2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 978-1-6654-6946-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CVPR52688.2022.01081

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

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

Abstract

We propose a robust and accurate method for estimating the 3D poses of two hands in close interaction from a single color image. This is a very challenging problem, as large occlusions and many confusions between the joints may happen. State-of-the-art methods solve this problem by regressing a heatmap for each joint, which requires solving two problems simultaneously: localizing the joints and *recognizing them. In this work, we propose to separate these tasks by relying on a CNN to first localize joints as 2D keypoints, and on self-attention between the CNN features at these keypoints to associate them with the corresponding hand joint. The resulting architecture, which we call "Keypoint Transformer", is highly efficient as it achieves state-of-the-art performance with roughly half the number of model parameters on the InterHand2.6M dataset. We also show it can be easily extended to estimate the 3D pose of an object manipulated by one or two hands with high performance. Moreover, we created a new dataset of more than 75,000 images of two hands manipulating an object fully annotated in 3D and will make it publicly available.*

Figure 1. Our approach accurately predicts 3D hand and object poses from a single RGB image in challenging scenarios including complex hand interactions (top) and 2 hands interacting with an object where the hands can be severely occluded (bottom). The bottom example is from the H_2O-3D dataset we also introduce in this paper, which contains challenging, fully and accurately 3D-annotated, video sequences of two hands manipulating objects.

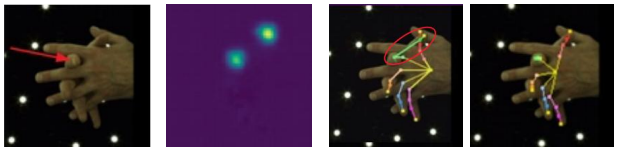
1. Introduction

3D hand pose estimation has the potential to make virtual reality, augmented reality, and interaction with computers and robots much more intuitive. Recently, significant progress has been made for single-hand pose estimation using depth maps and even single RGB images. Being able to deal with RGB images is particularly attractive as it does not require a power-hungry active sensor. Many approaches have been proposed, mostly based on direct prediction with different convolutional network architectures [15, 18, 29, 36, 44, 49, 61] of the 3D joint locations or angles, or relying on rendering for fine pose estimation and tracking [2, 12, 32, 40, 50].

In contrast to single-hand pose estimation, two-hand pose estimation has received much less attention. This problem is indeed significantly harder: The appearance similarities between the joints of the two hands make their identification extremely challenging. Moreover, in close interaction, some of the joints of a hand are likely to be occluded by the other hand or itself. Thus, first detecting the left and right hands and then independently predicting their 3D poses [13, 36] performs poorly in close interaction scenarios. Bottom-up approaches [30, 54] directly estimate the 2D joint locations and their depths using one heatmap per joint. However, as shown in Fig. 2, the similarity in appearances of the joints and severe occlusions degrade the quality of heatmaps failing to localize the joints accurately. More re-

978-1-6654-6946-3/22/$31.00 ©2022 IEEE
DOI 10.1109/CVPR52688.2022.01081

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:29:07 UTC from IEEE Xplore. Restrictions apply.



Input image | Heatmap from [30] for the Right Index Tip | Right hand pose recovered by [30] | Right hand pose recovered by our method

Figure 2. **Similar appearances between joints and partial occlusions make previous methods prone to failure.** The InterNet state-of-the-art method [30] predicts a heatmap for each joint but the predicted heatmaps can become ambiguous, resulting in failures when predicting the hand pose (the hand on the back in this example). Our approach explicitly models the relationship between keypoints resulting in more accurate poses. More examples can be found in the supplementary material.

cent works [10,22,58] have attempted to alleviate this problem by exploiting joint-segmentation, joint-visibility, or by adding more refinement layers increasing the overall complexity of the network. By exploiting only keypoints, our method outperforms these methods by a large margin with a significantly smaller model.

As shown in Fig. 3, instead of aiming to localize and recognize the hand joints simultaneously, we estimate the 3D poses of the hands in three stages: (1) We first detect "keypoints", which are potential joint locations in the image, by predicting a *single* heatmap. These keypoints do not have to exactly match all the hand joints: The 3D poses we predict are still correct if some joints are not detected as keypoints, and if some keypoints do not correspond to joints. (2) Then, we associate the keypoints with the corresponding joint or to the background in the case of false positives, on the basis of the keypoint locations and their image features. This is done for all the keypoints simultaneously to exploit mutual constraints, using the self-attention mechanism. (3) Finally, we predict the 3D hand poses using a cross-attention module, which selects keypoints associated with each of the hand joints. Our approach is agnostic to the parameterization of the pose and we consider three different hand pose representations.

Our architecture, which we call "Keypoint Transformer", is therefore designed to explicitly disambiguate the identity of the keypoints and performs very well even on complex configurations. Fig. 1 shows its output on two challenging examples, using the MANO [41] mesh as the output representation. Our architecture is related to the "Detection Transformer" (DETR) [8] architecture. DETR uses all the spatial features from a low-resolution CNN feature map, combined with learned location queries to detect objects in an image. The high computational complexity of the Transformer restricts DETR from using higher resolution CNN feature maps. As we show in our experiments, using the DETR-style architecture for hand pose estimation results in lower accuracy and we hypothesize that this is due to the use of lower resolution feature maps and features from the entire image.

We train and evaluate our architecture on the recent InterHand2.6M hand-hand [30] and HO-3D hand-object [12] interaction datasets. We also introduce a challenging dataset of videos with two hands interacting with an object with complete and accurate 3D annotations without markers. This dataset is based on the work of [12], and we call it H_2O-3D. Our method achieves state-of-the-art performance on existing hand-interaction datasets and serves as a strong baseline for the H_2O-3D dataset. Our experiments show that on InterHand2.6M, our method achieves state-of-the-art performance with roughly half the number of model parameters. We carry out several ablation studies and compare with strong baselines to prove the efficacy of our approach.

2. Related Work

Many approaches have already been proposed for hand or object pose estimation from either RGB images or depth maps. Here we focus mainly on works that estimate hand poses during hand-hand or hand-object interactions. We also discuss recent advances in Transformer architectures for computer vision as they are highly relevant to our work.

2.1. Interacting Hand Pose Estimation

Hand pose estimation methods can be broadly classified as generative, discriminative, or hybrid approaches. Generative methods [12, 26, 33–35, 52] fit a parametric hand model to an observed image or depth map by minimizing a fitting error under some constraints. Discriminative methods [5, 14, 15, 20, 30, 36, 49, 62] mostly directly predict the hand pose from a single frame. Generative methods often rely heavily on tracking and are prone to drift whereas discriminative methods tend to generalize poorly to unseen images [1]. Hybrid approaches [4, 7, 13, 31, 43, 45, 47, 48, 51, 54, 55] try to combine the best of these two worlds by using discriminative methods to detect visual cues in the image followed by model fitting.

Earlier methods [26,33,34] for generative hand pose estimation during interaction used complex optimization methods to fit a parametric hand model to RGBD data. [12] proposed multi-frame optimization to fit hand and object models to RGBD data from multiple RGBD cameras. Generative methods alone often lose tracking during close interactions or occlusions and are hence combined with discriminative methods to guide the optimization.

[4, 51] detect the fingertips as discriminative points and used them in the optimization along with a collision term and physical modelling. Recently, [43] proposed high-fidelity hand surface tracking of hand-hand interactions in a multi-view setup where the regressed 3D hand joint locations were used for initializing the tracking. [7,13,31,36,54] compute dense features or keypoints from a single RGB or

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:29:07 UTC from IEEE Xplore. Restrictions apply.

depth image and fit a hand model [41] to these estimates with physical constraints and joint angle constraints. Fully discriminative methods [14, 15, 30, 49] jointly estimate the 3D joint locations or hand model parameters of both the interacting hands or the interacting hand and the object by incorporating contacts and inter-penetrations in the training. [20] estimates the hand-object surface using implicit representation that naturally allows modelling of the contact regions between hand and object. [10, 22] improve the accuracy of 3D pose estimation during hand-hand interaction scenarios by incorporating joint-visibility and part-segmentation cues, whereas [58] utilize refinement layers to iteratively refine the estimated poses.

By contrast with the above mentioned approaches designed specifically for hand-hand or hand-object interaction scenarios, we propose in this work a unified discriminative approach for all hand interaction scenarios. Further, many previous discriminative methods perform poorly during close hand interactions due to similarity in appearance of the joints. In this work, we model relationship between all detected joints in the image resulting in more accurate pose estimation while keeping the model complexity low.

The success of discriminative methods depend on the variability of training data and several hand interaction datasets have been proposed. [11] first provided a marker-based hand-object interaction dataset using RGBD cameras. [12, 62] and [15] respectively proposed real and synthetic hand-object interaction dataset with a single hand manipulating an object, while [25] recently developed a two-hands and object interaction dataset. [5] proposed single and two-hand object interaction dataset using infrared camera for contact annotations. [30] developed a large-scale two-hand interaction dataset using semi-automatic annotation process with many close interactions. [46] used MoCap to obtain pose of full body, hand, and object during interaction and used it to generate realistic grasp on unseen objects.

In this work, we also introduce a challenging two-hands-and-object interaction dataset which we created using the optimization method of [12]. Our dataset is made of videos of two hands from different subjects manipulating an object from the YCB dataset [56], annotated with the 3D poses of the hands and the object. Our architecture already performs well on this dataset and constitutes a strong baseline.

2.2. Transformers in Computer Vision

Transformers have recently been increasingly gaining popularity for vision related problems [21]. Features are often extracted from a CNN backbone and different architectures have been proposed to solve object detection [8, 60], image classification [9], pose estimation [6, 17, 27, 28] and low-level image tasks [24, 57]. We refer the reader to [21] for a detailed survey.

[8, 60] proposed to combine a CNN backbone with a Transformer to detect objects in an image. [27] proposed to reconstruct the vertices of a single human body or hand from an RGB image using multiple Transformer encoder layers and achieved state-of-the-art performance. [28] improved [27] by using graph convolutions along with a Transformer encoder. [17] estimated a 3D pose from hand point-cloud data using a Transformer encoder-decoder architecture and proposed to generate query embeddings from input point-cloud instead of learning them as in [8, 60]. While these works are aimed at single hand pose estimation and their extension to two hands is non-trivial, our architecture is designed to estimate single and two-hands poses along with the object pose during hand-object interaction from the input RGB image.

In a closely related work, [6] solves the multi-person 2D pose estimation problem using detected keypoints and person centers, by associating the joint keypoints to the correct person center using attention. There are however several key differences: In our case, the hand centers get very close to each other during close interaction, and the approach in [6] would not be transferable. More importantly, hand joints are much more ambiguous than "body joints" as they look very similar to each other. Our method is also robust to undetected and falsely detected keypoints as we show in our discussions, while [6] cannot handle undetected keypoints. Further, we show that, by randomly sampling keypoints on the object, we can easily extend our method to 3D object pose estimation during hand-object interactions.

3. Method

As shown in Fig. 3, our architecture first detects keypoints that are likely to correspond to the 2D locations of hand joints and encodes them as input to the keypoint-joint association stage. The keypoints are encoded with their spatial locations and the image features at these locations. The self-attention layers in the Keypoint Transformer disambiguate the keypoints and associates them with different joint types and a background class. The (single) cross-attention layer then selects these "identity-aware keypoints" to predict root-joint-relative pose parameters of both hands, plus additional parameters such as the translation between the hands and hand shape parameters.

We detail below the keypoint detection and encoding step, how we use the Keypoint Transformer to predict the hands poses, the representations we considered for the 3D hand poses, and the loss used for training. We also explain how our approach can be extended to object pose estimation during hand-object interaction scenarios.

3.1. Keypoint Detection and Encoding

Given the input image, we first extract keypoints that are likely to correspond to 2D hand joint locations. To do this, we predict a single-channel heatmap $\mathcal{H}$ from the in-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:29:07 UTC from IEEE Xplore. Restrictions apply.

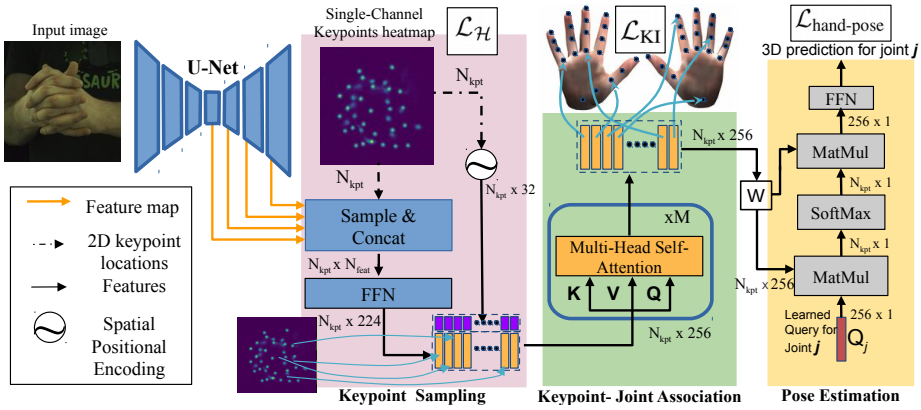


Figure 3. **Overview of our approach.** We detect keypoints which are potential locations of joints and encode them with CNN image features and spatial embedding (Section 3.1). From this information, the self-attention module creates context-aware keypoint features which are essential for associating each keypoint with the corresponding joint (Section 3.2). A cross-attention module finally predicts for each joint (using learned queries) the values required for computing the hand poses (Section 3.3). The exact nature of these values depends on the chosen representation of the hand pose (Section 3.4). Not all the keypoints have to correspond to a joint and not all the joints have to be detected as keypoints, which makes our approach very robust but still accurate as it relies on keypoints (as discussed in Section 2).

put image using a standard U-Net [42] architecture, and we keep its local maximums using a non-differentiable, non-maximum suppression operation. At this stage, we do not attempt to recognize which keypoint corresponds to which joint as it is a difficult task, and the predicted heatmap has only one channel. In practice, we keep a maximum of N_{hand} keypoints, with $N_{\text{hand}} = 64$, while the number of hand joints is 42 in total for 2 hands. The 2D keypoint locations are normalized to $[0, 1]$ range.

The ground truth heatmap $\mathcal{H}^*$ is obtained by applying a 2D Gaussian kernel of variance σ at each of the ground truth 2D joint locations and the U-Net is trained to predict the heatmap by minimizing the L2 loss.

We compute for each detected keypoint an appearance and spatial encoding to represent the keypoints as input to the next stage. As shown in Fig. 3, for the appearance part, we extract image features from the decoder of the U-Net network. More exactly, we sample feature maps at multiple layers of the U-Net decoder at the normalized keypoint locations using bilinear interpolation and concatenate them to form a 3968-D feature vector, which is then reduced down to a 224-D encoding vector using a 3-layer MLP. For the spatial encoding, we obtain 32-D sine positional encoding similar to [8] corresponding to the 2D location of the keypoint. We finally concatenate the appearance and spatial encodings to form a 256-D vector representation of the keypoint. The keypoint detector is pre-trained before fine-tuning it jointly with the rest of the pipeline.

3.2. Keypoint-Joint Association

For each keypoint K_i, we have now an encoding vector $\mathcal{F}_i \in \mathbb{R}^{256}$. We use these vectors as input to the multi-layer, multi-head self-attention module with N_{SA} layers. The self-attention [53] helps to model the relationship between the keypoints and create global context-aware feature $\mathcal{G}_i \in \mathbb{R}^{256}$, for each keypoint. Such context-aware features are necessary to associate the keypoints with different joint types using a "keypoint-joint association" loss we denote $\mathcal{L}_{\text{KI}}$. As a result of $\mathcal{L}_{\text{KI}}$, the keypoint features also now encode the joint identity information along with the localized CNN image features.

The identity of keypoint k is defined by (h_k, j_k), where h_k is the hand identity (left or right) and j_k is the joint index. We also use an additional 'background' identity for keypoints that are falsely detected. The keypoint identity is predicted using a feed-forward network (FFN) consisting of a 2-layer MLP, a linear projection layer and a softmax layer. We use the cross-entropy loss for $\mathcal{L}_{\text{KI}}$:

$$\mathcal{L}_{\text{KI}} = \sum_i \text{CE}((h_i, j_i), (h_i^*, j_i^*)) \,, \tag{1}$$

where (h_i^*, j_i^*) are the ground truth identities and CE denotes the cross-entropy loss. To obtain the ground truth identity for the detected keypoints, we associate them at training time with the closest reprojection of a ground truth 3D joint, if the distance is below a threshold γ. If there are no reprojected joints within a distance of γ, the keypoint is assigned to the background class. We empirically set $\gamma = 3$ pixels in our experiments. Similar to [8], the keypoint identities are predicted after each layer of self-attention module using FFNs with shared weights and the loss is applied to predictions of each layer.

The prediction can result in multiple keypoints assigned to the same joint identity and some keypoints assigned to the background class. As we discuss in Section 5, the keypoints associated to the background are ignored, while all the keypoints associated with a given joint are considered for estimating the pose of the corresponding joint by the

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:29:07 UTC from IEEE Xplore. Restrictions apply.

cross-attention module.

3.3. Pose Estimation from Identity-Aware Keypoints

The keypoint-joint association loss enables the keypoint features to also encode joint identity information along with the image features and spatial embeddings. We use a single cross-attention layer with learned joint queries to predict which keypoint(s) match the queried joint identities. The cross-attention operation [53] for a learned joint query $Q_j \in \mathbb{R}^{256}$ and features $\{\mathcal{G}_i\}_i$ is given by

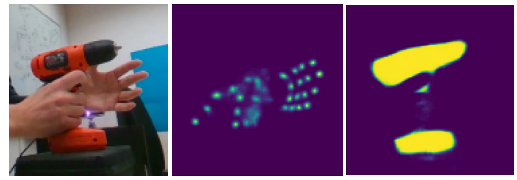
$$\mathrm{CA}\big(Q_j, G\big) = \mathrm{softmax}\left(\frac{Q_j^T W_K G}{16}\right)(W_V G)^T\,, \qquad (2)$$

where G is a matrix whose columns contain feature vectors $\{\mathcal{G}_i\}_i$, and W_K and W_V are learnable matrices of dimension 256×256. Similar to [53], the cross-attention features are added to Q_j to create a residual connection. The resulting features are transformed by a 3-layer MLP to map them to the pose space.

The number of joint queries depend on the pose representation. We consider 3 different representations and describe them in Section 3.4 and the suppl. mat. For example, we use 21 joint queries for each of the 21 joints per hand when using 2.5D pose representation. Along with the joint queries, one for each joint of the two hands, we use an additional learned query to predict the relative translation $T_{L\rightarrow R}$ between the hands, and the 10-D MANO hand shape parameters β. These are learned using the L1 loss. The MANO shape parameters are useful when predicting the pose using the MANO joint angle representation.

3.4. Hand Pose Representations and Losses

We consider three main hand pose representations: 3D joint locations, 2.5D [18, 30] joint locations, and MANO joint angles [41]. Previous methods [14, 15, 19, 38, 40] have noted that regressing model parameters such as joint angles is less accurate in terms of joint error than regressing the joint locations directly. However, regressing MANO joint angles provides access to the complete hand mesh required for modeling contacts and interpenetration during interactions [5, 15, 46] or for learning in a weakly supervised setup [3, 14, 23], which could be interesting for future extension of our method. As we show later in our experiments, the Keypoint Transformer enables the MANO joint angle representation to achieve competitive performance when compared to the joint location representation. We follow standard practice for these three representations. We detail them and their corresponding losses in the supplementary material for completeness.



(a) (b) (c)

Figure 4. **Keypoint detection for hands and object.** We train a U-net decoder to predict (b) a heatmap for all the joints together and (c) a segmentation map for the object from each we extract keypoints at random locations.

3.5. Object Pose Estimation

Our method generalizes easily to predict the 3D pose of an object together with 3D poses of hands. As shown in Fig. 4, along with the heatmap for the hand keypoints, we also predict a segmentation map of the object by adding an additional prediction head to the U-Net decoder. We then randomly select $N_{\text{obj}} = 20$ points from this segmentation map and refer to them as 'object keypoints'. We also tried estimating the heatmap of 2D reprojections of fixed points on the object mesh and selecting their local maximums as object keypoints and obtained similar results.

We encode the appearance and spatial locations of the object keypoints in a 256-D vector, exactly like the hand keypoints. Collectively, these keypoint encodings cover the object appearance, allowing us to predict the 3D rotation and translation of the object. The encodings of N_{obj} object keypoints and N_{hand} hand keypoints are provided *together* to the self-attention module.

Along with the hand keypoint identities (h_k, j_k) and the background identity described in Section 3.2, we rely on an additional identity for the object. During the keypoint association stage, all the keypoints originating from the object are associated with the 'object' identity, allowing the cross-attention module to only attend to object keypoints when estimating the object pose. Along with the joint queries that estimate the hand pose, we consider 2 additional queries in the cross-attention module and predict the 3D object rotation and 3D object translation relative to the right hand in a manner similar to that of hand pose. The object rotation is parameterized using the method proposed in [59].

We use a symmetry-aware object corner loss similar to [37] to train the network:

$$\mathcal{L}_{\text{obj-pose}} = \min_{R\in\mathcal{S}} \frac{1}{8}\sum_{i=1}^{8} ||P\cdot B_i - P^*\cdot R\cdot B_i||_2^2\,, \qquad (3)$$

where P and P^* denote the estimated and ground-truth object poses, B_i denotes the ith corner of the 3D bounding box for the object in rest pose, and $\mathcal{S}$ is the set of rotation matrices which, when applied to the object, does not change its appearance.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:29:07 UTC from IEEE Xplore. Restrictions apply.

3.6. End-to-End Training

We train our architecture end-end by minimizing the sum of the losses introduced above:

$$\mathcal{L} = \mathcal{L}_{\mathcal{H}} + \mathcal{L}_{\text{KI}} + \mathcal{L}_{\mathcal{T}} + \mathcal{L}_{\text{hand-pose}} + \mathcal{L}_{\text{obj-pose}} \,, \quad (4)$$

where $\mathcal{L}_{\text{hand-pose}}$ is the loss on the hand poses (detailed in the suppl. mat.) and $\mathcal{L}_{\mathcal{T}}$ is the L1 loss for relative translation between the two hands. Note that the keypoint detector is pretrained before training the entire network end-to-end. More optimization details are also given in the suppl. mat.

4. Evaluation

We evaluated our method on three challenging hand interaction datasets: InterHand2.6M, HO-3D, and our H_2O-3D dataset we introduce with this paper. We discuss them below.

4.1. InterHand2.6M

Training and test sets. InterHand2.6M [30] is a recently published two-hand interaction dataset with many challenging poses. It was annotated semi-automatically and contains 1.36M train images and 849K test images.

Metrics. We report the Mean Per Joint Position Error (MPJPE) and the Mean Relative-Root Position Error (MRRPE) to evaluate the root-relative hand pose and the translation between the hands respectively, as in [30].

Baselines. We consider two Transformer-based baseline architectures. The first baseline ('CNN+SA') provides the low-resolution (32× downsampled) CNN feature maps after flattening along the spatial dimensions as input to the Transformer encoder containing self-attention (SA) modules. The output tokens of the encoder are concatenated and the pose is predicted using an MLP. The second baseline ('CNN+SA+CA') is more similar to DETR [8], where the low-resolution CNN feature maps are provided to the Transformer encoder-decoder architecture. The Transformer decoder contains SA and cross-attention (CA) modules. The queries in the decoder are learnt and the pose is predicted using FFN similar to our Keypoint Transformer. We provide more details about the baselines in the suppl. mat. These baselines help to understand the importance of keypoint sampling and selection for pose estimation.

	MPJPE (mm)			MRRPE
	Single Hand	Two Hands	All	(mm)
CNN+SA	13.53	16.87	15.31	33.84
CNN+SA+CA (DETR [8])	12.81	15.94	14.48	32.87
InterNet [30]	12.16	16.02	14.22	32.57
Ours	**10.99**	**14.34**	**12.78**	**29.63**
Dong et al. [22]	-	-	12.08	-
Ours	9.10	11.98	**11.30**	21.89
Fan et al. [10]	11.32	15.57	-	**30.51**
Ours	**11.08**	**15.33**	13.41	30.87

Table 1. **Comparison with 2 baselines and the state-of-the-art methods on InterHand2.6M [30].** We compare with [10, 22, 30] using the different train/test splits reported in their works.

Results. Table 1 compares the accuracy of our method with the state-of-the-art method InterNet [30], and the two baselines, when using the 2.5D pose representation. Our method achieves 10% higher accuracy than InterNet, which is a CNN-based architecture, and 16% and 12% higher accuracy than the two baselines, respectively. The higher accuracy of 'CNN+SA+CA' w.r.t 'CNN+SA' baseline demonstrates that soft-selection of image features by the decoder improves the accuracy. Further, the higher accuracy (12%) of our Keypoint-Transformer w.r.t the 'CNN+SA+CA' architecture shows that use of keypoint features for pose estimation instead of features from the entire image increases the overall accuracy.

We compare our method with [22] and [10] using their train and test splits. [10, 22] use per-joint heatmaps coupled with joint visibility and segmentation guided features to improve the accuracy of the pose estimation, thus resulting in the model complexity that is higher than InterNet [30]. Our method with a model complexity same as [30] (see Section 5) still outperforms these state-of-the-art methods. We show qualitative results in Fig. 5 and in the suppl. mat.

4.2. HO-3D

Training and test sets. The HO-3D [12] dataset contains automatically annotated hand-object interaction sequences of a right hand and an object from the YCB [56] dataset. It contains 66K training images and 11K test images. We consider only objects seen in the training set for evaluation.

Metrics. As in [12], we report the mean joint error after scale-translation alignment of the root joint and the area-under-the-curve (AUC) metrics to evaluate the hand pose. The object pose is computed w.r.t to the hand frame of reference and is evaluated using the standard Maximum Symmetry-Aware Surface Distance (MSSD) [16], as it considers the symmetricity of objects.

Results. We use 3D joint representation to estimate the hand pose. Table 2 compares the accuracy of the proposed hand-object pose estimation method with other approaches. Keypoint Transformer performs significantly better than previous methods [12, 14, 15]. As [12, 14, 15] do not consider symmetricity of objects during training and evaluation, we also report our results in a similar setting. Qualitative results are shown in Fig. 6 and the suppl. mat.

4.3. H_2O-3D

Training and test sets. We introduce a dataset named H_2O-3D comprising sequences of two hands manipulating an object automatically annotated with the 3D poses of the

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:29:07 UTC from IEEE Xplore. Restrictions apply.

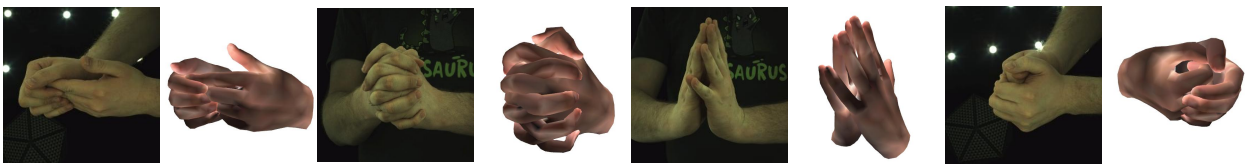
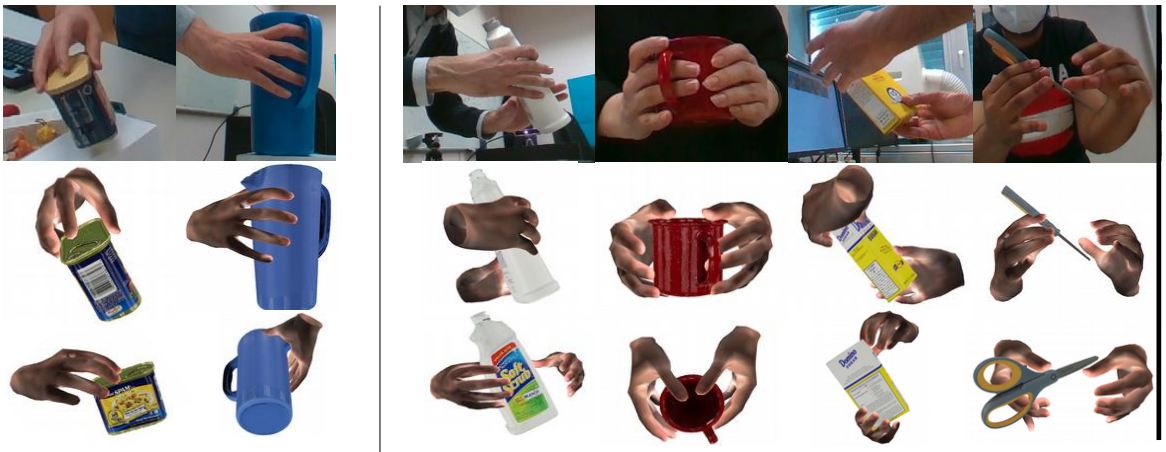


Figure 5. **Qualitative results on InterHand2.6M [30].** Our method obtains accurate poses of hands during complex interactions. We show the estimated MANO model from a different view.



HO-3D H_2O-3D

Figure 6. **Qualitative results for our method on the H_2O-3D and HO-3D datasets**. Our method recovers poses even under large occlusions by the object and achieves state-of-the-art results on HO-3D while serving as a strong baseline for our new dataset H_2O-3D. Note that some objects (columns 2&4) are considered to be rotationally symmetric along the z-axis.

	Camera Intrinsics	Image Crop	Joint Error	Mean Joint AUC	MSSD (Object Pose Error)
[12]	Yes	Yes	3.04	0.49	-
[15]	No	Yes	3.18	0.46	-
[14]	Yes	No	3.69	0.37	11.99
Ours	No	Yes	**2.57**	**0.54**	**7.02**

Table 2. **Accuracy of our method on the HO-3D dataset for hand and object pose estimation.** Our method outperforms previous methods by a large margin.

hands and the object, by extending the work of [12] to consider two hands. Figs. 1 and 6 show some images. Five different subjects manipulate 10 different objects from YCB using both hands with a functional intent. We captured 60'998 training images and 15'342 test images using a 5 RGBD cameras multi-view setup. The H_2O-3D test set contains 7 objects seen in the training set and 1 unseen object. More details are provided in the supplementary material. H_2O-3D is significantly more challenging than previous hand interaction datasets as there are many large occlusions between the hands and the objects.

Metrics and Results We use the 3D joint representation for the hand pose and evaluate the accuracy using the MPJPE and MRRPE metrics (see Section 4.1) for the hand and the MSSD metric for the object (see Section 4.2). Details about the angle of symmetry for different objects considered during training and evaluation is provided in the suppl. mat. Due to large occlusions of the object by the hands, a portion of images are unsuitable for object pose estimation. We identify these images as the ones whose ground truth object segmentation area is less than 2% of the cropped image area and exclude them from the object pose estimation during training and evaluation. We also used the HO-3D train split and mirrored the images randomly during training to obtain right hand- and left hand-only images, to later combine with the training set of H_2O-3D.

Our method achieves an MPJPE of 3.09 cm and an MRRPE of 8.28 cm on this dataset. Due to large occlusions by the object, estimating the translation between the hands is more challenging and the MRRPE is about 2.5 times worse than on InterHands2.6M which does not contain objects. On objects, our method achieves MSSD values of 7.96 cm. We provide object-specific MSSD values in the supplementary material. Fig. 6 shows qualitative results.

5. Discussion

We report here the results of experiments we perform using InterHand2.6M (V0.0) to understand better our method.

Visualization of cross-attention. We visualize the cross-attention weights in Fig. 7 for two joint queries of the left-index finger. When the joint is not occluded, as in the first row, each joint query attends to the keypoint whose location coincides with the corresponding joint location in the

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:29:07 UTC from IEEE Xplore. Restrictions apply.

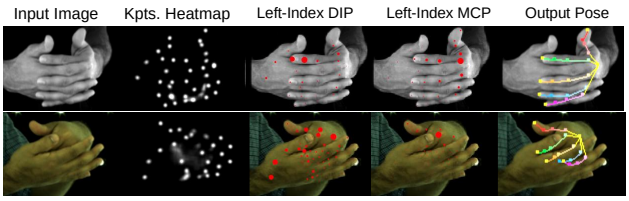


Figure 7. **Visualizing the cross-attention weights for two joint queries of the left hand.** The radius of the red circles are proportional to the weights. When the joint is occluded like the DIP joint on the second row, nearby visible keypoints are selected by the attention mechanism for pose estimation.

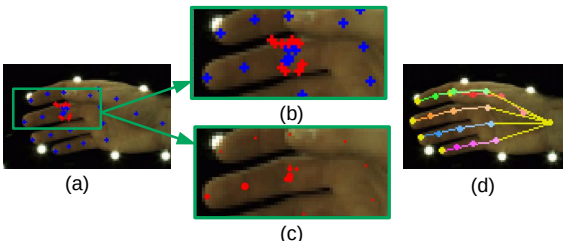


Figure 8. **Robustness to noisy keypoints.** In this example, we added noisy keypoints around the middle finger PIP joint. Most of the noisy keypoints are predicted to belong to background class (in red), while some are associated with the PIP joint (in blue). The noisy keypoints associated with the PIP joint have all higher cross-attention weights (c) and are considered for final pose estimation.

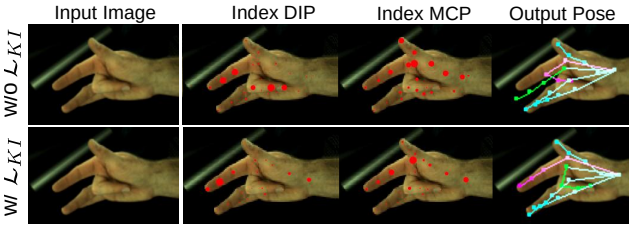


Figure 9. **Cross-attention with and without the keypoint-joint association loss $\mathcal{L}_{\mathbf{KI}}$.** $\mathcal{L}_{\text{KI}}$ makes the keypoints 'identity-aware', resulting in higher accuracy.

image. In other words, local image features at the location of the joint are used to estimate the pose of that joint. We believe this property of using local image features helps in achieving higher accuracy than other CNN-based approaches [14, 15, 30]. In the second row of Fig. 7, the left index finger is occluded except for the MCP joint and no keypoints are detected for the invisible joints. The cross-attention module *selects* nearby visible keypoints resulting in a more global-level feature for estimating the joint pose.

Robustness to noisy keypoints. To demonstrate the robustness of our method, we added incorrect keypoints around the detected keypoints. As shown in Fig. 8 and supplementary material, most of these keypoints are labeled as background and all the keypoints that are assigned to the same joint are considered equally for the pose estimation.

Importance of the keypoint-joint association loss $\mathcal{L}_{\mathbf{KI}}$. $\mathcal{L}_{\text{KI}}$ helps the cross-attention module to select the appropriate features for pose estimation as visualized in Fig. 9. Further, $\mathcal{L}_{\text{KI}}$ improves MPJPE by 10% (17.08 mm v/s 18.91 mm) and MRRPE by 15% (33.14 mm v/s 38.96 mm) on interacting hand images.

	Camera Intrinsics	MPJPE (mm) Single Hand	Two Hands	All	MRRPE (mm)
3D	No	12.42	**17.08**	14.76	**33.14**
2.5D	Yes	**11.73**	17.69	**14.73**	34.40
θ	No	15.36	20.61	18.01	37.91

Table 3. **Accuracy obtained with the 3 different pose representations.**

	$N_{CA}=1$, Varying N_{SA} 0	3	6	$N_{SA}=6$, Varying N_{CA} 1	3	6	[30]
Single Hand	12.34	11.77	11.24	11.24	11.14	11.08	12.63
Two Hands	16.93	15.55	15.44	15.44	15.35	15.33	17.36

Table 4. **3D pose accuracy (MPJPE, in mm) for different numbers of self-attention (N_{SA}) and cross-attention (N_{CA}) layers.**

	Resnet-18	Resnet-34	Resnet-50	[30] (Resnet-50)
Total Params	28M	38M	48M	48M
Single Hand	11.67	11.99	11.28	12.63
Two Hands	16.78	16.41	15.32	17.36

Table 5. **3D pose accuracy (MPJPE, in mm) for different backbones.**

Accuracy with different pose representations. Table 3 compares the accuracy of the 3 hand pose representations that we consider. While the accuracy of the 3D and 2.5D representations are similar, the joint angle representation results in lower accuracy, in line with the observation from previous works [14, 15, 19, 38, 40].

Effect of the number of self-attention (SA) and cross-attention (CA) layers. Table 4 reports the MPJPE with different combinations of SA and CA layers. Even in the absence of any SA layers, our method outperforms [30]. Adding more CA layers has little effect on the accuracy.

Effect of the number of parameters. Table 5 reports the MPJPE for different CNN backbones. While larger backbones improve the accuracy, our method outperforms [30] even with a Resnet-18 backbone with approximately half the total number of parameters.

6. Conclusion

We showed that, by integrating a keypoint detector into a Transformer architecture, we could predict 3D poses of hands and objects from very challenging images, in a much more accurate way than a standard Transformer architecture does. As we rely on keypoints, we believe that our approach is more general and could be applied to other problems, such as human and other articulated objects pose prediction and object category pose prediction [39].

Acknowledgments. This work was supported by the Christian Doppler Laboratory for Semantic 3D Computer Vision, funded in part by Qualcomm Inc, and Chistera IPalm.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:29:07 UTC from IEEE Xplore. Restrictions apply.

References

[1] Anil Armagan, Guillermo Garcia-Hernando, Seungryul Baek, Shreyas Hampali, Mahdi Rad, Zhaohui Zhang, Shipeng Xie, Ming-xiu Chen, Boshen Zhang, F. Xiong, Yang Xiao, Zhiguo Cao, Junsong Yuan, Pengfei Ren, Weiting Huang, Haifeng Sun, Marek Hrúz, Jakub Kanis, Zdenek Krnoul, Qingfu Wan, Shile Li, Linlin Yang, Dongheui Lee, Angela Yao, Weiguo Zhou, Sijia Mei, Yunhui Liu, Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Philippe Weinzaepfel, Romain Brégier, Grégory Rogez, Vincent Lepetit, and Tae-Kyun Kim. Measuring Generalisation to Unseen Viewpoints, Articulations, Shapes and Objects for 3D Hand Pose Estimation Under Hand-Object Interaction. In *European Conference on Computer Vision*, 2020. 2

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the Envelope for RGB-Based Dense 3D Hand Pose Estimation via Neural Rendering. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1

[3] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-Supervised Domain Adaptation via GAN and Mesh Model for Estimating 3D Hand Poses Interacting Objects. In *Conference on Computer Vision and Pattern Recognition*, 2020. 5

[4] Luca Ballan, Aparna Taneja, Juergen Gall, Luc Van Gool, and Marc Pollefeys. Motion Capture of Hands in Action Using Discriminative Salient Points. In *European Conference on Computer Vision*, 2012. 2

[5] 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. 2, 3, 5

[6] Guillem Braso, Nikita Kister, and Laura Leal-Taixé. The Center of Attention: Center-Keypoint Grouping Attention for Multi-Person Pose Estimation. In *International Conference on Computer Vision*, 2021. 3

[7] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing Hand-Object Interactions in the Wild. In *arXiv Preprint*, 2020. 2

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*, 2020. 2, 3, 4, 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *arXiv Preprint*, 2020. 3

[10] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael Black, and Otmar Hilliges. Learning to Disambiguate Strongly Interacting Hands via Probabilistic Per-Pixel Part Segmentation. In *International Conference on 3D Vision*, 2021. 2, 3, 6

[11] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations. In *Conference on Computer Vision and Pattern Recognition*, 2018. 3

[12] 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. 1, 2, 3, 6, 7

[13] Shangchen Han, Beibei Liu, R. Cabezas, Christopher D. Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, Asaf Nitzan, Gang Dong, Yuting Ye, Lingling Tao, Chengde Wan, and Robert Wang. MEgATrack: Monochrome Egocentric Articulated Hand-Tracking for Virtual Reality. *IEEE Transactions on Robotics and Automation*, 39, 2020. 1, 2

[14] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging Photometric Consistency over Time for Sparsely Supervised Hand-Object Reconstruction. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3, 5, 6, 7, 8

[15] 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, 2, 3, 5, 6, 7, 8

[16] Tomás Hodan, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiri Matas. BOP Challenge 2020 on 6D Object Localization. In *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, 2020. 6

[17] Lin Huang, Jianchao Tan, Ji Liu, and Junsong Yuan. Hand-Transformer: Non-Autoregressive Structured Modeling for 3D Hand Pose Estimation. In *European Conference on Computer Vision*, 2020. 3

[18] 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, 5

[19] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-End Recovery of Human Shape and Pose. In *Conference on Computer Vision and Pattern Recognition*, 2018. 5, 8

[20] 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. 2, 3

[21] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Khan, and Mubarak Shah. Transformers in Vision: A Survey. In *arXiv Preprint*, 2021. 3

[22] Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-End Detection and Pose Estimation of Two Interacting Hands. In *International Conference on Computer Vision*, 2021. 2, 3, 6

[23] Dominik Kulon, Riza Alp Güler, Iasonnas Kokkinos, Michael Bronstein, and Stefanos Zafeiriou. Weakly-Supervised Mesh-Convolutional Hand Reconstruction in the Wild. In *Conference on Computer Vision and Pattern Recognition*, 2020. 5

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:29:07 UTC from IEEE Xplore. Restrictions apply.

[24] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization Transformer. In *arXiv Preprint*, 2021. 3

[25] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two Hands Manipulating Objects for First Person Interaction Recognition. In *International Conference on Computer Vision*, 2021. 3

[26] Nikolaos Kyriazis and Antonis A. Argyros. Scalable 3D Tracking of Multiple Interacting Objects. In *Conference on Computer Vision and Pattern Recognition*, 2014. 2

[27] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-End Human Pose and Mesh Reconstruction with Transformers. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3

[28] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh Graphormer. In *International Conference on Computer Vision*, 2021. 3

[29] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image. In *European Conference on Computer Vision*, 2020. 1

[30] 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, 2, 3, 5, 6, 7, 8

[31] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel Otaduy, Dan Casas, and Christian Theobalt. Real-Time Pose and Shape Reconstruction of Two Interacting Hands with a Single Depth Camera. *IEEE Transactions on Robotics and Automation*, 38, 2019. 2

[32] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Training a Feedback Loop for Hand Pose Estimation. In *International Conference on Computer Vision*, 2015. 1

[33] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Full DOF Tracking of a Hand Interacting with an Object by Modeling Occlusions and Physical Constraints. In *International Conference on Computer Vision*, 2011. 2

[34] I. Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Tracking the Articulated Motion of Two Strongly Interacting Hands. In *Conference on Computer Vision and Pattern Recognition*, 2012. 2

[35] Paschalis Panteleris, Nikolaos Kyriazis, and Antonis A. Argyros. 3D Tracking of Human Hands in Interaction with Unknown Objects. In *British Machine Vision Conference*, 2015. 2

[36] Paschalis Panteleris, Iason Oikonomidis, and Antonis A. Argyros. Using a Single RGB Frame for Real Time 3D Hand Pose Estimation in the Wild. In *IEEE Winter Conference on Applications of Computer Vision*, 2018. 1, 2

[37] Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation. In *International Conference on Computer Vision*, 2019. 5

[38] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising Human Mesh Estimation with Texture Consistency. In *International Conference on Computer Vision*, 2019. 5, 8

[39] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G Derpanis, and Kostas Daniilidis. 6-dof object pose from semantic keypoints. In *International Conference on Robotics and Automation (ICRA)*, 2017. 8

[40] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to Estimate 3D Human Pose and Shape from a Single Color Image. In *Conference on Computer Vision and Pattern Recognition*, 2018. 1, 5, 8

[41] 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. 2, 3, 5

[42] 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. 4

[43] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica Hodgins, and Takaaki Shiratori. Constraining Dense Hand Surface Tracking with Elasticity. *IEEE Transactions on Robotics and Automation*, 39, 2020. 2

[44] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly Supervised 3D Hand Pose Estimation via Biomechanical Constraints. In *European Conference on Computer Vision*, 2020. 1

[45] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-Time Joint Tracking of a Hand Manipulating an Object from RGB-D Input. In *European Conference on Computer Vision*, 2016. 2

[46] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *European Conference on Computer Vision*, 2020. 3, 5

[47] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, Arran Topalian, Erroll Wood, Sameh Khamis, Pushmeet Kohli, Shahram Izadi, Richard Banks, Andrew Fitzgibbon, and Jamie Shotton. Efficient and Precise Interactive Hand Tracking through Joint, Continuous Optimization of Pose and Correspondences. *IEEE Transactions on Robotics and Automation*, 35, 2016. 2

[48] Jonathan Taylor, Vladimir Tankovich, Danhang Tang, Cem Keskin, David Kim, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. Articulated Distance Fields for Ultra-Fast Tracking of Hands Interacting. *IEEE Transactions on Robotics and Automation*, 36, 2017. 2

[49] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 3

[50] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-Supervised Learning of Motion Capture. In *Advances in Neural Information Processing Systems*, 2017. 1

[51] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing Hands

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:29:07 UTC from IEEE Xplore. Restrictions apply.

in Action Using Discriminative Salient Points and Physics Simulation. *International Journal of Computer Vision*, 118, 2016. 2

[52] Dimitrios Tzionas and Juergen Gall. 3D Object Reconstruction from Hand-Object Interactions. In *International Conference on Computer Vision*, 2015. 2

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017. 4, 5

[54] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel Otaduy, Dan Casas, and Christian Theobalt. RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video. *IEEE Transactions on Robotics and Automation*, 39, 2020. 1, 2

[55] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular Total Capture: Posing Face, Body, and Hands in the Wild. In *Conference on Computer Vision and Pattern Recognition*, 2019. 2

[56] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. *Science*, 2018. 3, 6

[57] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning Texture Transformer Network for Image Super-Resolution. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3

[58] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting Two-Hand 3D Pose and Shape Reconstruction from Single Color Image. In *International Conference on Computer Vision*, 2021. 2, 3

[59] Y. Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *Conference on Computer Vision and Pattern Recognition*, 2019. 5

[60] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference for Learning Representations*, 2021. 3

[61] Christian Zimmermann and Thomas Brox. Learning to Estimate 3D Hand Pose from Single RGB Images. In *International Conference on Computer Vision*, 2017. 1

[62] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan C. Russell, Max Argus, and Thomas Brox. FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape from Single RGB Images. In *International Conference on Computer Vision*, 2019. 2, 3

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:29:07 UTC from IEEE Xplore. Restrictions apply.