising applications across diverse domains such as augmented reality, virtual reality, robotics, and other industries.

However, our current framework has limitations, each providing clear avenues for future improvements: (1) While MastSAM provides strong empirical results, further rigorous experiments on diverse benchmarks are needed to generalize its efficacy; (2) The computational complexity of our pipeline necessitates task-specific adaptations for downstream applications, which may limit plug-and-play usability. (3) Our experiments focus on video sequences, leaving open questions about performance in static multi-view settings (e.g., ScanNet, indoor scenes).

Future work will directly address these challenges by optimizing computational efficiency and broadening MastSAM's practical viability through experimentation and evaluation across various 3D datasets and settings. Moreover, additional promising directions include integrating advanced foundation models and adopting human-in-the-loop methods for enhanced precision and adaptability. By pursuing these future paths, MastSAM can evolve and generalize to real world applications, bringing practical impact in day-to-day tasks.