Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation

Supplementary Material

Shreyas Hampali$^{(1)}$, Sayan Deb Sarkar$^{(1)}$, Mahdi Rad$^{(1)}$, Vincent Lepetit$^{(2,1)}$

$^{(1)}$Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria
$^{(2)}$LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France
{<firstname>.<lastname>}@icg.tugraz.at, vincent.lepetit@enpc.fr

In this supplementary material, we discuss the limitations of our method, provide more details about the experiments and also show several qualitative results and comparisons. We also refer the reader to the **Supplementary Video** for visualization of results on different action sequences.

1. Hand Pose Representations and Losses

We detail the three possible representations mentioned in Section 3.4 of the paper. We assume 21 3D-joint locations per hand as in the MANO [10] model. The losses for each of the 3 representations are summarized in Table 1.

3D representation. In this representation, each joint j is associated with a parent-relative joint vector $V(j) = J_{\text{3D}}(j) - J_{\text{3D}}(p(j))$, where J_{3D} is the 3D joint location and $p(j)$ refers to the parent joint index of joint j. We estimate 20 joint vectors per hand using 20 joint queries, one for each skeletal bone (40 queries for two hands), from which we can compute the root-relative 3D location, J_{3D}^r of each joint by simple accumulation. The advantage of this representation is that it defines the hand pose relative to its root without requiring knowledge of the camera intrinsics.

2.5D representation [5, 9]. In this representation, each joint is parameterised by its 2D location J_{2D}, and the difference ΔZ^p between its depth and the depth of its parent joint. The camera intrinsics matrix K and the absolute depth Z_{root} of the root joint (the wrist) [9] or the scale of the hand [5] are then required to reconstruct the 3D pose of the hand in camera coordinate system as $J_{\text{3D}} = K^{-1} \cdot (Z_{\text{root}} + \Delta Z^r) \cdot \left[J_{\text{2D}_x}, J_{\text{2D}_y}, 1\right]^T$, where ΔZ^r is the root-relative depth of the joint computed from its predicted ΔZ^p and the predicted ΔZ^p for its parents. $J_{\text{2D}_x}, J_{\text{2D}_y}$ are the predicted x and y coordinates of J_{2D}.

When using this representation, we also predict the root depth Z_{root} separately using RootNet [8] as in [9]. Each joint query estimates the J_{2D} and ΔZ^r for that joint and we require a total of 21 joint queries (42 for two hands), one for each joint location to estimate the 2.5D pose per hand.

Representation	$\mathcal{L}_{hand-pose}$
3D	$\sum_j \lvert\lvert V(j) - V(j)^* \rvert\rvert_1 + \sum_j \lvert\lvert J_{3D}^r(j) - J_{3D}^{r^*}(j) \rvert\rvert_1$
2.5D	$\sum_j \lvert\lvert J_{2D}(j) - J_{2D}^*(j) \rvert\rvert_1 + \sum_j \lvert \Delta Z^r(j) - \Delta Z^{r^*}(j) \rvert$
θ	$\sum_j \lvert\lvert J_{3D}^r(j) - J_{3D}^{r^*}(j) \rvert\rvert_1 + \sum_j \lvert\lvert \theta(j) - \theta^*(j) \rvert\rvert_1$

Table 1. Hand pose losses for different pose representations. x^* denotes the ground-truth values for variable x and $x(j)$ the value of x at joint j.

MANO joint angles, θ [10]. In this representation, each 3D hand pose is represented by 16 3D joint angles in the hand kinematic tree and is estimated using 16 joint queries per hand, one for each joint. The MANO hand shape parameter is estimated along with the relative translation between the hands using an additional query. Given the predicted 3D joint angles θ for each hand and the shape parameters β, it is possible to compute the root-relative 3D joint locations, J_{3D}^r of each hand.

2. Method Limitations

Though our method results in accurate poses during interactions, the results are sometimes not plausible as we do not model contacts and interpenetration [1, 3, 6] between hands and objects. Further, during highly complex and severely occluded hand interactions as we show in the last row of Fig. 9, our method fails to obtain reasonable hand poses. We believe these problems can be tackled in the future by incorporating temporal information and physical modeling into our architecture.

3. Hand-Object Pose Estimation Pipeline

In Fig. 1, we show the complete pipeline of our Keypoint-Transformer architecture for estimating poses of two hands and object during interaction.

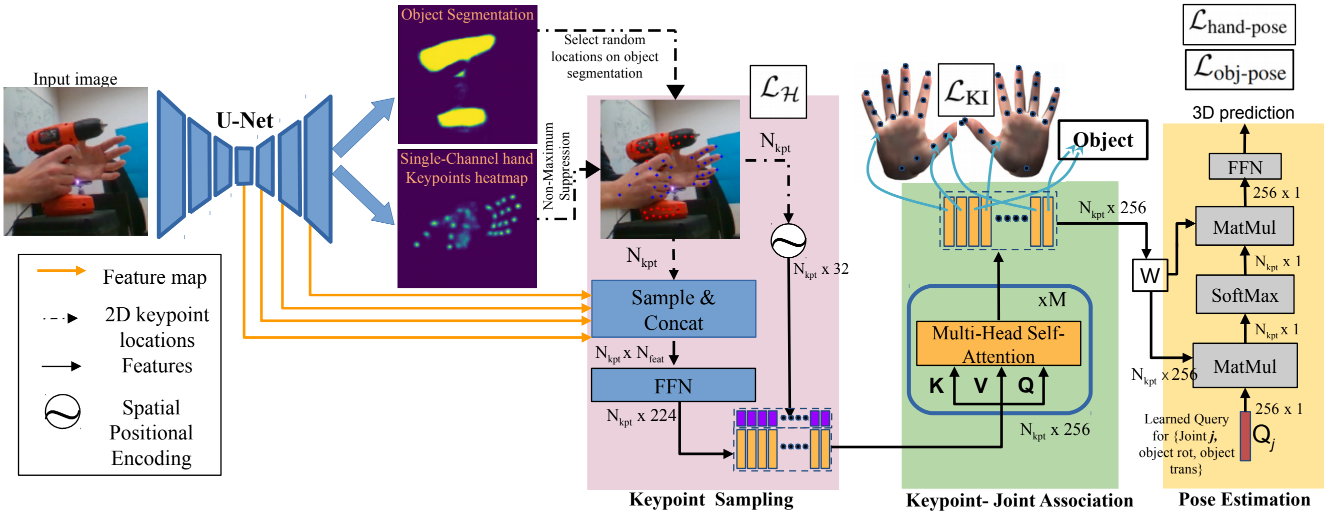

Figure 1. **Pipeline for hands and object pose estimation.** The object keypoints are selected by randomly sampling 2D locations on the object segmentation map regressed by the U-Net (Section 3.5 of main paper). The hand keypoints are selected from the single-channel keypoints heatmap, also regressed by the U-Net (Section 3.1 of main paper). Each of the detected keypoints are encoded using CNN image features and spatial embedding. The keypoints are associated with one of the 42 hand joints (21 joints per hand), the object class or the background class in the keypoint-joint association stage (Section 3.2 of main paper). The object rotation and translation w.r.t the right hand is estimated in the pose estimation stage using 2 different learned object queries, while the pose of each hand-joint is estimated using per-joint learned queries (Section 3.3 of main paper).

4. Implementation details

The encoder of our U-Net [11] is based on ResNet-50 [4] architecture while a series of upsampling and convolutional layers with skip connections forms the U-Net decoder. We use 256×256 pixels as input image resolution, 128×128 pixels as heatmap resolution, and set the 2D Gaussian kernel variance, σ to 1.25 during training. The 256×256 pixel input image patch is loosely cropped around the hand and object. We use Adam [7] optimizer with a learning rate of 10^{-4} and 10^{-5} for the attention modules and CNN backbone, respectively. The network is trained for 50 epochs on 3 Titan V GPUs with a total batch size of 78 and uses on-line augmentation techniques such as rotation, scale and mirroring during training.

5. Baseline Architectures

We detail here the two baselines, 'CNN+SA' and 'CNN+SA+CA' considered in Section 4.1 of the main paper. Figures 2 and 3 show their architectures. We used 256 × 256 cropped images as input to the CNN resulting in a feature map of spatial dimensions 8 × 8 and 2048 channels. The features are flattened along the spatial dimensions and the 64 features are converted to 224 dimensions using 3 MLP layers. These features are then concatenated with 32-D positional embeddings resulting in 256-D features and are provided to the Transformer encoder. The networks were trained to output the 2.5D pose representation for 50 epochs on 3 Titan V GPUs with a batch size of 78. The joint queries in 'CNN+SA+CA' are learned in a similar way as for our Keypoint Transformer.

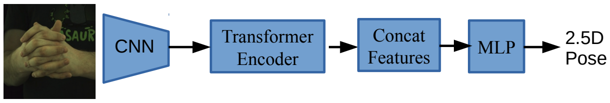

Figure 2. The 'CNN+SA' baseline architecture.

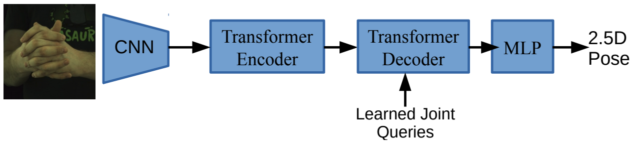

Figure 3. The 'CNN+SA+CA' baseline architecture.

6. Robustness to Noisy Keypoints

We show more examples to demonstrate the robustness of our method to noisy keypoints. We consider two scenarios, adding noisy keypoints to the set of detected keypoints, and randomly removing some keypoints from the set of detected keypoints. We show results in Figures 4 and 5, respectively. The number of detected keypoints for these cases were 48 and we added 30 additional noisy keypoints for the former scenario and retained only 30 keypoints for the latter scenario.

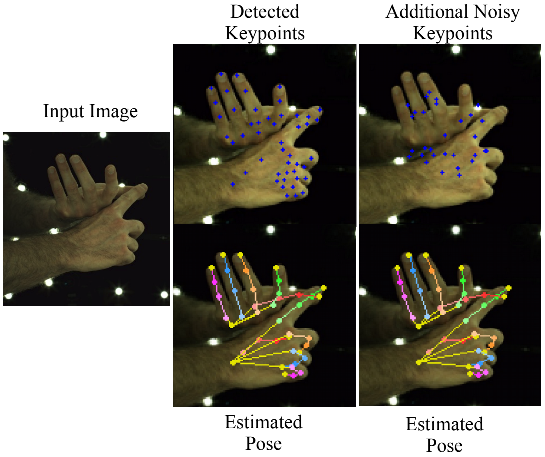

Figure 4. Effect of adding additional noisy keypoints. Our method predicts accurate poses even with noisy keypoints.

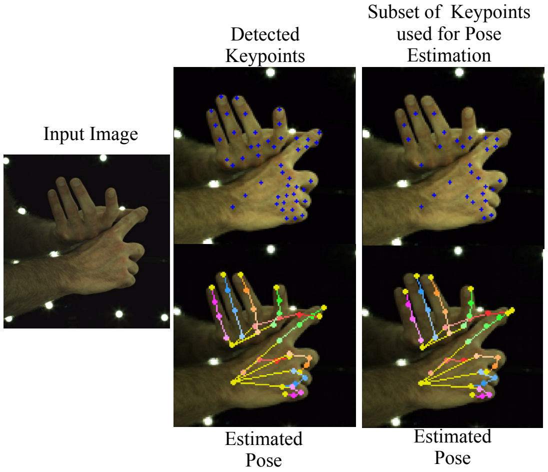

Figure 5. Effect of using a subset of detected keypoints for pose estimation. We consider only 30 of the 48 detected keypoints for pose estimation and still estimate an accurate pose.

7. H_2O-3D Dataset

Our dataset contains sequences of two hands interacting with an object, captured on a multi-view setup with 5 RGBD cameras. We collected data from six different subjects and considered ten objects from the YCB dataset with each subject manipulating the object with a functional intent. The dataset is automatically annotated with 3D poses of hands and objects using the optimization method of [2]. The dataset contains 60'998 training images and 15'342 test images from 17 different multi-view sequences in total. As explained in the main paper, we only consider 9'098 images from the set of 15'342 test images for object pose evaluation as the objects in the remaining images are barely visible due to occlusion by the hands. We show some sample annotations from the dataset in Fig. 6. Table 2 shows the list of YCB objects and their axis and angle of symmetry considered during our training and evaluation.

7.1. Per-Object MSSD Values with Keypoint Transformer

Table 3 shows the accuracy of the object poses estimated by our Keypoint Transformer on the H_2O-3D dataset using the MSSD metric as described in Section 4.3 of the main paper.

8. Qualitative Results and Comparisons

We provide here more qualitative results on HO-3D, H_2O-3D and InterHand2.6M.

8.1. HO-3D and H_2O-3D Qualitative Results

Fig. 7 shows qualitative results on H_2O-3D and HO-3D. Note that as we do not model contacts and interpenetration between hands and object, our method sometimes results in implausible poses as we show in the last example of Fig. 7.

8.2. InterHand2.6M Qualitative Results

Fig. 8 compares the estimated poses using the InterNet method from [9] and our proposed approach. As noted in Section 1 and Table 3 of the main paper, purely CNN-based

Object	Axis	Angle
Mustard Bottle	Z	180^o
Bleach Cleanser	Z	180^o
Cracker Box	Z	180^o
Sugar Box	Z	180^o
Potted Meat Can	Z	180^o
Bowl	Z	∞
Mug	Z	∞
Pitcher Base	Z	∞
Banana	-	-
Power Drill	-	-

Table 2. H_2O-3D objects and their axis and angle of symmetry considered during training and evaluation with our Keypoint Transformer.

Object	MSSD (cm)
Bleach Cleanser	7.7
Mug	6.5
Banana	9.8
Pitcher Base	7.9
Bowl	7.8
Scissors	13.5
Power Drill	8.5
All	7.9

Table 3. Object pose estimation accuracy of our Keypoint Transformer on the H_2O-3D dataset.

approaches do not explicitly model the relationship between image features of joints and tend to *confuse* joints during complex interactions. Our method performs well during complex interactions and strong occlusions (see last row of Fig. 8).

We show more qualitative results using the MANO angle representation in Fig. 9. Our retrieved poses are very similar to ground-truth poses. As we show in the last row of Fig. 9, our method fails during scenarios where the hand is severely occluded during complex interaction.

9. Attention Visualization

In Fig. 10, we show more visualization of the cross-attention weights for three different joint queries. More specifically, the cross-attention weights represent the multiplicative factor on each of the keypoint features for a given joint query. We observe that the cross-attention learns to select keypoint(s) from respective joint location for each joint query when the joint is visible. For occluded joints, features from nearby visible joints are selected.

References

[1] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A Dataset of Grasps with Object Contact and Hand Pose. In *European Conference on Computer Vision*, 2020. 1

[2] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A Method for 3D Annotation of Hand and Object Poses. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3, 5

[3] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning Joint Reconstruction of Hands and Manipulated Objects. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1

[4] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016. 2

[5] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand Pose Estimation via Latent 2.5D Heatmap Regression. In *European Conference on Computer Vision*, 2018. 1

[6] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning Implicit Representations for Human Grasps. In *International Conference on 3D Vision*, 2020. 1

[7] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations*, 2015. 2

[8] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera Distance-Aware Top-Down Approach for 3D Multi-Person Pose Estimation from a Single RGB Image. In *International Conference on Computer Vision*, 2019. 1

[9] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *European Conference on Computer Vision*, 2020. 1, 3, 6, 7

[10] Javier Romero, Dimitrios Tzionas, and Michael J. Black. EMbodied Hands: Modeling and Capturing Hands and Bodies Together. *IEEE Transactions on Robotics and Automation*, 36, 2017. 1

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Conference on Medical Image Computing and Computer Assisted Intervention*, 2015. 2

[12] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. *Science*, 2018. 5

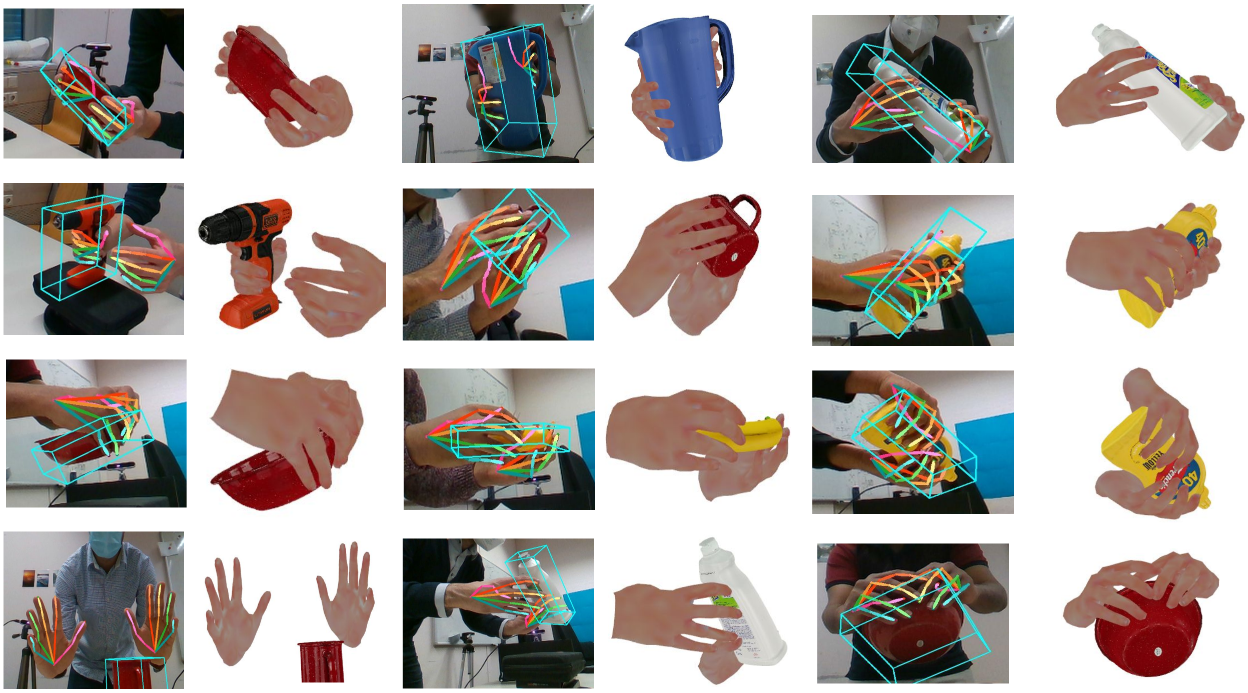

Figure 6. Samples from H_2O-3D dataset. Our dataset contains sequences with complex actions performed by both hands on YCB [12] objects.

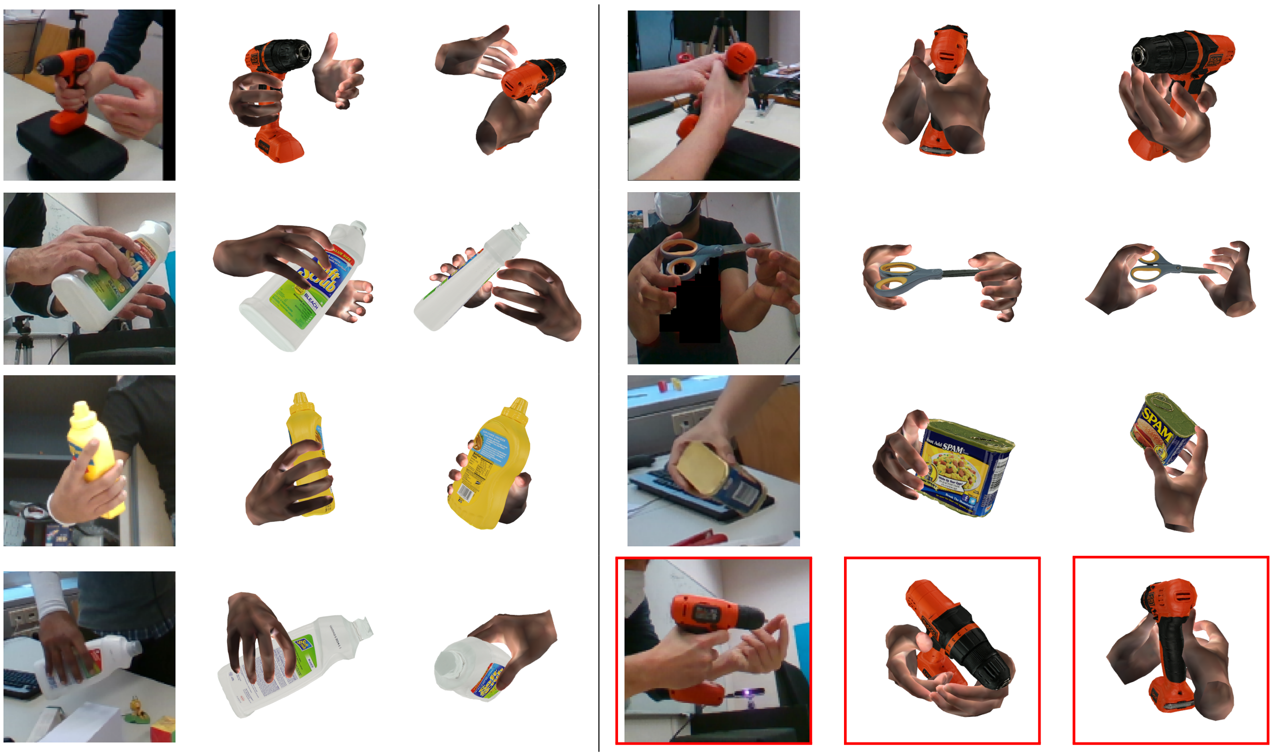

Figure 7. Qualitative results on H_2O-3D and HO-3D [2]. Our method obtains state-of-the-art results on HO-3D while predicting reasonable results on H_2O-3D. The last example is a failure case where the predicted relative translations are inaccurate.

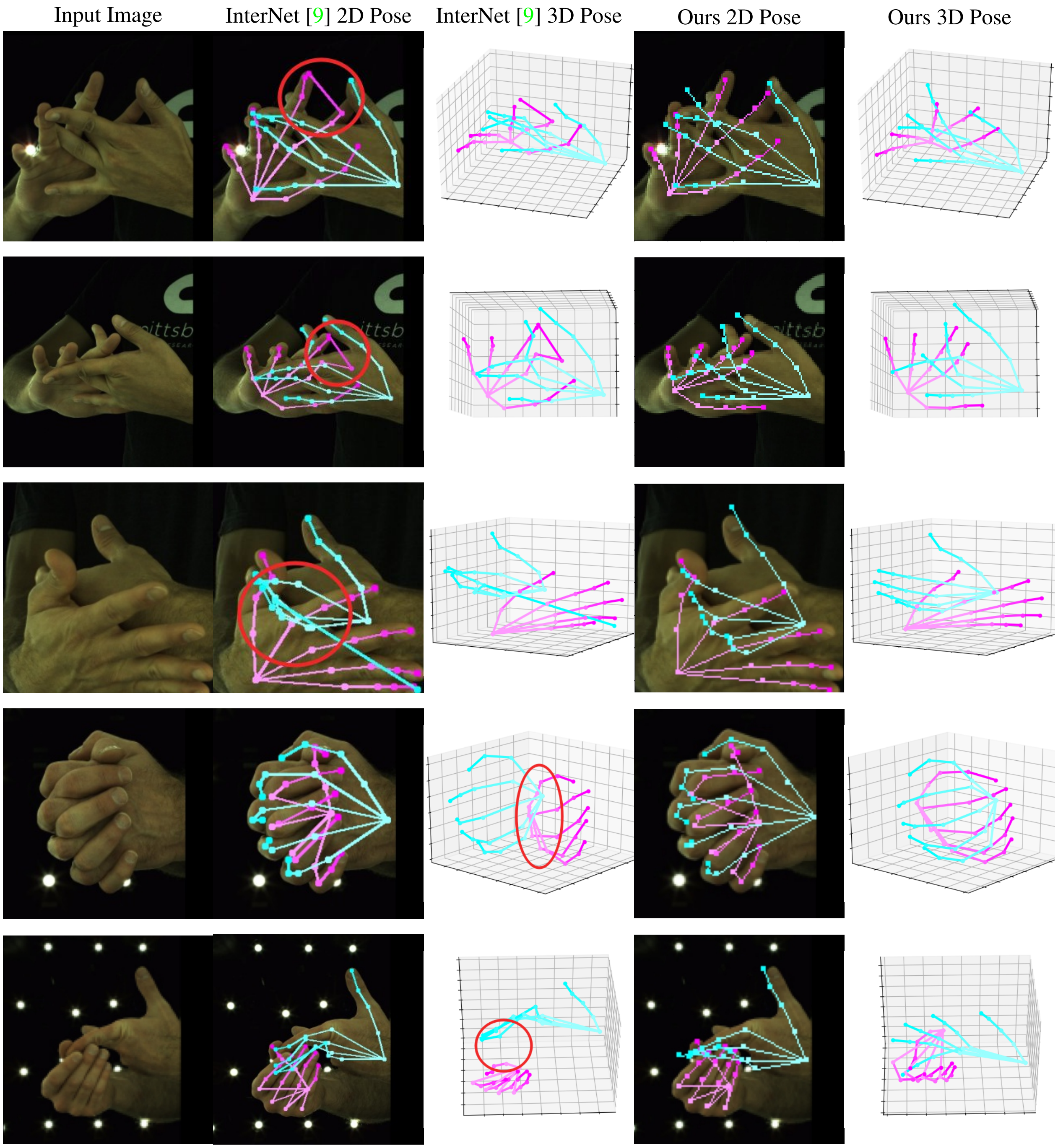

Figure 8. Qualitative comparison between InterNet [9] and our proposed method. Our method outputs more accurate poses even during strong occlusions. Red circles indicate regions where InterNet results are inaccurate.

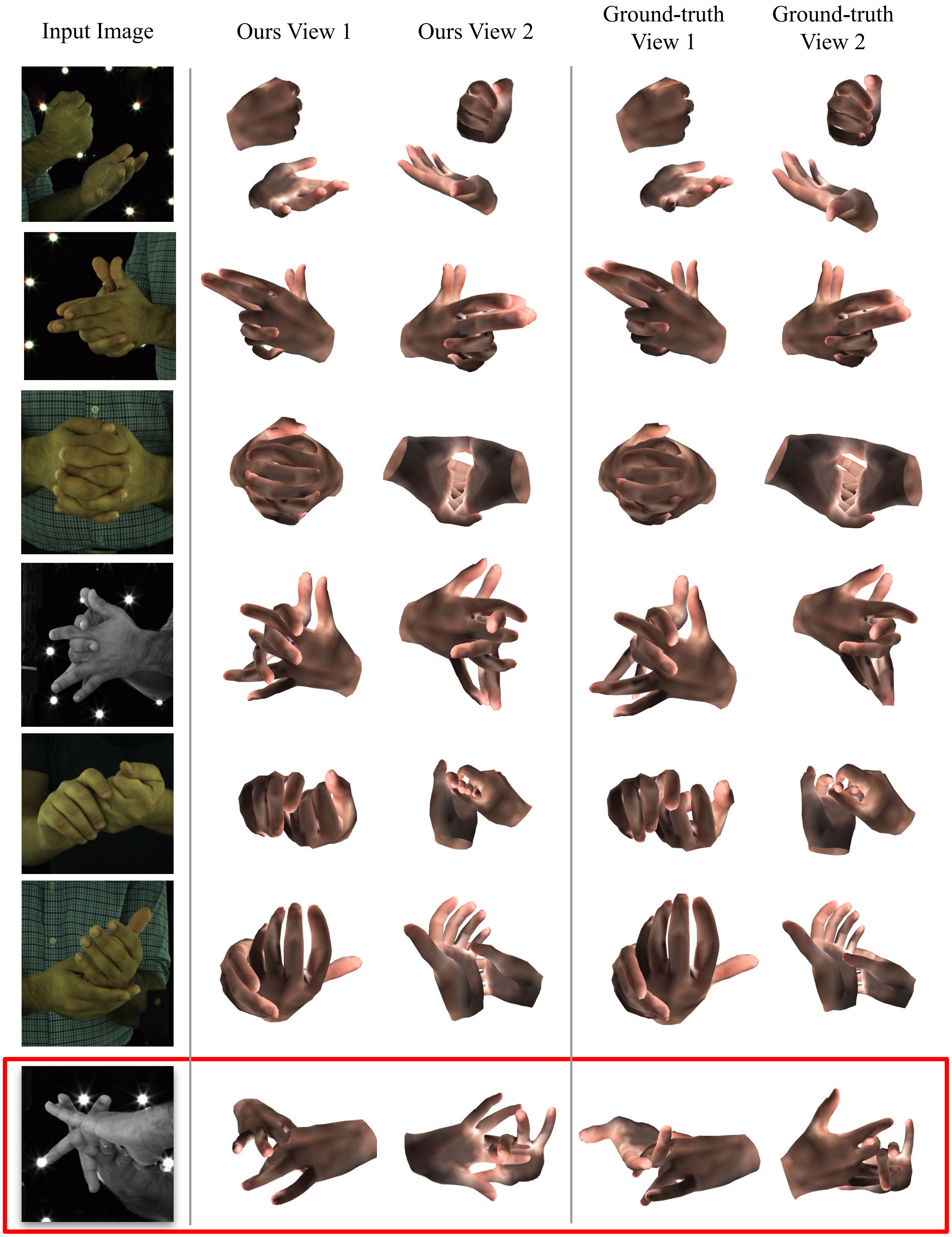

Figure 9. Qualitative results of our method on InterHand2.6M [9] compared to ground-truth poses. Our method predicts accurate poses in most scenarios. The last row shows a failure case where our method cannot recover the accurate pose due to complex pose and severe occlusion.

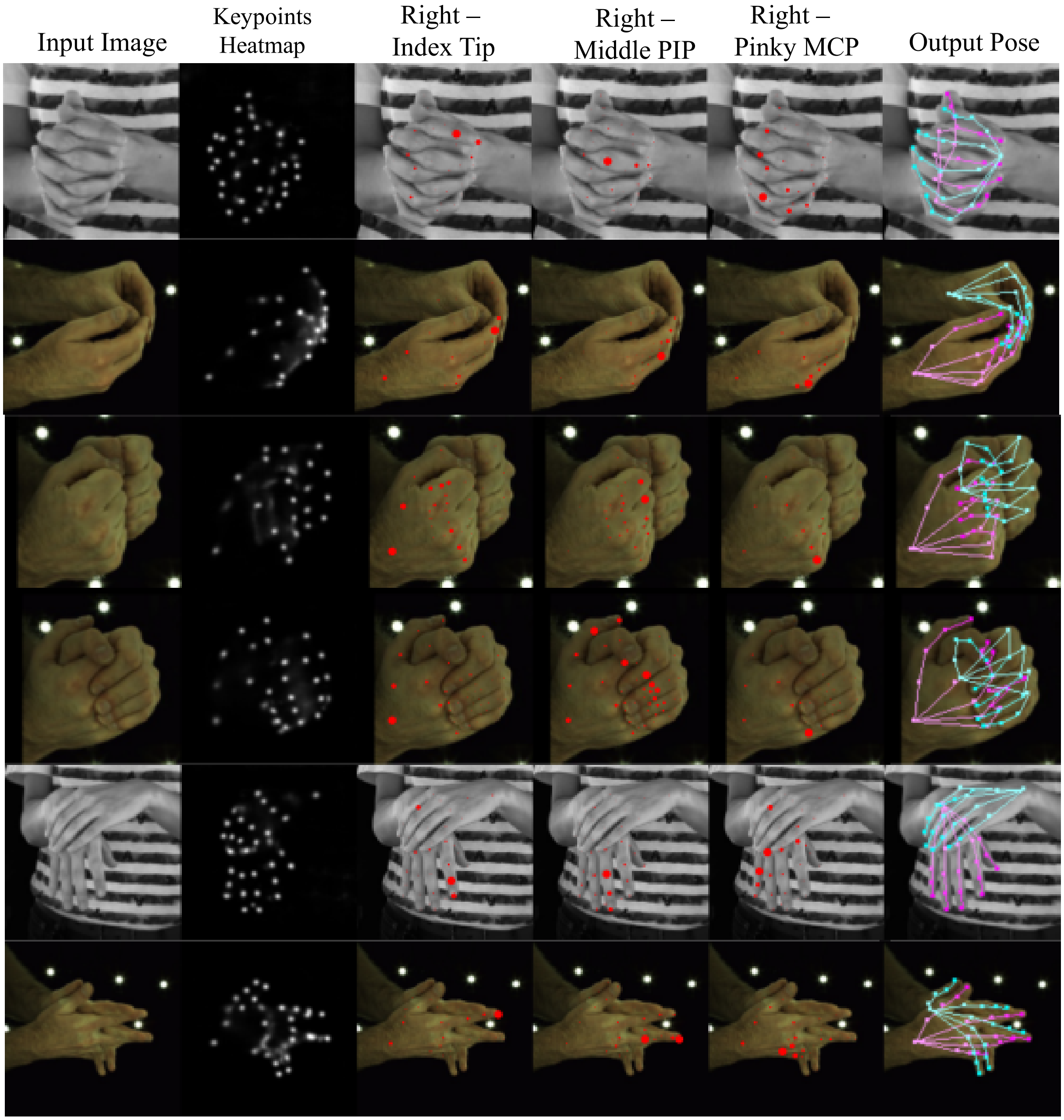

Figure 10. Attention visualization for 3 joint queries. Each joint query attends to the image feature from the respective joint location.