*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**

I. Introduction

## 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.**??**, such inconsistencies occur in both object-level and scene-level segmentation, ultimately causing semantic corruption in 3D. For example in Fig.**??** (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.**??** 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.**??**, 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. 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.