

PROCEEDINGS

Hybrid-AOI: Reliable Arm-Object Interaction Perception for On-Orbit Robotic Operational Behavior Recognition

Zhihao Zhang¹, Zhaohui Dang^{1,*}, and Bohan Jiao¹

¹Northwestern Polytechnical University, Xi'an, 710129, China

*Corresponding Author: Zhaohui Dang. Email: dangzhaohui@nwpu.edu.cn

ABSTRACT

Reliable perception of non-cooperative on-orbit robotic operations requires more than identifying spacecraft, components, or poses. Operational behavior recognition further asks which platform is acting, through which manipulator, on which target, and with what operational semantics. Inspired by Human-Object Interaction and visual semantic role labeling, Arm-Object Interaction (AOI) addresses this need by representing a robotic operation as a structured visual event with four roles: Agent, Instrument, Object, and Verb. By explicitly binding these roles to an operational verb, AOI extends conventional object-centric space perception toward role-centric operational behavior recognition. Existing AOI studies have demonstrated the value of this formulation for single-image on-orbit and embodied robotic scenarios and have established dedicated benchmarks with representative behaviors such as observe, carry, crawl, contract, extend, and other robot-arm-assisted operations. However, their outputs remain predominantly box-level. This coarse grounding is still vulnerable in observe-like behaviors, self-inspection scenes, and tightly coupled arm-object configurations, where the manipulated target is visually similar to the acting platform, partially occluded, or distinguishable mainly through local contact, attachment, and assembly cues rather than coarse bounding boxes.

This paper introduces Hybrid-AOI, a role-selective Arm-Object Interaction modeling framework for on-orbit operational behavior recognition. The central idea is selective role segmentation. Hybrid-AOI preserves box-level grounding for the Agent and assigns mask-level grounding to the Instrument and Object, yielding a hybrid box-mask role representation under the original AOI semantic schema. This design is motivated by the error characteristics observed in AOI: in on-orbit imagery, the Agent usually occupies a large portion of the field of view and provides relatively stable global contextual support, whereas the dominant ambiguity arises on the manipulated side of the interaction, especially when behavior discrimination depends on fine local geometry. Uniformly segmenting all roles would unnecessarily increase representation complexity on the less critical side of the event, while selective role segmentation concentrates modeling capacity on the semantically decisive part of robotic manipulation. Under constrained viewpoints, severe illumination variation, rigid attachments, and incomplete visibility, precise Instrument/Object delineation provides more discriminative evidence for operational behavior recognition than uniform box-level grounding and makes the contact structure of the interaction more explicit.

Specifically, this study makes three contributions. (1) We define Hybrid-AOI as a new role-grounding setting for on-orbit operational behavior recognition. Compared with conventional object detection and existing box-level AOI formulations, Hybrid-AOI retains the structured semantic schema (Agent, Instrument, Object, Verb) while reformulating role localization as a hybrid box-mask representation. The Agent remains box-grounded to preserve global role context, while the Instrument and Object are selectively mask-grounded to encode contact-sensitive manipulation geometry. (2) We develop a role-selective mask-aware interaction network for Hybrid-AOI. Built upon a transformer-based AOI decoder, the network integrates box-prompted role mask refinement with mask-aware interaction reasoning, so that verb prediction is jointly conditioned on role semantics, local interaction boundaries, and manipulated-object geometry. Compared with conventional detection-style AOI baselines, the proposed network is designed to better resolve operated-object ambiguity in observe-like, self-inspection, and



rigidly coupled arm-object scenes, where conventional box-level outputs often blur the distinction between the acting platform and the manipulated target. (3) We establish a Hybrid-AOI benchmark and evaluation protocol by augmenting existing role-verb annotations with hybrid box-mask supervision and by organizing systematic comparisons between box-only grounding and hybrid grounding. The benchmark builds on approximately 7.5K AOI images with three superclasses, six subclasses, and five representative operational behaviors. This organization preserves the space-domain focus of the task while providing morphology diversity and complementary interaction semantics for robustness analysis.

Based on this benchmark, we conduct structured evaluation of role grounding, role-verb matching, and ambiguity-sensitive operational behaviors under constrained space views. Beyond conventional AOI role-level matching, we introduce a hybrid evaluation protocol that jointly measures box-grounded Agent correctness together with mask-grounded Instrument/Object consistency, making manipulated-side boundary quality part of the operational behavior recognition criterion rather than a detached segmentation diagnostic. We further examine representative failure factors, including extreme illumination, partial visibility, constrained viewpoints, and rigid arm-object attachment, and analyze how selective role segmentation makes interaction evidence more localized, explicit, and interpretable at the manipulated side. By explicitly linking manipulated-side geometry to role-verb prediction, Hybrid-AOI provides a dedicated role-selective solution to coarse grounding in current AOI-based operational behavior recognition and supplies more reliable semantic evidence for the perception layer of on-orbit servicing, non-cooperative target handling, and debris-related robotic operations.

KEYWORDS

Behavior recognition; robotic arm operation; space robotic arms; human-object interaction

Acknowledgement: The authors would like to thank the reviewers and editors for their valuable comments and constructive suggestions, which have helped improve the quality of this paper.

Funding Statement: This research was supported by National Key R&D Program of China: Gravitational Wave Detection Project (No.2024YFC2207900) and National Natural Science Foundation of China (No.12472046).

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.