Original Manuscript ID: Access-2024-41310

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,

Dinh-Cuong Hoang, et al.

Reviewer#1, Concern # 1: The paper are well organized and ready for publication.

Author response: Thank you for your positive feedback. No specific changes are needed for this concern. We will ensure all minor edits and enhancements suggested by other reviewers are incorporated to maintain the manuscript's overall clarity and quality.

Reviewer#2, Concern # 1: While the manuscript demonstrates improvements over existing methods, the comparison lacks depth. Consider expanding your discussion to include additional baselines or related work.

Author response: Author Response:

Thank you for pointing out this limitation in our manuscript. We appreciate your suggestion to expand the discussion and include additional baselines and related work to provide a more comprehensive comparison. To address this concern, we have revised the Related Work section to include a discussion of additional transformer-based methods [67-70]. These methods employ advanced transformer architectures for 6D object pose estimation, and the revised section highlights their strengths and limitations, particularly in addressing challenges such as occlusions and motion blur. Please see section II.C. TRANSFORMER-BASED METHODS.

Additionally, we enhanced the quantitative evaluation by including these methods in Table 3, comparing their performance to our proposed approach across three benchmark datasets: DexYCB, FPHAB, and HO-3D. The table now includes both standalone performances of these methods and their enhanced versions when combined with CosyPose [86] for cross-frame consistency. The accompanying discussion demonstrates the superior performance of our method in terms of AUC and AP metrics, emphasizing its robustness and computational efficiency. Please see section IV.E. COMPARISON WITH TRANSFORMER-BASED METHODS.

Reviewer#2, Concern # 2: Some equations and the overall algorithm are not adequately described. Ensure that each mathematical element is defined and that the algorithm is presented with enough detail for reproducible.

Author response: Thank you for highlighting this concern. To address the issue, we have revised the Method section to provide a more detailed explanation of the equations and the overall algorithm. Each mathematical element has been clearly defined, and we have ensured that the sequence of computations is presented in a step-by-step manner to enhance clarity and reproducibility. The updated section now provides a comprehensive description of visibility estimation and object pose prediction under heavy occlusion, explicitly detailing the intermediate computations, loss functions, and network components. Please see section “III.B. VISIBILITY-AWARE OBJECT POSE ESTIMATION UNDER OCCLUSIONS”.

Reviewer#2, Concern # 3: The methodology omits several key details, such as specific hyperparameters used, dataset preprocessing steps, and training settings. Providing this information would enhance the clarity and reproducibility of your work.

Author response: Thank you for highlighting this important concern. To address this concern, we have revised the section IV.A DATASETS to include specific preprocessing steps for all datasets used in the experiments, including image resizing, data augmentation techniques (e.g., random rotation, scaling, color jitter, and horizontal flipping), and the split ratios for training, validation, and testing. For each dataset, we have also outlined the criteria for constructing test sets with varying levels of occlusions to ensure consistency and robustness in evaluation. In the section IV. B. IMPLEMENTATION DETAILS, we have added detailed hyperparameter settings for the model. This includes the architecture of the spatial and temporal transformers (number of layers, number of attention heads, feed-forward network dimensions, and query dimensionality), learning rate schedules, batch size, and the optimizer used during training. Additionally,

we have specified the number of training epochs, learning rate decay schedule, adversarial training strategy, and the hardware configuration used for training and inference.

Author action:

- We revised section IV.A. DATASETS

"DexYCB Dataset [20]. The DexYCB dataset is highly suitable for evaluating hand-held object pose tracking in videos due to its comprehensive set of annotated hand-object interactions. The dataset includes over 582 video sequences capturing the dynamic manipulation of 20 different objects from the YCB dataset by 10 subjects. Each sequence provides 6D object poses, 3D hand poses, and RGB-D data, which are crucial for testing the robustness and accuracy of tracking algorithms in real-world conditions. The variety of objects, coupled with naturalistic hand movements, makes DexYCB an ideal benchmark for assessing how well a model can maintain accurate pose estimates over time, particularly in scenarios involving rapid hand movements and frequent occlusions. To prepare the dataset, we follow the benchmark setup S0 (default) as defined in [20]. The train split contains sequences from all 10 subjects, all 8 camera views, and all 20 grasped objects. We ensure that no sequences are shared between the training, validation, and test splits. Images are resized to 256 × 256 pixels, and data augmentation techniques, including random rotation, scaling, and brightness adjustments, are applied during training to enhance model generalization.

FPHAB Dataset [22]. The First-Person Hand Action Benchmark (FPHAB) dataset is particularly relevant for evaluating pose tracking in first-person view scenarios. It offers 1175 RGB-D video sequences across 45 different action classes, all annotated with 3D hand poses. This dataset is valuable for testing hand-held object pose tracking models because the egocentric viewpoint introduces challenges such as motion blur, self-occlusion, and varying lighting conditions, which are common in real-world applications. By providing a diverse set of actions and interactions, FPHAB enables the assessment of a model's ability to track objects consistently despite the inherent complexities of first-person perspectives. In our setup, the dataset is preprocessed by resizing images to 256 × 256 pixels. We split the dataset into approximately 70,000 frames for training and 30,000 frames for testing. Data augmentation, including random flipping, cropping, and color jitter, is applied to the training set. To evaluate the model's robustness to occlusions, more than 70% of the test set images involve occlusions, with at least 50% of those under heavy occlusion.

HO-3D Dataset [79]. This dataset is specifically designed to evaluate 3D hand-object pose estimation, making it an excellent benchmark for hand-held object pose tracking in videos. With over 80,000 annotated frames, this dataset captures intricate hand-object interactions, often under challenging conditions like severe occlusions and complex background scenes. The sequences in HO-3D are recorded in real-world environments, providing a realistic testbed for assessing the effectiveness of pose tracking algorithms. The dataset's emphasis on naturalistic hand movements and object manipulation in cluttered environments is critical for evaluating how well a model can maintain accurate pose estimates across frames when faced with partial or full occlusions. We preprocess the dataset by resizing all images to 256 × 256 pixels and applying data augmentation such as random rotation, scaling, and brightness adjustments to enhance diversity in the training data. The dataset is split into approximately 65,000 frames for training and 10,000 frames for testing, with over 70% of the test set images involving occlusions and at least 50% of those under heavy occlusion."

- We revised section IV.B. IMPLEMENTATION DETAILS

"Our implementation employs ResNet-50 [77] combined with ROIAlign [80] to extract object features at an output resolution of 32×32 with 256 feature channels. The spatial transformer module, which includes both the encoder and decoder, is built with 3 layers of multi-head self-attention. Each attention layer contains 8 heads, with a feed-forward network dimension of 256. The transformer decoder uses query vectors with a dimensionality of 256, and the temporal transformer shares the same architecture as the spatial transformer. The ResNet-50 backbone is initialized with ImageNet-pretrained weights, while the transformer components are trained from scratch using an end-to-end approach. Training is performed over 120 epochs using the Adam optimizer with an initial learning rate of 1×10^{-5}. The learning rate is reduced by 30% every 20 epochs. To enhance robustness, adversarial training is employed, where the transformers are alternately updated at each training step. Batch size is set to 64, and training is conducted across 8 NVIDIA RTX 2080Ti GPUs. During training, the dataset is augmented with random rotation, scaling, color jitter, and horizontal flipping to improve generalization. The model is evaluated using standard metrics, including ADD(- S) and AUC for 6D pose estimation, to measure accuracy across different datasets. For inference, the trained model is deployed on a single NVIDIA RTX 2080Ti GPU."

Reviewer#2, Concern # 4: Some figures and plots lack sufficient annotations or context. Enhance these visual aids to ensure they are self-explanatory and clearly support the narrative.

Author response: Thank you for pointing this out. We have enhanced the clarity of figures and plots to ensure they are fully self-explanatory. Please see Figure 2 and 3.

Reviewer#2, Concern # 5: Discuss the computational demands of your approach, particularly with regard to the transformer architecture, and propose strategies for real-time implementation.

Author response: Thank you for raising this important concern. We acknowledge the need to address the computational demands of our approach and to discuss strategies for enabling real-time implementation. In response, we have added a new section IV.G. RUNTIME ANALYSIS AND COMPUTATIONAL EFFICIENCY to the manuscript. This section provides a detailed analysis of our method's runtime performance and computational efficiency.

Author action: Added the new section IV.G. RUNTIME ANALYSIS AND COMPUTATIONAL EFFICIENCY

"Our method achieves a balance between computational efficiency and accuracy, making it suitable for real-world applications. The average inference time for a single frame is 66ms, measured on a single NVIDIA RTX 2080Ti GPU, corresponding to a processing speed of approximately 15 frames per second (FPS). This runtime includes computations for both spatial and temporal transformers as well as the visibility-aware module for handling occlusions. While transformer-based architectures are inherently more computationally demanding than simpler models, our framework incorporates optimizations that reduce latency without sacrificing performance. Several factors contribute to the efficiency of our approach. The spatial transformer encoder utilizes positional embeddings and multi-head self-attention layers to process frame-wise features efficiently while maintaining a fixed computational budget. Similarly, the temporal transformer encoder integrates long-range dependencies across sequences in parallel, avoiding the iterative bottlenecks of sequential processing. The visibility-aware module further enhances computational efficiency by dynamically focusing on frames with low visibility, thereby bypassing redundant computations for highly visible frames. These optimizations collectively ensure that our model operates effectively within real-time constraints.

To further enhance runtime performance, we propose strategies such as model compression, which includes pruning and quantization to reduce the size and computation requirements of the model. For instance, lowering the precision of weights and activations to 16-bit floating point (FP16) can reduce memory usage and increase GPU throughput. Additionally, spatial and temporal transformer computations can be executed in parallel pipelines to leverage modern multi-core GPUs, minimizing latency through concurrent processing. Frame grouping into mini-batches can also benefit GPU optimizations, allowing multiple frames to be processed in parallel. Furthermore, selectively invoking the visibility-aware module only for frames with low visibility scores reduces unnecessary computation, and precomputing static spatial features for non-dynamic sequence elements can offload redundant operations."

Reviewer#2, Concern # 6: Provide a clearer distinction between your visibility-aware module and similar mechanisms in the literature to highlight the unique contributions of your work.

Author response: Thank you for your insightful feedback. To address this concern, we have revised the Related Work section to provide a more detailed comparison and highlight the novel aspects of our approach. The revised section now clearly explains how existing methods utilizing visibility information, such as those applied in scene reconstruction and multi-view stereo, primarily focus on static scenes. In contrast, our module is specifically designed to handle the challenges of 6D object pose estimation in dynamic scenarios, such as occlusions, motion blur, and rapid hand movements. We emphasize that our module dynamically estimates visibility scores in real time for each frame, leveraging these scores to adjust pose predictions dynamically. Moreover, we have highlighted the distinctive use of a Pose Transformer in our framework, which aggregates pose information from neighboring visible frames using cross-attention mechanisms. This innovative approach ensures accurate pose estimation even under severe occlusions by fusing temporal information from both occluded and visible frames. This combination of dynamic visibility estimation and temporal aggregation sets our method apart from existing visibility-aware mechanisms, which often lack the capacity to address temporal dependencies or real-time adjustments. These revisions enhance the manuscript by explicitly delineating the contributions of our visibility-aware module, ensuring that its novelty and impact are well-articulated. We believe this addresses the reviewer's concerns comprehensively and strengthens the manuscript's contribution to the field.

Author action: - We revised Related Work section

- *"While existing transformer-based methods demonstrate promising results, their performance often deteriorates in scenarios involving occlusions, motion blur, and rapid hand movements, which complicate pose estimation. To address these challenges, we propose a novel transformer-based neural network that explicitly incorporates object visibility and motion information to leverage neighboring frames for predicting the poses of occluded objects. To the best of our knowledge, this is the first study to introduce a visibility- aware module specifically for object pose estimation. Although visibility information has been utilized in related tasks such as scene reconstruction [72] and multi-view stereo [73], [74], our approach differs significantly in its design and application. Unlike existing visibility-aware mechanisms, which primarily focus on static scene modeling, our module dynamically estimates and adjusts pose predictions in real-time based on an object's visibility in each frame. This is achieved by generating a visibility score through a series of fully connected layers and employing it to identify frames with low visibility. The module then aggregates pose information from more visible frames using a Pose Transformer, which leverages cross-attention mechanisms to fuse temporal information from both occluded and visible frames. This dynamic aggregation ensures accurate pose estimation even under severe occlusions, enabling robust performance in highly dynamic and challenging environments."*

Reviewer#2, Concern # 7: While the references are generally relevant and up-to-date, consider including additional recent works that directly address visibility-aware tracking or transformer-based object pose estimation to strengthen the contextual foundation.

Author response:

- We agree with your suggestion. In the revised manuscript, we have updated the references to include more recent works that address visibility-aware tracking and transformer-based pose estimation. Please see section II.C.

Reviewer#2, Concern # 8: A section discussing the limitations of your approach, such as scalability to larger datasets or performance in real-time scenarios, would provide a more balanced and transparent evaluation of your work.

Author response: In the revised manuscript, we have added a new section IV.H. LIMITATIONS AND FUTURE WORK.

"While our approach demonstrates significant improvements in object pose tracking under occlusions, there are limitations that warrant discussion. One of the primary challenges is scalability to larger datasets. Although our method has been evaluated on benchmarks such as DexYCB, FPHAB, and HO-3D, the computational demands of transformer-based architectures may pose challenges when scaling to significantly larger datasets with higher resolution images or longer video sequences. The increased sequence length could lead to higher memory consumption and slower training times due to the quadratic complexity of attention mechanisms in transformers. Efficient adaptations, such as sparse attention mechanisms or hierarchical transformers, could be explored in future work to address this limitation.

Another limitation lies in real-time performance. While our approach achieves a processing speed of 15 FPS on an NVIDIA RTX 2080Ti GPU, this may not meet the stringent requirements of some real-time applications, particularly in resource-constrained environments or systems requiring higher frame rates. Although we have proposed strategies such as model compression, selective module invocation, and parallel processing to improve runtime efficiency, implementing and validating these optimizations in practical scenarios remains an area for future investigation.

Additionally, the reliance on labeled datasets for training and evaluation presents another limitation. Current datasets used in our experiments provide well-annotated frames for training and testing; however, real-world deployment may involve objects or environments not represented in these datasets. Exploring methods for unsupervised or semi-supervised learning could enhance the adaptability of our framework to new settings without requiring extensive labeled data."

Reviewer#2, Concern # 9: While the manuscript is written in readable English, there are minor grammatical errors and phrasing inconsistencies. A thorough language review is recommended.

Author response: We have conducted a thorough language review to correct any inconsistencies.