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, South Korea [2]Michigan State University, USA

AAAI-25 / IAAI-25 / EAAI-25

Abstract

- Presents QORT-Former, a real-time Transformer-based framework for estimating 3D poses of two hands and an object.
- Reduces reliance on heavy encoders by generating semantically meaningful queries and refining both image and query features within a single decoder stage.
- Integrates contact map features to enhance hand-object interaction cues.
- Achieves state-of-the-art performance on H2O and FPHA datasets at 53.5 FPS on an RTX 3090TI.

Contributions

- **Real-Time Transformer:** Introduces QORT-Former, a fast, Transformer-based architecture for two-hand-and-object 3D pose estimation.
- **Efficient Query Strategy:** Constrains query number (108) and decoder count (1) to speed up inference.
- **Contact Map Integration:** Incorporates contact map features into object queries for robust interaction modeling.
- **Three-Step Feature Update in Decoder:** Reduces heavy decoder dependency, limits decoder count, and ensures lightweight efficiency.
- **Superior Accuracy:** Outperforms existing methods on H2O and FPHA with a 5.3–27.2% margin, maintaining real-time speeds

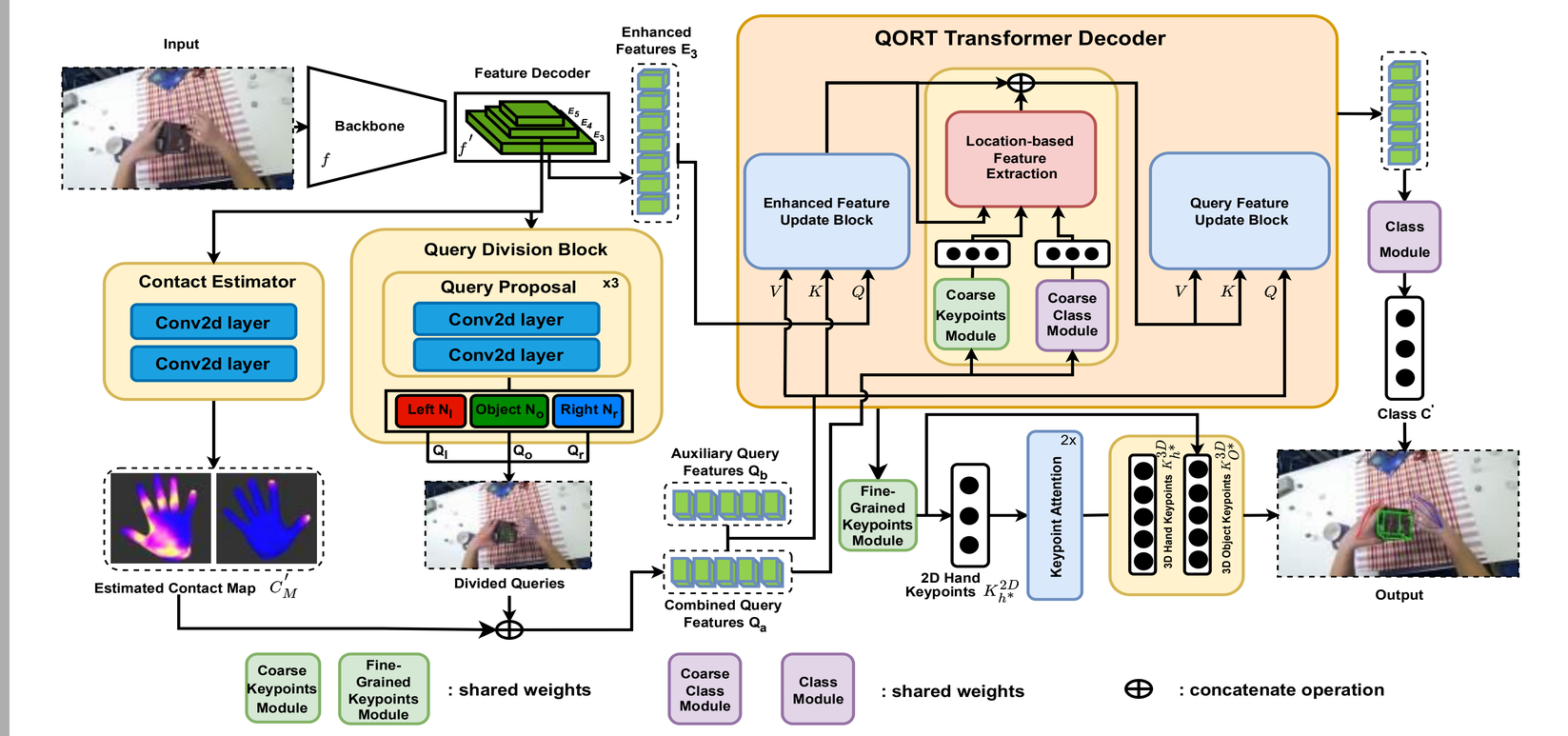
Method

- **Backbone & Feature Decoder**: Utilizes ResNet-50 + PPM-FPN (Zhao et al. 2017) to produce multi-scale features.
- **Query Division Block**: Divides queries into three classes (left hand, right hand, object), selecting top semantic locations from mid-level features.
- **Contact Estimator**: Learns contact maps for hands and incorporates them into object queries for refined interactions.
- **QORT Transformer Decoder**
  - **Enhanced Feature Update:** Cross-attention refines image features with integrated queries.
  - **Location-Based Enhancement**: Zooms into 3×3 patches around coarse keypoint predictions.
  - **Query Feature Update**: Final cross- and self-attention layers capture fine details.

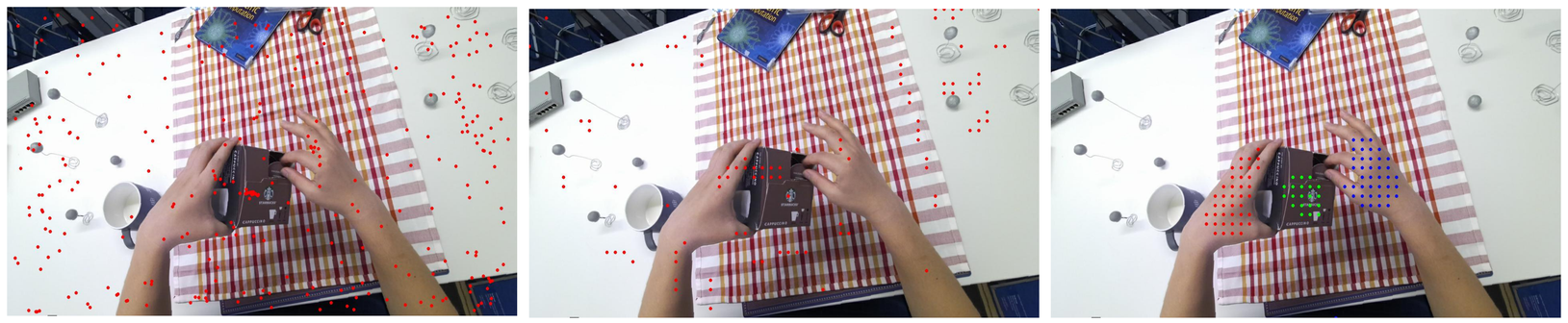
References

- Cho, H., Kim, C., Kim, J., Lee, S., Ismayilzada, E., & Baek, S. (2023). Transformer-based unified recognition of two hands manipulating objects. In CVPR 2023
- Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *CVPR 2017*

Overall Architecture

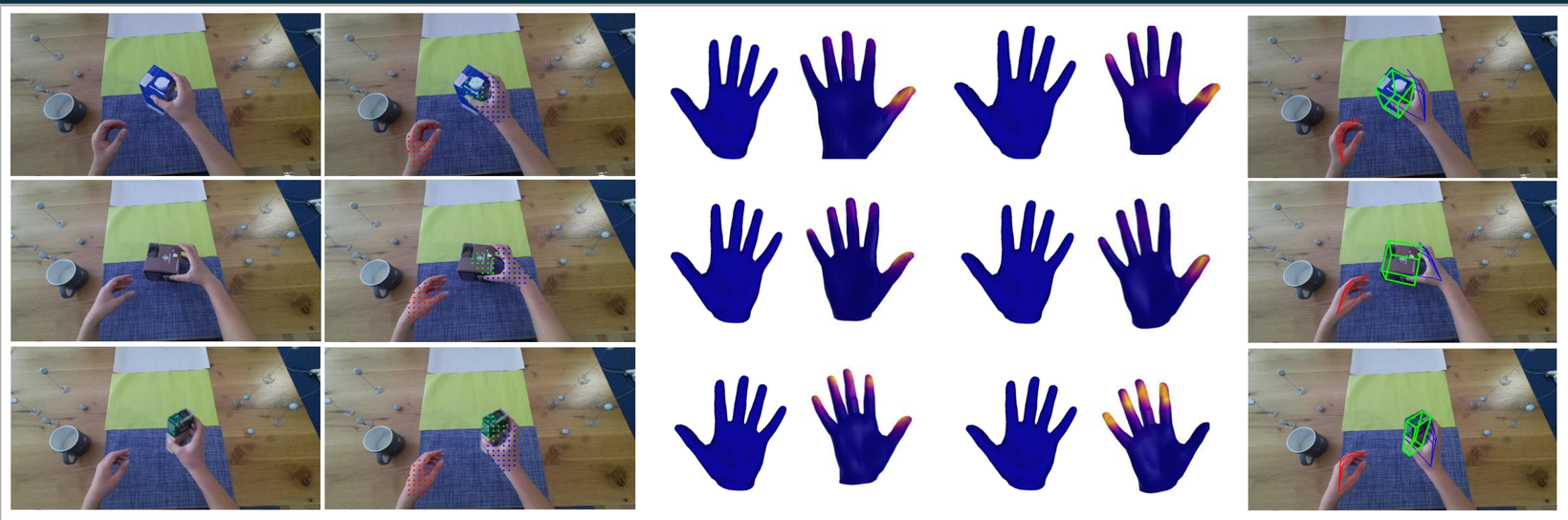


Query Location Visualization



- **Left:** query locations of H2OTR (Cho et al. 2023), employing 300 queries**. Middle:** Our hand-object query locations w/o Query division block. **Right:** Our hand-object query locations.
- Demonstrates how our method ensures queries precisely focus on left/right hands and the object, avoiding random dispersion in backgrounds.
- Improves coverage of key interaction regions, boosting hand-object pose accuracy.

Results



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.