RISeg: Robot Interactive Object Segmentation via Body Frame-Invariant Features

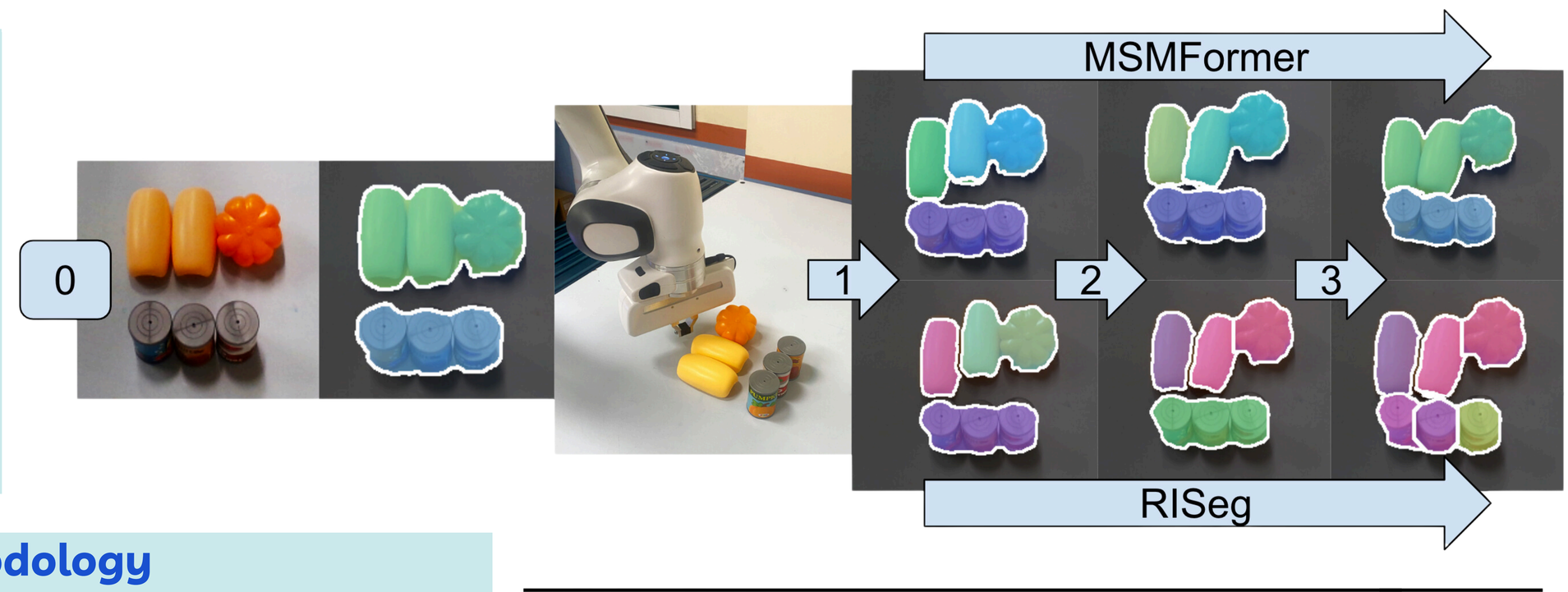
Howard Qian[1], Yangxiao Lu[2], Kejia Ren[1], Gaotian Wang[1], Ninad Khargonkar[2], Yu Xiang[2], and Kaiyu Hang[1]
Rice University[1], University of Texas at Dallas[2]

Abstract

Robots must be proficient in segmenting unseen objects to execute tasks in new environments. Previous works train deep neural networks on large-scale data, where cluttered scenes often result in under segmentation. We **introduce a novel approach to correct inaccurate segmentation by using robot interaction** at regions of uncertainty and a designed body frame-invariant feature based on rigid body motions. We demonstrate the effectiveness of RISeg in accurately segmenting cluttered scenes by achieving an **average object segmentation accuracy rate of 80.7%, an increase of 28.2%** when compared with other state-of-the-art UOIS methods.

Objective

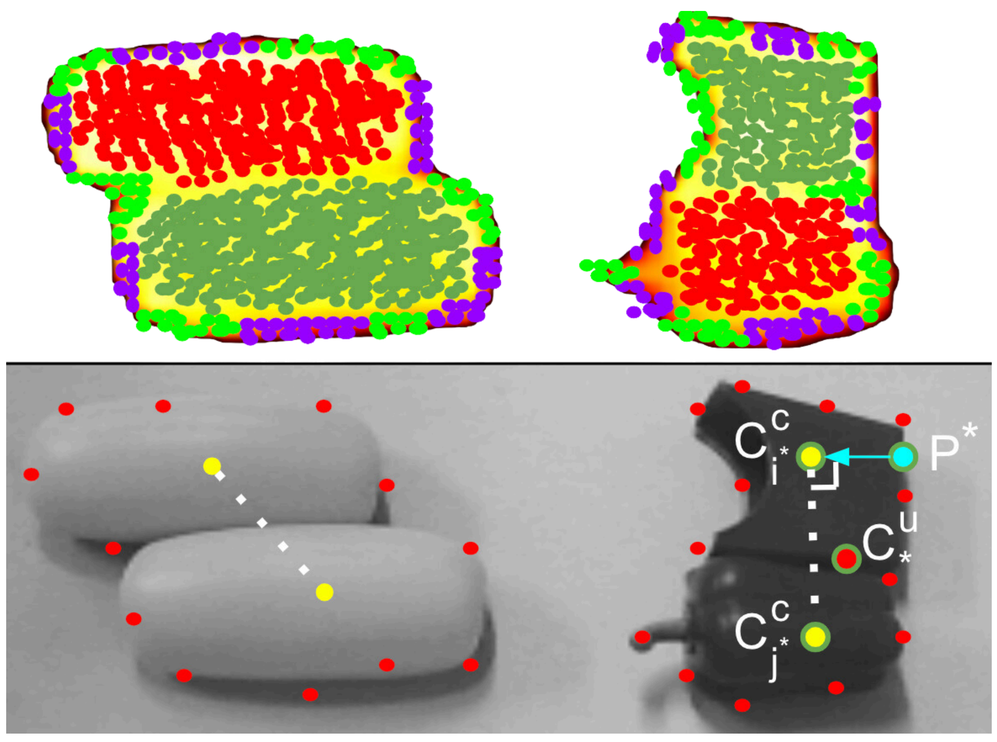
Improve unseen object instance segmentation (UOIS) in cluttered environments without the need for object singulation and by interacting with the scene in a **minimal and non-disruptive manner**.



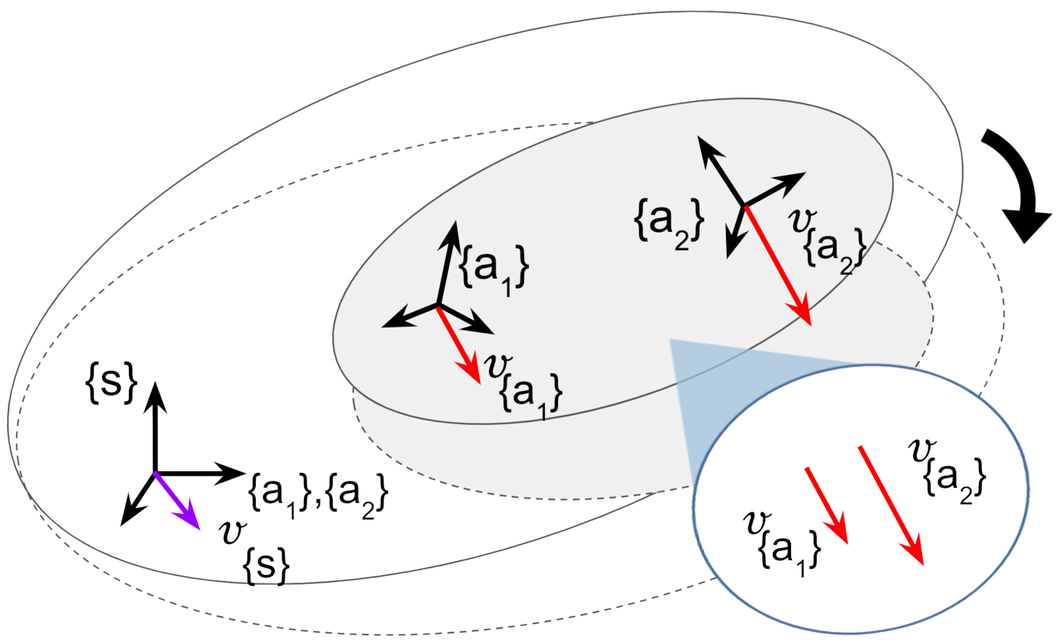
Methodology

The RISeg interactive perception framework makes 2 main contributions in **uncertainty-driven action planning** and **rigid body motion-based segmentation mask correction**. RISeg uses static segmentation masks and pixelwise uncertainty scores from MSMFormer for both action selection and baseline segmentation masks. Furthermore, body frame-invariant feature (BFIF) analysis is done on static images taken before and after each robot interaction.

Heuristic Action Selection (Alg. 2) using Uncertainty Heatmap and K-Means Clustering



Body Frame-Invariant Feature (Alg. 3) based on Spatial Twists and Rigid Body Motions



Algorithm 1 RISeg

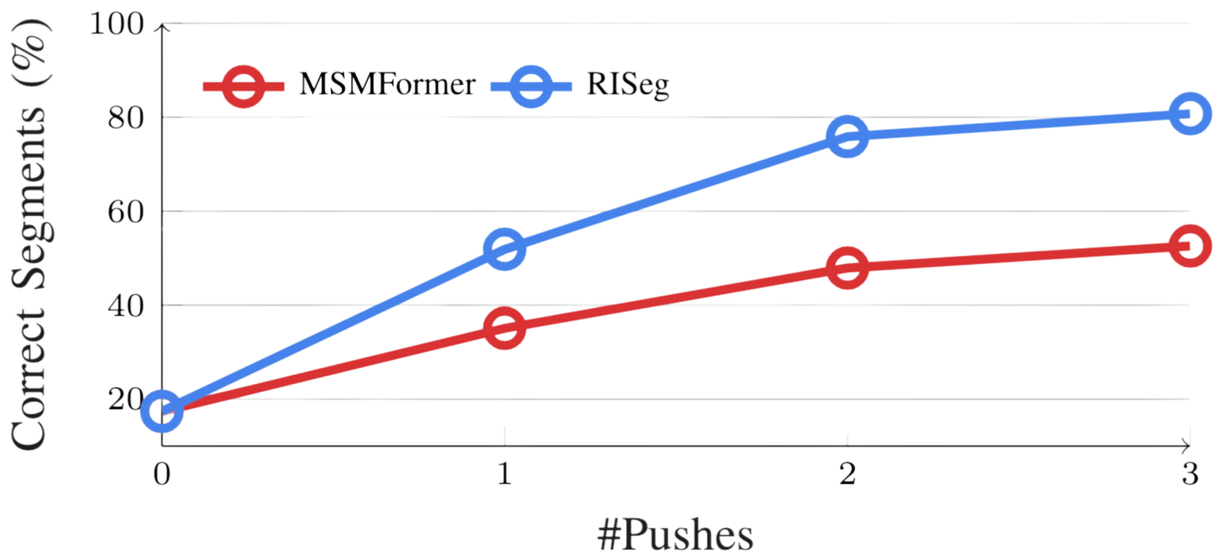
Input: I_0, StaticSeg$(\cdot)$
Output: $\hat{L}_{t+1}$

$t \leftarrow 0$
$L_t \leftarrow \text{StaticSeg}(I_t)$
$\hat{L}_t \leftarrow L_t$
while $a_t \leftarrow \text{FindAction}(I_t)$ **not** *null* **do** ▷ Alg. 2
  $I_{t+1} \leftarrow \text{Interact}(a_t)$
  $L_{t+1} \leftarrow \text{StaticSeg}(I_{t+1})$
  $\hat{L}_{t+1} \leftarrow \text{UpdateMask}(I_t, I_{t+1}, \hat{L}_t, L_{t+1})$ ▷ Alg. 3
  $t \leftarrow t + 1$
return $\hat{L}_{t+1}$

Results and Discussion

After all robot interactions on cluttered tabletop scenes, RISeg achieves an **average segmentation accuracy rate of 80.7%**, while MSMFormer achieves 52.5%. Pixelwise accuracy metrics also reflect RISeg outperforming MSMFormer across interactions.

Since BFIF analysis is done on static images of the scene before and after each interaction, we believe future work on continuous video analysis of BFIFs can further increase performance.



Related literature: MSMFormer

Y. Lu, Y. Chen, N. Ruozzi, and Y. Xiang, "*Mean Shift Mask Transformer for Unseen Object Instance Segmentation*," in IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024.