Original Manuscript ID: Access-2024-51179

Original Article Title: **“**Visibility Aware In-Hand Object Pose Tracking in Videos with Transformers”

To: IEEE Access Editor

Re: Response to reviewers

Dear Editor,

Thank you for allowing a resubmission of our manuscript, with an opportunity to address the reviewers’ comments.

We are uploading (a) our point-by-point response to the comments (below) (response to reviewers, under “Author’s Response Files”), (b) an updated manuscript with yellow highlighting indicating changes (as “Highlighted PDF”), and (c) a clean updated manuscript without highlights (“Main Manuscript”*)*.

Best regards,

Phan Xuan Tan, et al.

Reviewer#1, Concern # 1: I think they have addressed all concerns.

Author response: Thank you for your positive feedback. No specific changes are needed for this concern.

Reviewer#2, Concern # 1: The paper is very lengthy. The authors have unnecessarily referred historical papers with no relevance to current work. Consider omitting it.

Author response: We have reviewed the manuscript and identified sections where historical references could be trimmed without affecting the context or completeness of the discussion. Specifically, we have removed references [2, 4, 6, 8, 9, 10, 11, 12, 18, 19, 24, 28, 30, …] .

Reviewer#2, Concern # 2: The idea of introducing 'visibility base function is good' and appears to improve the performance noticeably. I see it as 'best effort approach' with current state of occlusion.

Author response: We appreciate your positive remarks about our visibility-aware module.

Reviewer#2, Concern # 3: On page 28 the authors say that the visibility estimation 'is trained' with binary cross entropy method'. it is unclear (not likely) on how they got the training data. In absence of it, binary-cross entropy is very unreliable.

Author response: Thank you for highlighting this concern. The ground truth visibility score for each frame was computed using the following process:

1. Ground Truth Poses and Camera Calibration: For each frame, we utilized the ground truth poses of the object and hand, along with the camera calibration parameters. These inputs allowed us to accurately project 3D models of the hand and object into the 2D image plane.
2. Generating 2D Masks: Using the projected 3D hand and object models, we created 2D binary masks representing the visible regions of both the object and the hand in the image frame.
3. Computing Visibility Scores:
    - First, we determined the total number of pixels in the object mask without considering the hand (i.e., the full projected object model).
    - Next, we computed the number of pixels in the object mask that remained visible when the hand model was also projected into the 2D frame, simulating occlusion.
    - The visibility score for the frame was then calculated as the ratio of the number of visible object pixels (after accounting for hand occlusion) to the total number of object pixels.

We ensure that the visibility scores accurately reflect the level of occlusion in each frame. These scores were then used as the ground truth for training the visibility estimation module with the binary cross-entropy loss. To enhance clarity, we have updated the manuscript to include this explanation in detail, along with a Figure to further aid understanding. Please see section IV. A and Figure 2.

Reviewer#2, Concern # 4: Rational behind Sigmoid activation, on page 28 and multi layered perception (MLP), on page 29 is not explained.

Author response: Thank you for pointing out the need for clarification. The Sigmoid activation function was chosen for the visibility-aware module because it outputs values in the range [0, 1], which naturally aligns with the interpretation of visibility as a probability score. Similarly, the multi-layer perceptron (MLP) was used as it effectively maps the high-dimensional feature space from the transformer outputs to the desired pose parameters, providing the flexibility to model complex relationships. We have revised the manuscript to explicitly detail these design choices and their theoretical motivations to enhance the reader's understanding.