QORT-Former: Query-optimized Real-time Transformer for Understanding Two Hands Manipulating Objects

Elkhan Ismayilzada[1, 2*†], **MD Khalequzzaman Chowdhury Sayem**[1*], **Yihalem Yimolal Tiruneh**[1], **Mubarrat Tajoar Chowdhury**[1], **Muhammadjon Boboev**[1], **Seungryul Baek**[1‡]

[1]UNIST, Ulsan, South Korea
[2]Michigan State University, MI, USA
{elkhan, khalequzzamansayem, yihalemyimolal, mubarrattajoar, muhammad, srbaek}@unist.ac.kr

Abstract

Significant advancements have been achieved in the realm of understanding poses and interactions of two hands manipulating an object. The emergence of augmented reality (AR) and virtual reality (VR) technologies has heightened the demand for real-time performance in these applications. However, current state-of-the-art models often exhibit promising results at the expense of substantial computational overhead. In this paper, we present a query-optimized real-time Transformer (QORT-Former), the first Transformer-based real-time framework for 3D pose estimation of two hands and an object. We first limit the number of queries and decoders to meet the efficiency requirement. Given limited number of queries and decoders, we propose to optimize queries which are taken as input to the Transformer decoder, to secure better accuracy: (1) we propose to divide queries into three types (a left hand query, a right hand query and an object query) and enhance query features (2) by using the contact information between hands and an object and (3) by using three-step update of enhanced image and query features with respect to one another. With proposed methods, we achieved real-time pose estimation performance using just 108 queries and 1 decoder (53.5 FPS on an RTX 3090TI GPU). Surpassing state-of-the-art results on the H2O dataset by 17.6% (left hand), 22.8% (right hand), and 27.2% (object), as well as on the FPHA dataset by 5.3% (right hand) and 10.4% (object), our method excels in accuracy. Additionally, it sets the state-of-the-art in interaction recognition, maintaining real-time efficiency with an off-the-shelf action recognition module.

Project Page — https://kcsayem.github.io/QORT-Former/

Introduction

Estimating poses and actions in egocentric videos involving two hands and an object is crucial for applications like AR, VR, and HCI. Significant progress has been made in hand pose estimation (Zimmermann and Brox 2017; Baek, Kim, and Kim 2018, 2019, 2020; Kim, Kim, and Baek 2021; Garcia-Hernando et al. 2018; Wang et al. 2020; Moon et al. 2020; Lin, Wang, and Liu 2021; Chen et al. 2021, 2022; Lee et al. 2023) and object 6D pose estimation (Xiang et al. 2017; Li, Wang, and Ji 2019; Tekin, Sinha, and Fua 2018; Kehl et al. 2017; Chen et al. 2020; Iwase et al. 2021; Xu et al. 2022), which were previously addressed independently. Given that hands frequently interact with objects, there's a growing demand for methods that jointly estimate the poses of both hands and objects, along with hand interaction classes (Hasson et al. 2020, 2019; Armagan et al. 2020; Liu et al. 2021; Cho et al. 2023; Fan et al. 2024, 2025). Recent frameworks like H2OTR (Cho et al. 2023) have shown impressive performance in estimating two-hand and object poses, object types, and hand interaction classes in a single Transformer-based framework. However, H2OTR's use of the deformable DETR architecture (Zhu et al. 2021) for frame-by-frame pose estimation leads to significant compu-

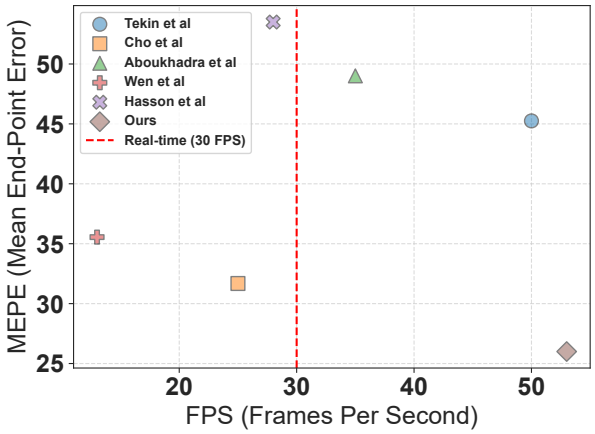

Figure 1: Comparisons to competitive state-of-the-art algorithms (Cho et al. 2023; Tekin, Bogo, and Pollefeys 2019; Aboukhadra et al. 2023; Wang, Mao, and Li 2023; Hasson et al. 2020) on the two hands and an object pose estimation task on an RTX 3090TI GPU. Even with the Transformer architecture, we achieved the fastest speed (53.5 FPS) while obtaining the best accuracy among the methods.

*These authors contributed equally.
†This work was conducted when Elkhan Ismayilzada was graduate student at UNIST.
‡Corresponding author.

tational overhead, making it unsuitable for real-time applications, since pose estimation accounts for over 97% of the total inference time.

This paper aims to achieve real-time 3D pose estimation for two hands and an object by improving both computational efficiency and accuracy. In deformable DETR (Zhu et al. 2021), the encoder consumes 49% of the overall GFLOPs while contributing only 11% of the AP (Lin et al. 2022), mainly due to the multi-scale deformable attention mechanism and numerous decoder layers. Inspired by recent efficient encoder designs (Cheng et al. 2022b; He et al. 2023; Lv et al. 2023), we developed a feature decoder based on the pyramid pooling module (PPM) (Zhao et al. 2017) to expand receptive fields and reduce computational costs. We also optimized the number of queries by introducing hand-object queries tailored for estimating poses, explicitly utilizing 2D locations of hands and objects. To minimize reliance on heavy feature decoder, we update enhanced features and query features within the Transformer decoder. Unlike H2OTR (Cho et al. 2023), which randomly initializes object queries, we select queries with high semantics from multi-scale feature maps. Despite using fewer queries (108 compared to H2OTR's 300), our method yields higher-quality queries, enhancing both speed and accuracy (Refer to Fig 3).

An essential facet of two hands and an object pose estimation is the complex phenomenon of grasping, which involves intricate hand configurations and contact regions between the hands and the object. Despite progress in estimating hand-object interaction poses, identifying contact points remains challenging. To address this, we estimate a contact map from the feature decoder's output and integrate it into our query features before passing them to the Transformer decoders.

To reduce dependence on a heavy feature decoder, in our proposed decoder both image and query features are co-optimized for enhanced hand-object pose estimation. We combine query features for the hands and object with auxiliary background queries. Unlike traditional methods that update only query features, our decoder refines image features and query features through a three-step process: 1) cross-attention improves spatial and contextual relationships, 2) location-based enhancement focuses on key areas around the hands and object, and 3) further refinement captures fine details like finger joints and contact points, leading to more accurate pose and class estimation.

Our proposed modification ensures superior performance with just 100 hand-object queries (+ 8 auxiliary queries) and 1 decoder, compared to the 300 randomly initialized queries and 6 decoders in H2OTR (Cho et al. 2023); which enables us to achieve real-time performance at 53.5 FPS, significantly outperforming the 26 FPS in H2OTR (Cho et al. 2023) on an RTX 3090TI GPU. We also demonstrate the state-of-the-art two hands and an object pose and action recognition performance on H2O (Kwon et al. 2021) and FPHA (Garcia-Hernando et al. 2018) datasets. To summarize, our main contributions are as follows:

- We present the query-optimized real-time Transformer (QORT-Former), to the best of our knowledge, the first Transformer-based real-time framework for 3D pose estimation of two hands and an object.
- For the real-time speed, we proposed to constrain the query numbers (as 108) and the number of decoders (as 1). See Figure 1 for FPS vs. Error comparison with other methods.
- For robust accuracy with a reduced number of queries and decoders, we propose a novel method of dividing object queries into three sections: a left hand, a right hand, and an object to optimize the location of queries. We also introduce the incorporation of contact map features into query features, enhancing the query's awareness of contact dynamics in two hands and an object interactions.
- To reduce the dependency on heavy feature decoders, we introduce a three-step feature update in the transformer decoder, simultaneously constraining the decoder count.
- Our proposed method outperforms current state-of-the-art by an impressive margin (5.3%-27.2%) in pose estimation on H2O (Kwon et al. 2021) and FPHA (Garcia-Hernando et al. 2018) datasets while ensuring real-time performance (53.5 FPS on an RTX 3090TI GPU).

Related Works

In this section, we discuss the previous related works in the domain of hand-object pose estimation. Building upon the success of transformers (Vaswani et al. 2017) and the subsequent emergence of ViT (Dosovitskiy et al. 2021), numerous transformer-based methodologies (Carion et al. 2020; Wang et al. 2022; Yao et al. 2022; Li et al. 2022; Liu et al. 2022; Cha et al. 2024) have been successfully applied across multiple vision-related tasks (Han et al. 2022), including hand pose estimation (Huang et al. 2020; Jiang et al. 2023; Fu et al. 2023; Zhang and Kong 2024; Pavlakos et al. 2024) and hand-object pose estimation (Hampali et al. 2022; Liu et al. 2021; Cho et al. 2023). Hampali et al. introduced a transformer-based 3D hand-object pose estimation methodology that performs self-attention between 2D hand-keypoint features. Fu et al. propose deformer, a dynamic fusion Transformer that leverages spatial relationships within an image and temporal correlations between nearby frames to learn hand deformations. More recently, A2J-Transformer (Jiang et al. 2023) extends the state-of-the-art depth-based 3D single hand pose estimation method A2J (Xiong et al. 2019) to the RGB domain under interacting hand conditions. (Jiang et al. 2023) enhances A2J (Xiong et al. 2019) by incorporating non-local encoding-decoding framework of transformers, enabling global spatial context awareness and adaptive feature learning for each anchor point located in 3D space. Cho et al. introduce a Transformer-based unified framework to estimate the poses of two hands and an object, and their interaction classes in a single inference step. Although this model is state-of-the-art in terms of accuracy, it does not perform in real-time. The major drawback in terms of speed in Cho et al.'s work is in the pose estimator network, which takes more than 97% of the total inference time. This is contributed by factors such as a high number of queries, heavy encoder and the use of a large number of decoder layers. To reduce the complexity of encoder, we

employ PPM-FPN (Cheng et al. 2022b) and use one feature map instead of using all three feature maps as in (Cho et al. 2023). But this makes our encoder less feature enriched compared to (Cho et al. 2023). To tackle this, we propose two simple but effective modifications. Firstly, we construct semantically meaningful queries. Which are then divided into left-hand, right-hand, and object categories, similar to (Hampali et al. 2022). However, we deviate by using a dedicated query proposal network to suggest locations based on semantic relevance, eliminating the need for non-maximal suppression for reduced inference time and improved efficiency. Another significant aspect of hand-object interaction is the point of contact between hands and objects (Karunratanakul et al. 2020; Yang et al. 2021). To enrich the query, in our work, we combine the contact map features along with the semantic features to be able to catch intricate details while hands and objects are in contact. Additionally, since we utilize a single feature map, thereby creating discrepancies with conventional decoders employed in complex models (Cho et al. 2023; Cheng et al. 2022a), modifications to the decoder become necessary. Therefore, we opt for a three-step image and query features co-optimization strategy in the decoder, involving cross-attention twice. While this three-step update increases the decoder's complexity, it enables achieving comparable performance with a smaller number of decoder layers compared to other models. Combining all the modifications allows us to get state-of-the-art performance and real-time inference speed.

Method

We propose the hand-object interaction recognition framework that inputs an RGB image and outputs 3D poses of two hands and an object.

Query-Optimized Real-Time Transformer

To tackle the challenging task of recognizing 3D poses of two hands and an object, we proposed the real-time Transformer-based framework, query-optimized real-time Transformer (QORT-Former). In this section, we will explain each component of our method in detail.

Backbone. Our model uses the ImageNet pre-trained ResNet-50 (He et al. 2016) architecture to extract features from an input image $\mathbf{x}$. Specifically, we acquire three distinct feature maps $\mathbf{f}$, each of resolutions $1/8$, $1/16$, and $1/32$ of the input image, respectively. The projection of these feature maps involves a 1×1 convolutional layer, resulting in them being represented with 256 channels. Subsequently, these projected feature maps are used as input to the feature decoder for further processing.

Feature Decoder. The feature decoder takes in the projected feature maps $\mathbf{f}$ and generates enhanced multi-scale contextual feature maps $\mathbf{f}' = \{\mathbf{E}_3 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times256}, \mathbf{E}_4 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times256}, \mathbf{E}_5 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times256}\}$ where H and W are the height and the width of the input image, respectively. As the feature decoder, we use the PPM-FPN (Cheng et al. 2022b), which employs pyramid pooling module (PPM) (Zhao et al. 2017) to enlarge the receptive fields.

Query Division Block. The role of object queries in Transformer architecture is paramount. Unlike H2OTR (Cho et al. 2023), where the object queries are randomly initialized, we propose to select queries with high semantics from underlying multi-scale feature maps by adding a query proposal module on top of the feature map to output the class probability prediction for each pixel. The output of the query proposal module is a $(N_c+1)-$dimensional probability simplex, where N_c is the number of classes and one dimension is added for the "no object" class. We use three query proposal modules, two for classifying each hand and one for objects. As a result, each pixel is classified with three classifiers. In the context of hand-object pose estimation, classifying left and right hands is especially difficult due to the similarities in their underlying feature maps, and this query proposal module aids in reducing the uncertainty. Given the classification probability from the respective module, we adopt a strategy wherein the top N_l location of the pixel corresponds to the left hand, the top N_r location to the right hand, and the top N_o location to objects. Here, N_l, N_r, and N_o represent the number of queries designated for the left and right hands and objects, respectively. Subsequently, we generate $\boldsymbol{Q}_{\mathrm{l}} \in \mathbb{R}^{N_l\times256}$, $\boldsymbol{Q}_{\mathrm{r}} \in \mathbb{R}^{N_r\times256}$ and $\boldsymbol{Q}_{\mathrm{o}} \in \mathbb{R}^{N_o\times256}$ query features for each selected location utilizing our $\mathbf{E}_4$ feature map from the feature decoder, where $\boldsymbol{Q}_{\mathrm{l}}$, $\boldsymbol{Q}_{\mathrm{r}}$, $\boldsymbol{Q}_{\mathrm{o}}$ denote query features for left hand, right hand and objects. During the training phase, we implement a matching-based Hungarian (Kuhn 1955) loss to supervise each module. This involves leveraging class predictions along with a location cost, indicating whether the pixel is located in the region of interest for the target object. Further details regarding the loss calculation are elaborated in the Loss function section.

Contact Estimator. In the context of 3D hand-object pose estimation, a crucial aspect revolves around the nuanced dynamics of grasping—a foundational element in the interaction between two hands and an object. This complex process involves sophisticated hand configurations leading to contact zones between two hands and an object. Consequently, the integration of contact zone information into the pose estimation of interacting two hands and an object holds substantial promise for improving accuracy. With the goal of achieving the objective, our initial step involves constructing a contact map $\mathbf{C}_{\mathbf{M}} \in \mathbb{R}^{2\times778\times1}$ for two hands encoding the vertex regions close to 1 if they are contacting with an object, following the approach outlined in (Cho et al. 2023). For contact map estimation, we utilize the mid-sized (*i.e.*, $\mathbf{E}_4$ in Fig. 2.) feature map from the feature decoder. Opting for the mid-sized feature map aims to balance computational efficiency with information retention. Further details regarding the loss calculation for the contact estimator can be found in Loss function section. Once the contact maps of two hands are estimated, we add them as the contact map feature to object queries, to further improve the semantics of integrated query features.

QORT Transformer Decoder. Upon obtaining query features $\boldsymbol{Q}_{\mathrm{l}}$, $\boldsymbol{Q}_{\mathrm{r}}$, and $\boldsymbol{Q}_{\mathrm{o}}$ with the contact map features, we concatenate them to form combined query features, $\boldsymbol{Q}_{\mathrm{a}} \in \mathbb{R}^{(N_l+N_r+N_o)\times256}$. These are then integrated with $\boldsymbol{Q}_{\mathrm{b}} \in \mathbb{R}^{N_b\times256}$ auxiliary query features, where N_b is the number of auxiliary queries, strategically designed to facilitate

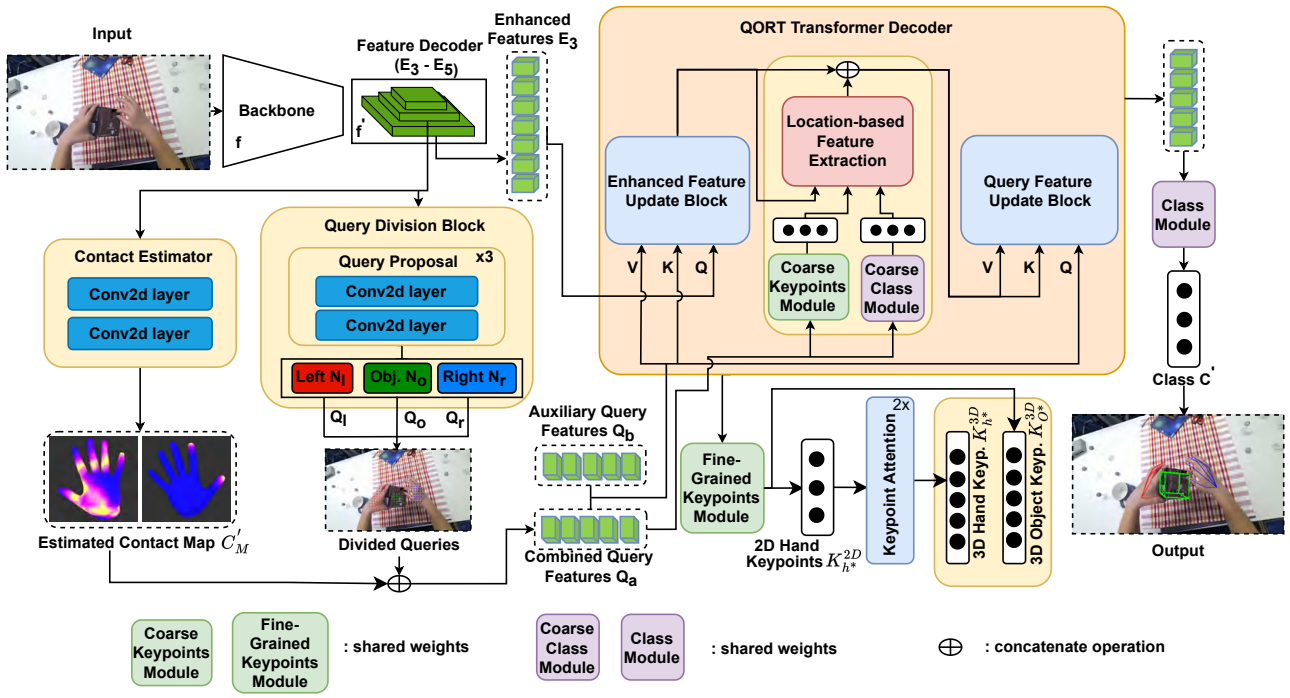

Figure 2: Our architecture begins with extracting a multi-scale feature $\mathbf{f}$ from an image using ResNet-50 (He et al. 2016), which is then refined into $\mathbf{f}'$ by our feature decoder. We propose queries aligned with hand and object locations, incorporating contact map features, while auxiliary queries capture background details. In the QORT Transformer decoder, enhanced and query features undergo three steps: 1) Cross-attention updates the enhanced feature based on integrated query features in Enhanced Feature Update Block, 2) Location-based Feature Extraction module adds feature maps of 3×3 patches around coarse 2D hand and object keypoints to Enhanced Feature, and 3) Cross and self-attention layers update the integrated query features based on updated enhanced features in Query Feature Update Block. Finally, the heads estimate poses for both hands and the object.

the aggregation of background features and provide general image-independent cues during the update process. The integrated queries are fed into the QORT Transformer decoder alongside the flattened enhanced features, $\mathbf{E}_3$.

Unlike traditional Transformer architectures (Cho et al. 2023; Cheng et al. 2022a), where only query features are updated, our proposed decoder co-optimizes both image features (Enhanced Features, $\mathbf{E}_3$) and query features. This reduces reliance on heavy encoders and allows for a lightweight feature decoder. The co-optimization process involves three steps:

- Enhanced Feature ($\mathbf{E}_3$) Update: Enhanced features and integrated query features undergo cross-attention in the "Enhanced Feature Update Block", where enhanced features act as queries and integrated query features as keys and values. This refines the holistic representation of the spatial configuration and contextual relationships between the hands and the object in the scene.
- Location-Based Enhancement: The enhanced feature is further refined to focus more on the areas around the hands and object. First, coarse 2D keypoints and probability simplex for each class are estimated using the "coarse keypoints module" and "coarse class module" from the combined query features, Q_a. The keypoints with the highest probability for both hands and the object are then fed into the "Location-based feature extraction" module along with the updated $\mathbf{E}_3$. This module generates feature maps of 3×3 patches around the coarse 2D keypoints, which are then concatenated with the updated $\mathbf{E}_3$. This procedure allows to refine the integrated query features with added attention around the area of both hands and object in the next step.
- Query Feature Update: In the "Query Feature Update Block", integrated query features are further refined using cross-attention (with updated $\mathbf{E}_3$ as keys and values) and self-attention layers. This block allows to capture the fine-grained details such as finger joints, palm surface characteristics, object contact points, and specific regions on the hand and object that contribute to a detailed understanding of their poses. Leveraging the refined query features, we proceed to estimate target classes, hand poses, and object poses.

Prediction. On top of the refined query features at each decoder layer, we apply three 3-layer MLPs and a linear layer to output fine-grained 2D keypoints for left and right hands, 3D object poses and target classes, respectively. Notably, the same sets of linear layers with shared weights were employed to estimate coase 2D keypoints for hands & object

Method	H2O			FPHA	
	Left	Right	Obj. (L/R)	Right	Obj.
Hasson et al. (2020)	39.6	41.9	67.5/66.1	18.0	22.3
Tekin et al. (2019)	41.4	38.9	48.1/52.6	15.8	24.9
Kwon et al. (2021)	41.5	37.2	47.9	-	-
Wen et al. (2023)	35.0	36.1	-	15.8	-
Aboukhadra1 et al. (2023)	36.8	36.5	73.9	-	-
Cho et al. (2023)	24.4	25.8	45.2	15.0	21.0
Ours	**20.1**	**19.9**	**32.9**	**14.2**	**18.8**

Table 1: Mean End-Point Error comparison across SOTA pose estimation pipelines have been shown. Experiments are performed on test sets of H2O and FPHA datasets. Single-hand methodologies (Hasson et al. 2019; Tekin, Bogo, and Pollefeys 2019) are tested for left and right hand object interactions separately. Our method outperforms others by a significant margin. Best results are in bold.

Method	H2O
	Accuracy
C2D (Wang et al. 2018)	70.7
I3D (Carreira and Zisserman 2017)	75.2
SlowFast (Feichtenhofer et al. 2019)	77.7
Tekin et al. (2019)	68.9
Kwon et al. (2021)	79.3
HTT (Wen et al. 2023)	86.4
H2OTR (Cho et al. 2023)	90.9
Ours	**91.3**

Table 2: Comparison of Top-1 Accuracy for Hand-Object Interaction Recognition in H2O dataset.

and coarse classification in the previous step in the decoder. 2D keypoints of object were extracted from estimated 3D keypoints before further processing in the Location-based Feature Extraction module.

Motivated by the performance and high efficiency of graph-oriented attention in 3D hand pose estimation (Zhao, Wang, and Tian 2022), by leveraging the skeleton structure of hands to be as graph-structured data, we further refine hand poses by passing the 2D input keypoints through a series of keypoint attention and ChebGConv (Zhao, Wang, and Tian 2022) layers. The keypoint attention block integrates multi-head attention and graph convolution layers, while the ChebGConv block incorporates Chebyshev graph convolutional layers. This strategic combination exploits the inherent connectivity among keypoints, enabling the model to effectively capture and estimate 3D poses from 2D coordinates. This methodology proves to be a robust and highly efficient solution in (Zhao, Wang, and Tian 2022) for overcoming the inherent challenges associated with direct 3D pose prediction from 2D feature maps.

Loss Function

The overall loss function $\mathcal{L}$ for training QORT-Former can be written as follows:

$$\mathcal{L} = \lambda_{\text{CE}}\mathcal{L}_{\text{CE}} + \lambda_{\text{KP}}\mathcal{L}_{\text{KP}} + \lambda_{\text{QP}}\mathcal{L}_{\text{QP}} + \lambda_{\text{CM}}\mathcal{L}_{\text{CM}} \tag{1}$$

where, $\mathcal{L}_{\text{CE}}$ is the classification loss for the final classification head, $\mathcal{L}_{\text{QP}}$ is the query proposal loss, $\mathcal{L}_{\text{CM}}$ is the contact map estimation loss and $\mathcal{L}_{\text{KP}}$ is the loss for estimating keypoints of two hands and an object. λ_{CE}, λ_{QP}, λ_{CM} and λ_{KP} are the hyper-parameters to balance the weights of losses during the training.

Classification Loss. For classification loss, $\mathcal{L}_{\text{CE}}$, of our model, similar to (Cho et al. 2023) we use Hungarian algorithm (Kuhn 1955) to match our predicted output classes, $\boldsymbol{C}'$ to the ground truth classes, $\boldsymbol{C}$. The output of this classification head forms a probability simplex for each class, encompassing left hand, right hand, and individual object categories. To train the model effectively, we apply cross-entropy loss, aiming to maximize the likelihood of the true class occurrences within the scene.

Keypoints Estimation Loss. We apply L1 loss between predicted keypoints and ground-truth keypoints as follows:

$$\begin{aligned}\mathcal{L}_{\text{KP}} = \|\boldsymbol{K}_{\text{h}}^{\text{3D}} - \boldsymbol{K}_{\text{h}^*}^{\text{3D}}\|_1 \quad &+\|\boldsymbol{K}_{\text{h}}^{\text{2D}} - \boldsymbol{K}_{\text{h}^*}^{\text{2D}}\|_1 \\ &+\|\boldsymbol{K}_{\text{o}}^{\text{3D}} - \boldsymbol{K}_{\text{o}^*}^{\text{3D}}\|_1\end{aligned} \tag{2}$$

where $\boldsymbol{K}_{\text{h}}^{\text{3D}} \in \mathbb{R}^{2\times21\times3}$, $\boldsymbol{K}_{\text{h}}^{\text{2D}} \in \mathbb{R}^{2\times21\times2}$ and $\boldsymbol{K}_{\text{o}}^{\text{3D}} \in \mathbb{R}^{21\times3}$ denote ground truth 3D hand keypoints, 2D hand keypoints and 3D object keypoints, respectively. And, $\boldsymbol{K}_{\text{h}^*}^{\text{3D}}$, $\boldsymbol{K}_{\text{h}^*}^{\text{2D}}$ and $\boldsymbol{K}_{\text{o}^*}^{\text{3D}}$ represent the estimated 3D hand keypoints, 2D hand keypoints and 3D object keypoints, respectively.

Query Proposal Loss. We use the binary cross-entropy loss for left and right hand query proposal modules and cross-entropy loss for supervising object query proposal module. We apply binary classification loss to the left and right hand query proposal modules as for both of them, the objective is to classify hand or "no object". For each query, we find the optimal match using the Hungarian algorithm (Kuhn 1955) by computing classification and location costs (*i.e.*, whether the estimated query position is in the region of the object of interest or not). To make sure queries from each segment are focused on a different area of the region of interest, we do not match with another query location to a previously matched ground truth location. The query proposal loss is defined as follows,

$$\mathcal{L}_{\text{QP}} = \mathcal{L}_{\text{qo}} + \mathcal{L}_{\text{ql}} + \mathcal{L}_{\text{qr}} \tag{3}$$

where $\mathcal{L}_{\text{ql}}$, $\mathcal{L}_{\text{qr}}$, $\mathcal{L}_{\text{qo}}$ are proposal losses for the left and right hand and an object query proposal modules, respectively. Our method often fails to distinguish between the highly similar features of the left and right hands. To address the challenge and increase the likelihood of capturing features for both hands, we segment a set of queries for each specific region—left hand, right hand, and object and to channel each set of queries towards its corresponding area of interest, we employ the respective query proposal losses.

Contact Map Estimation Loss. To guide our model to estimate the ground truth contact map, we obtain the left-hand contact map, $\mathbf{C}_{\text{m}}^{\text{left}} \in \mathbb{R}^{778\times1}$ and right-hand contact map, $\mathbf{C}_{\text{m}}^{\text{right}} \in \mathbb{R}^{778\times1}$ from hands and object meshes by following (Cho et al. 2023). For our task, we combine the contact map of both hands to obtain the ground truth contact map, $\mathbf{C}_{\text{M}} \in \mathbb{R}^{2\times778\times1}$. For the loss, we calculate the L1 loss between the predicted contact map, $\mathbf{C}'_{\text{M}} \in \mathbb{R}^{2\times778\times1}$, and the ground truth contact map, $\mathbf{C}_{\text{M}}$. Therefore, the contact map

Figure 3: Query location visualization: (a) **Left:** query locations of H2OTR (Cho et al. 2023), employing 300 queries. Notably, a substantial amount of queries are distributed in backgrounds. (b) **Middle:** Our hand-object query locations w/o Query division block. Due to feature similarities between two hands, a considerable number of queries concentrate on the left hand than the right hand, which reduces the accuracy of the right hand. (c) **Right:** Our hand-object query locations. Queries for left and right hands are highlighted in red and blue, respectively. Queries for objects are denoted as green. The query proposal loss ensures that each query concentrates on its specific region of interest.

Model	H2O			FPS
	Left	Right	Object	
Ours w/o EFU & LFE	26.4	25.8	36.1	**58.2**
Ours w/o LFE	22.3	21.8	33.9	56.5
Ours	**20.1**	**19.9**	**32.9**	53.5

Table 3: Ablation studies on each component of QORT Transformer decoder. EFU denotes Enhanced Feature update and LFE denotes location-based feature extraction.

estimation loss, $\mathcal{L}_{\text{CM}}$ can be defined as follows,

$$\mathcal{L}_{\text{CM}} = \|\mathbf{C}_{\text{M}} - \mathbf{C}'_{\text{M}}\|_1. \quad (4)$$

Experiments

Datasets and Evaluation Metrics

We conducted evaluations on two distinct datasets: H2O (Kwon et al. 2021) and FPHA (Garcia-Hernando et al. 2018), both of which include annotations for 3D hand poses, object 6D poses, object types and interaction classes. For hand-object pose estimation, we measure the mean end-point error (in mm) across 21 joints, and for our extended experiment interaction recognition, we use top-1 accuracy with an off-the-shelf network (Cho et al. 2023). Further details on datasets, implementation details and evaluation metrics are available in our supplemental materials.

Experiment Results

Pose Estimation. We compare our method with SOTA hand-object pose estimation methods that use a single RGB image as input on the H2O and FPHA datasets. As experiment results in Table 1 demonstrate, our method achieves SOTA results and outperforms all previous methods by a significant margin. Compared to Cho et al., we achieve substantial gains on the H2O dataset: 17.6% for the left hand, 22.8% for the right hand, and 27.2% for the object. On the FPHA dataset, our method outperforms the state-of-the-art by a decisive margin: 5.3% for the right hand and 10.4% for the object. Hasson et al. and Tekin et al. (Tekin, Bogo, and Pollefeys 2019) predict pose for only a single hand. Wen et al.'s method does not predict object poses. Figs. 4 and 5 show the example estimated 3D poses of hands and an object on H2O and FPHA datasets, respectively.

-		w/o HOQ	w/ HOQ	w/ HOQ + CM
H2O	Left	27.3	22.6	**20.1**
	Right	33.5	21.4	**19.9**
	Object	36.5	33.1	**32.9**
FPHA	Right	22.3	15.8	**14.2**
	Object	24.2	20.3	**18.8**

Table 4: Ablation studies on different components of our framework. In the table, HOQ denotes hand-object queries and CM denotes contact map features.

Interaction Recognition. In our extended experiment on hand-object interaction recognition, we utilized the action recognition module from Cho et al. to assess the impact of our improved inference speed of our pose estimator on interaction classification. By replacing their pose estimator with our proposed method, QORT-Former, we observed an increase in performance from 90.9 to 91.3 on the H2O dataset, as shown in Table 2. Additionally, incorporating our method led to a significant boost in inference speed, achieving 53.3 FPS compared to the 24.97 FPS of Cho et al.'s framework on an RTX 3090TI GPU.

Ablation Study

In this section, we conduct ablation studies on our improved queries and decoder of our proposed model, QORT-Former. Further ablations of other configurations of our architecture are available in our supplementary mat. Each component is evaluated on the test sets of the H2O and/or FPHA datasets.

Analysis of Our Improved Queries. To enhance the efficiency of our QORT Transformer decoder, we introduce hand-object queries combined with a query proposal loss, aimed at improving query location accuracy by ensuring queries focus on regions of interest, such as the left and right hands and the interacting object. This approach addresses the challenge of imprecise query location estimation, which often arises in hand pose estimation due to the semantic similarity between the left and right hands. The improved query distribution leads to a more balanced model performance across both hands, as evidenced by a substantial performance increase demonstrated in Table 4 and illustrated in Figure 3. Additionally, to further refine the model's accuracy in interacting hands and object pose estimation, we incorporate contact map features that capture the spatial relationships between the hands and the object. By adding these features to the query inputs, our model achieves state-of-the-art results in hand and object pose estimation from a single RGB image, as shown in Table 4.

Analysis on QORT Transformer Decoder. We adopted a three-step update approach in our proposed decoder to compensate for the reduced feature maps compared to heavier architectures (Cho et al. 2023; Cheng et al. 2022a). Unlike

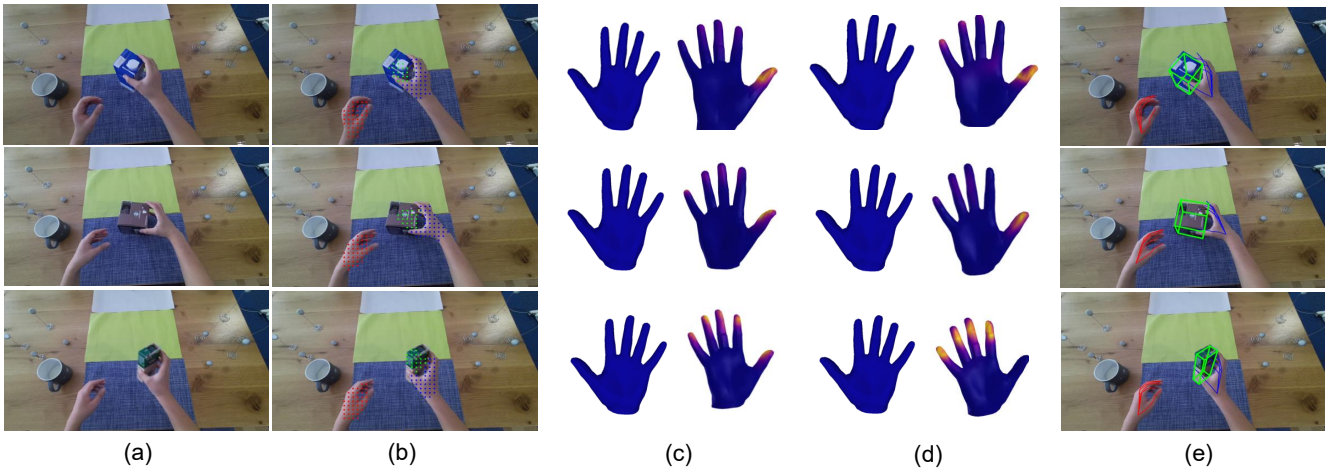

Figure 4: Examples of estimated 3D poses on H2O dataset: For a separate example in each row, the figure represents (a) input RGB image, (b) our hand-object queries, (c) ground-truth contact map, (d) predicted contact map, and (e) final 3D pose estimation results, respectively.

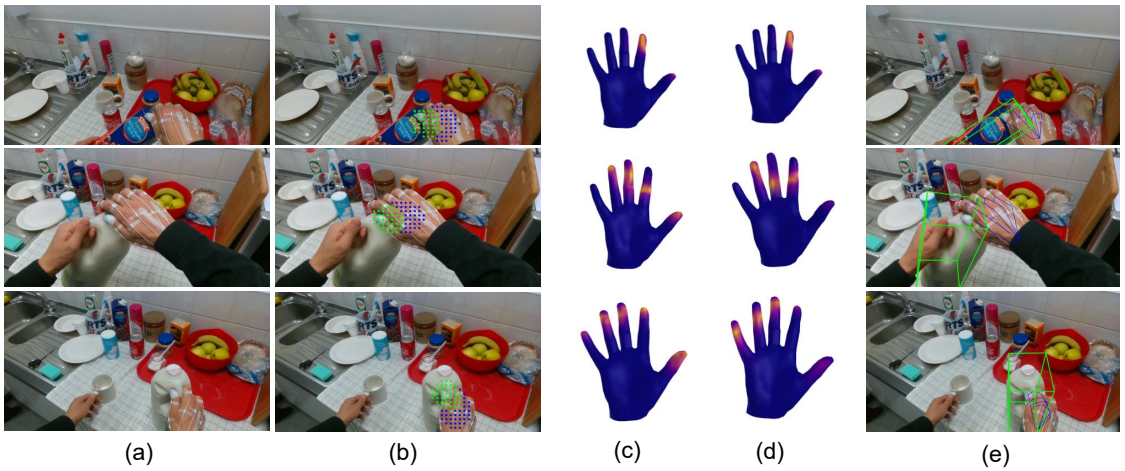

Figure 5: Examples of estimated 3D poses on FPHA dataset. For a separate example in each row, the figure represents (a) input RGB image, (b) our hand-object queries, (c) ground-truth contact map, (d) predicted contact map, and (e) final 3D pose estimation results, respectively.

(Cheng et al. 2022a), which relies solely on query feature updates, we first refine enhanced features $\mathbf{E}_3$ using cross-attention and then apply location-based feature extraction to focus on hands and objects. This refined $\mathbf{E}_3$ then, guides the final query feature update. Although this method slightly reduces FPS, it improves overall pose estimation performance, as shown in Table 3.

Conclusion

In this work, we present QORT-Former, the first real-time Transformer-based framework designed specifically for two hands and interacting object pose estimation. Our approach introduces a lightweight feature decoder with pyramid pooling, significantly reducing the number of queries to just 108 while effectively incorporating contact map features to model intricate hand-object interactions. By leveraging a novel three-step update strategy (Enhanced Feature Update, Location-based Enhancement, Query Feature Update, QORT-Former minimizes the computational overhead of the encoder and simplifies the architecture by utilizing a single decoder. These innovations collectively enable QORT-Former to achieve state-of-the-art performance on widely-used hand-object interaction benchmarks such as the H2O and FPHA datasets. Furthermore, QORT-Former operates at an impressive speed of 53.5 frames per second (FPS) on an RTX 3090TI GPU, demonstrating its practical viability for real-time applications. The combination of accuracy, efficiency, and real-time performance positions QORT-Former as a significant advancement in the field of hands-object pose estimation, paving the way for new possibilities in applications, such as human-computer interaction, robotics, augmented reality and virtual reality.

Acknowledgements

This work was supported by IITP grants (No. RS-2020-II201336 Artificial intelligence graduate school program (UNIST) 10%; No. RS-2021-II212068 AI innovation hub 10%; No. RS-2022-II220264 Comprehensive video understanding and generation with knowledge-based deep logic neural network 20%) and the NRF grant (No. RS-2023-00252630 20%), all funded by the Korean government (MSIT). This work was also supported by Korea Institute of Marine Science & Technology Promotion (KIMST) funded by Ministry of Oceans and Fisheries (RS-2022-KS221674) 20% and received support from AI Center, CJ Corporation (20%).

References

Aboukhadra, A. T.; Malik, J.; Elhayek, A.; Robertini, N.; and Stricker, D. 2023. THOR-Net: End-to-end Graformer-based Realistic Two Hands and Object Reconstruction with Self-supervision. In *WACV*.

Armagan, A.; Garcia-Hernando, G.; Baek, S.; Hampali, S.; Rad, M.; Zhang, Z.; Xie, S.; Chen, N.; Zhang, B.; Xiong, F.; Xiao, Y.; Cao, Z.; Yuan, J.; Ren, P.; Huang, W.; haifeng sun; Hrúz, M.; Kanis, J.; Krňoul, Z.; Wan, Q.; Li, S.; Lee, D.; Yang, L.; Yao, A.; Liu, Y.-H.; Spurr, A.; Molchanov, P.; Iqbal, U.; Weinzaepfel, P.; Brégier, R.; Rogez, G.; Lepetit, V.; and Kim, T.-K. 2020. Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3D hand pose estimation under hand-object interaction. In *ECCV*.

Baek, S.; Kim, K. I.; and Kim, T.-K. 2018. Augmented skeleton space transfer for depth-based hand pose estimation. In *CVPR*.

Baek, S.; Kim, K. I.; and Kim, T.-K. 2019. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *CVPR*.

Baek, S.; Kim, K. I.; and Kim, T.-K. 2020. Weakly-Supervised Domain Adaptation via GAN and Mesh Model for Estimating 3D Hand Poses Interacting Objects. In *CVPR*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Cha, J.; Kim, J.; Yoon, J. S.; and Baek, S. 2024. Text2HOI: Text-guided 3D Motion Generation for Hand-Object Interaction. In *CVPR*.

Chen, D.; Li, J.; Wang, Z.; and Xu, K. 2020. Learning canonical shape space for category-level 6d object pose and size estimation. In *CVPR*.

Chen, L.; Lin, S.-Y.; Xie, Y.; Lin, Y.-Y.; and Xie, X. 2021. Temporal-aware self-supervised learning for 3d hand pose and mesh estimation in videos. In *WACV*.

Chen, X.; Liu, Y.; Dong, Y.; Zhang, X.; Ma, C.; Xiong, Y.; Zhang, Y.; and Guo, X. 2022. MobRecon: Mobile-Friendly Hand Mesh Reconstruction from Monocular Image. In *CVPR*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022a. Masked-Attention Mask Transformer for Universal Image Segmentation. In *CVPR*.

Cheng, T.; Wang, X.; Chen, S.; Zhang, W.; Zhang, Q.; Huang, C.; Zhang, Z.; and Liu, W. 2022b. Sparse instance activation for real-time instance segmentation. In *CVPR*.

Cho, H.; Kim, C.; Kim, J.; Lee, S.; Ismayilzada, E.; and Baek, S. 2023. Transformer-Based Unified Recognition of Two Hands Manipulating Objects. In *CVPR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Fan, Z.; Ohkawa, T.; Yang, L.; Lin, N.; Zhou, Z.; Zhou, S.; Liang, J.; Gao, Z.; Zhang, X.; Zhang, X.; et al. 2025. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *ECCV*.

Fan, Z.; Parelli, M.; Kadoglou, M. E.; Chen, X.; Kocabas, M.; Black, M. J.; and Hilliges, O. 2024. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *CVPR*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *ICCV*.

Fu, Q.; Liu, X.; Xu, R.; Niebles, J. C.; and Kitani, K. M. 2023. Deformer: Dynamic Fusion Transformer for Robust Hand Pose Estimation. In *ICCV*.

Garcia-Hernando, G.; Yuan, S.; Baek, S.; and Kim, T.-K. 2018. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*.

Hampali, S.; Sarkar, S. D.; Rad, M.; and Lepetit, V. 2022. Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation. In *CVPR*.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *PAMI*.

Hasson, Y.; Tekin, B.; Bogo, F.; Laptev, I.; Pollefeys, M.; and Schmid, C. 2020. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *CVPR*.

He, J.; Li, P.; Geng, Y.; and Xie, X. 2023. FastInst: A Simple Query-Based Model for Real-Time Instance Segmentation. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Huang, L.; Tan, J.; Liu, J.; and Yuan, J. 2020. Hand-transformer: Non-autoregressive structured modeling for 3d hand pose estimation. In *ECCV*.

Iwase, S.; Liu, X.; Khirodkar, R.; Yokota, R.; and Kitani, K. M. 2021. RePOSE: Real-time iterative rendering and refinement for 6d object pose estimation. In *ICCV*.

Jiang, C.; Xiao, Y.; Wu, C.; Zhang, M.; Zheng, J.; Cao, Z.; and Zhou, J. T. 2023. A2J-Transformer: Anchor-to-Joint Transformer Network for 3D Interacting Hand Pose Estimation From a Single RGB Image. In *CVPR*.

Karunratanakul, K.; Yang, J.; Zhang, Y.; Black, M. J.; Muandet, K.; and Tang, S. 2020. Grasping field: Learning implicit representations for human grasps. In *3DV*.

Kehl, W.; Manhardt, F.; Tombari, F.; Ilic, S.; and Navab, N. 2017. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *ICCV*.

Kim, D. U.; Kim, K. I.; and Baek, S. 2021. End-to-end detection and pose estimation of two interacting hands. In *ICCV*.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*.

Kwon, T.; Tekin, B.; Stühmer, J.; Bogo, F.; and Pollefeys, M. 2021. H2o: Two hands manipulating objects for first person interaction recognition. In *CVPR*.

Lee, S.; Park, H.; Kim, D. U.; Kim, J.; Boboev, M.; and Baek, S. 2023. Image-free Domain Generalization via CLIP for 3D Hand Pose Estimation. In *WACV*.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*.

Li, Z.; Wang, G.; and Ji, X. 2019. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *CVPR*.

Lin, J.; Mao, X.; Chen, Y.; Xu, L.; He, Y.; and Xue, H. 2022. Dˆ 2ETR: Decoder-Only DETR with Computationally Efficient Cross-Scale Attention. *arXiv preprint arXiv:2203.00860*.

Lin, K.; Wang, L.; and Liu, Z. 2021. Mesh graphormer. In *ICCV*.

Liu, S.; Jiang, H.; Xu, J.; Liu, S.; and Wang, X. 2021. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *ICLR*.

Lv, W.; Xu, S.; Zhao, Y.; Wang, G.; Wei, J.; Cui, C.; Du, Y.; Dang, Q.; and Liu, Y. 2023. DETRs beat YOLOs on real-time object detection. *ArXiv:20304.08069*.

Moon, G.; Yu, S.-I.; Wen, H.; Shiratori, T.; and Lee, K. M. 2020. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*.

Pavlakos, G.; Shan, D.; Radosavovic, I.; Kanazawa, A.; Fouhey, D.; and Malik, J. 2024. Reconstructing hands in 3d with transformers. In *CVPR*.

Tekin, B.; Bogo, F.; and Pollefeys, M. 2019. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*.

Tekin, B.; Sinha, S. N.; and Fua, P. 2018. Real-time seamless single shot 6d object pose prediction. In *CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*.

Wang, J.; Mueller, F.; Bernard, F.; Sorli, S.; Sotnychenko, O.; Qian, N.; Otaduy, M. A.; Casas, D.; and Theobalt, C. 2020. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. In *SIGGRAPH*.

Wang, R.; Mao, W.; and Li, H. 2023. Interacting Hand-Object Pose Estimation via Dense Mutual Attention. In *WACV*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *CVPR*.

Wang, Y.; Zhang, X.; Yang, T.; and Sun, J. 2022. Anchor detr: Query design for transformer-based detector. In *AAAI*.

Wen, Y.; Pan, H.; Yang, L.; Pan, J.; Komura, T.; and Wang, W. 2023. Hierarchical Temporal Transformer for 3D Hand Pose Estimation and Action Recognition From Egocentric RGB Videos. In *CVPR*.

Xiang, Y.; Schmidt, T.; Narayanan, V.; and Fox, D. 2017. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv:1711.00199*.

Xiong, F.; Zhang, B.; Xiao, Y.; Cao, Z.; Yu, T.; Zhou, J. T.; and Yuan, J. 2019. A2J: Anchor-to-Joint Regression Network for 3D Articulated Pose Estimation From a Single Depth Image. In *ICCV*.

Xu, Y.; Lin, K.-Y.; Zhang, G.; Wang, X.; and Li, H. 2022. RNNPose: Recurrent 6-DoF Object Pose Refinement with Robust Correspondence Field Estimation and Pose Optimization. In *CVPR*.

Yang, L.; Zhan, X.; Li, K.; Xu, W.; Li, J.; and Lu, C. 2021. CPF: Learning a contact potential field to model the hand-object interaction. In *CVPR*.

Yao, Z.; Ai, J.; Li, B.; and Zhang, C. 2022. Efficient detr: improving end-to-end object detector with dense prior. In *ICLR*.

Zhang, P.; and Kong, D. 2024. Handformer2T: A Lightweight Regression-Based Model for Interacting Hands Pose Estimation From a Single RGB Image. In *WACV*.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*.

Zhao, W.; Wang, W.; and Tian, Y. 2022. Graformer: Graph-oriented transformer for 3d pose estimation. In *CVPR*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*.

Zimmermann, C.; and Brox, T. 2017. Learning to estimate 3D hand pose from single RGB images. In *ICCV*.