HOT-Net: Non-Autoregressive Transformer for 3D Hand-Object Pose Estimation

Lin Huang
University at Buffalo, SUNY
Buffalo, New York
lhuang27@buffalo.edu

Jianchao Tan
Y-tech, Kwai Inc.
Seattle, Washington
jianchaotan@kuaishou.com

Jingjing Meng
University at Buffalo, SUNY
Buffalo, New York
jmeng2@buffalo.edu

Ji Liu
Y-tech, Kwai Inc.
Seattle, Washington
jiliu@kuaishou.com

Junsong Yuan
University at Buffalo, SUNY
Buffalo, New York
jsyuan@buffalo.edu

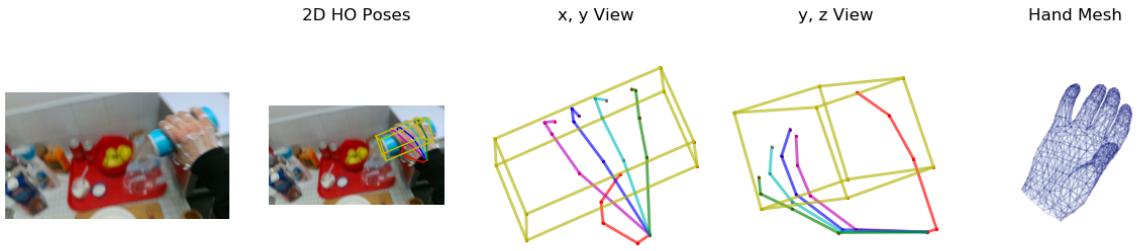

Figure 1: 3D hand-object pose estimation: our proposed method fully leverages correlations among hand joints and object bounding box corners for 3D hand-object pose estimation from a single RGB image. From left to right: input image, 2D hand-object (HO) pose, multi-views of 3D HO pose, and the reconstructed hand mesh. Results in Fig.4 are ordered in the same way.

ABSTRACT

As we use our hands frequently in daily activities, the analysis of hand-object interactions plays a critical role to many multimedia understanding and interaction applications. Different from conventional 3D hand-only and object-only pose estimation, estimating 3D hand-object pose is more challenging due to the mutual occlusions between hand and object, as well as the physical constraints between them. To overcome these issues, we propose to fully utilize the structural correlations among hand joints and object corners in order to obtain more reliable poses. Our work is inspired by structured output learning models in sequence transduction field like Transformer encoder-decoder framework. Besides modeling inherent dependencies from extracted 2D hand-object pose, our proposed Hand-Object Transformer Network (HOT-Net) also captures the structural correlations among 3D hand joints and object corners. Similar to Transformer's autoregressive decoder, by considering structured output patterns, this helps better constrain the output space and leads to more robust pose estimation. However, different from Transformer's sequential modeling mechanism, HOT-Net adopts a novel non-autoregressive decoding strategy for 3D hand-object pose estimation. Specifically, our model removes the Transformer's dependence on previously generated results and explicitly feeds a reference 3D hand-object pose into the decoding process to provide equivalent target pose patterns for parallely localizing each 3D keypoint. To further improve physical validity of estimated hand pose, besides anatomical constraints, we propose a cooperative pose constraint, aiming to enable the hand pose to cooperate with hand shape, to generate hand mesh. We demonstrate real-time speed and state-of-the-art performance on benchmark hand-object datasets for both 3D hand and object poses.

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**.

KEYWORDS

3D Hand and Object Poses; Structured Learning; Transformer

ACM Reference Format:
Lin Huang, Jianchao Tan, Jingjing Meng, Ji Liu, and Junsong Yuan. 2020. HOT-Net: Non-Autoregressive Transformer for 3D Hand-Object Pose Estimation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413775

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413775

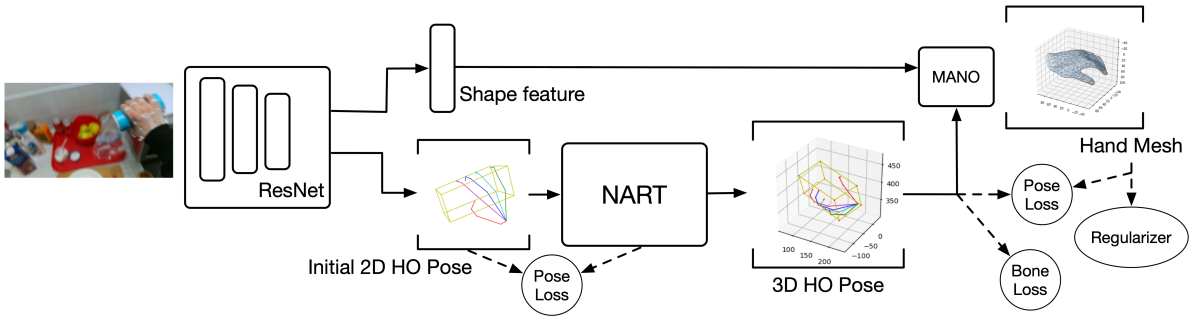

Figure 2: The architecture of HOT-Net. The model consists of 3 components. The ResNet first encodes image features and regresses 2D HO pose. Then, both information gets passed into our proposed Non-AutoRegressive Transformer (NART) module to generate 3D HO pose. The hand pose will further cooperate with shape feature in order to generate a MANO mesh.

1 INTRODUCTION

Interaction with surrounding objects has always been an essential part of human daily behavior, especially through hands sensing and manipulating everyday objects. Thus, understanding hand-object interactions (HOI) and recognizing 3D hand and object poses are vital for analyzing and imitating human behavior, leading to a great many applications in human-computer interactions, virtual reality, augmented reality, and robotics, *etc.*

In the last decade, we have witnessed a rapid advance towards both 3D hand pose estimation [4, 5, 11, 13, 14, 24, 30, 33, 41, 42, 47, 47, 51, 56, 57, 60] and object pose estimation [25, 26, 37, 38, 46, 52, 53, 55] in isolation. However, joint estimation for both 3D hand and object poses from a single RGB frame has received far less attention and remains a challenging task. Besides the common issues with articulated and rigid pose estimation from RGB images, including complex pose variations, depth/scale ambiguities, clutter, and self-occlusions, the complex HOI scenarios bring in another challenge: the hand and manipulated object would occlude each other, possibly leading to severe mutual occlusion. Nonetheless, we also observe that jointly accounting for the presence of hand and object helps overcome the above issues since hand and object are highly correlated under HOI scenarios. Specifically, different object poses and categories induce different hand grasps while the hand pose can also provide hints on the object pose and category. Thus, the problem boils down to: how to capture and use the correlations between hand and object poses in order to jointly model a kinematically feasible hand-object pose configuration space?

Early works [7, 16, 19, 30, 31, 39] estimate both 3D hand and object poses separately, thus neglecting the constraints between the hand pose and the pose of object being interacted. Recent works [9, 21, 34, 45, 48] start to exploit the existence of hand and object as effective evidence and jointly model hand and object poses. However, they tend to rely on a coarse modeling without a finer exploration of useful correlations or ignore the inherent dependencies of the 3D target pose, which causes unrealistic pose configurations.

To tackle this problem, we make use of the connection between the structured pose prediction problem and the structured sequence transduciton task in Natural Language Processing (NLP) field. Typical transduction algorithms [1, 50], following an encoder-decoder framework, autoregressively condition each output token generation on the relevant input tokens' features and the inherent dependencies among previously generated output tokens. By considering the structured output patterns, this comprehensive modeling strategy has led to drastic improvements in generating semantically and syntactically valid results, such as image captions and language translations. Therefore, we propose to leverage the state-of-the-art transduction model, the Transformer network, as our central building block, aiming to exploit the structured output learning mechanism in NLP field for reliable 3D hand-object pose estimation.

Following the Transformer encoder-decoder framework, we first capture the structural correlations among extracted 2D hand joints and object corners. Specifically, we convert the input HOI image into a concatenation of estimated 2D hand-object pose with image context features and we feed them into an attention-based encoder for capturing 3D information embedded in the 2D spatial configuration and image features. The encoding process yields point-wise features for subsequent decoding.

Then, we pay attention to the inherent dependencies of the 3D hand-object pose. Human hands are inherently structured and highly correlated with manipulated objects. For instance, the ring finger can constrain the motion of middle finger by bending backward and the hand-object contact points should always remain at the rigid object surface without interpenetration [21]. However, most works [11, 12, 32, 33] simply regard the complex 3D pose as a set of independent 3D keypoints while a few studies have enforced geometrical constraints [21, 23, 58, 59]. Nonetheless, due to the large variations in HOI, an algorithm that can adaptively model the inherent correlations among 3D keypoints should be more helpful compared with pre-defined constraints. Inspired by the autoregressive decoding mechanism used in Transformer decoder, we can impose necessary knowledge of 3D output pose by conditioning each 3D keypoint generation on previously generated results. However, the autoregressive factorization causes high inference latency. Additionally, the sequential decoding mechanism would only provide dependencies among "previous" keypoints for current 3D keypoint generation, given a specified order of 3D hand and object keypoints. However, we can find cases where each hand joint and object corner are inter-correlated with both "previous" and

"future" keypoints. Therefore, this might lead to inferior and invalid results if we only take sequential dependencies into consideration.

To speed up the inference while feeding necessary 3D hand-object pose patterns to the decoder, we propose to replace the autoregressive factorization with a novel non-autoregressive structured learning mechanism designed for joint estimation of 3D hand-object pose. Unlike recently proposed Non-AutoRegressive Transformer (NART) models [17, 18, 44, 54], most of which simply expose a modified copy of input tokens to decoder resulting in loss of knowledge from structured output, we design a structured-reference extractor to feed a 3D reference hand-object pose into the decoder. Our goal is to exploit the inherent dependencies of the reference pose as an approximation to that of the target pose.

Beyond drawing features from extracted 2D hand-object pose and reference hand-object pose, our decoder imitates the Transformer to further utilize the captured reference pose dependencies as queries to attend over the point-wise features output from encoder. This step helps prioritize the set of informative features, towards each 3D keypoint generation. Finally, we can merge the attention-weighted information with the reference pose dependencies to find precise 3D hand joint and object corner locations.

Moreover, since hands are highly articulated, to further optimize the geometric validity of our resulting hand pose space besides using the help from object, we adopt a hand anatomical constraint [29] composed of a bone length loss and a bone direction loss in order to further constrain the structural correlations between different 3D hand joints. We also propose a novel cooperative pose constraint. Previous works tend to ignore the correlations between articulated pose and other visual factors such as shape within a given frame. For example, the hand pose feature should always cooperate with the hand shape feature in order to generate a reasonable 3D hand mesh. Besides using target 3D coordinates as supervision, this can impose another form of supervision to aid the articulated pose estimation task.

Our major contributions are summarized as follows: (1) We propose a novel structured modeling framework, HOT-Net, which is based on Non-Autoregressive Transformer, for joint 3D articulated and rigid object pose estimation. Our method models the strong correlations among hand joints and object corners in a fine-grained and comprehensive manner. (2) We introduce a novel cooperative pose constraint to improve physical validity of hand structure. Our scheme depends on the cooperative relationship between hand pose and hand shape to reconstruct hand mesh. (3) Our extensive experimental results on benchmarks show that our method consistently outperforms previous methods. We also conduct comprehensive ablation studies to gain better understandings of our approach.

2 RELATED WORK

We now review relevant works on hand-object interaction.

Hand-Object Interaction. Early works [7, 16, 19, 30, 31, 39] tend to ignore the strong correlations between hand and object under HOI scenarios. Subsequent approaches [2, 34–36, 43, 48, 49] start to show the effectiveness to exploit the interaction of hand and object as constraints especially for depth input or multi-view camera system. Oberweger et al. [34] adopts an iterative depth reconstruction strategy for joint hand-object pose estimation. Due to the high cost to set up a multi-view camera system and the huge power consumption using active depth sensors, researchers turn to RGB frames. Recent works [9, 21, 45] have shown various methods for jointly understanding hand-object poses, which however do not yet fully leverage the correlations between the hand and the manipulated rigid object. Tekin et al. [45] proposes a unified YOLO-based framework for jointly regressing 3D hand-object pose. Nonetheless, they directly output both poses without considering the fine-grained correlations between hand joints and object corners, which might lead to physically invalid results. Hasson et al. [21] instead employs a novel contact constraint to enforce valid hand-object configuration. However, for this precise contact modeling, it usually requires dense annotations, which is difficult to obtain. Doosti et al. [9] directly models the relations among 2D hand and object keypoints using an adaptive graph convolutional network, but it ignores utilizing the inherent dependencies among 3D hand and object keypoints.

3 METHODOLOGY

Our proposed HOT-Net for joint 3D hand-object pose estimation is illustrated in Fig. 2, which comprises three modules. Given a HOI RGB frame, the first module serves as a backbone network for image context feature extraction and estimation of 2D hand and object poses. Our second module, the proposed Non-AutoRegressive Transformer (NART), fully captures the strong correlations between hand joints and object corners for robust 3D hand-object pose estimation. Our third module further improves the physical validity of resulting hand articulated pose via anatomy inspired constraints and cooperation with hand shape features for mesh reconstruction using the parametric MANO hand model [40].

3.1 Revisiting Transformer

Transformer [50], which is established as state-of-the-art sequence transduction model, adopts a comprehensive structured learning mechanism. Unlike typical deep learning models, most of which tend to ignore the inherent dependencies of structured output data, the AutoRegressive Transformer (ART), following an encoder-decoder framework, utilizes attention-based autoregressive decoding mechanism to condition each token generation on previously generated output patterns and relevant input features selected by the output patterns. These strategies effectively help constrain the structured output space and achieve superior performance.

Thus, in the goal to obtain physically valid 3D hand-object pose configuration space, we translate this structured modeling framework into our case. While keeping the attention mechanism, we extend the autoregressive decoding to a more suited strategy for joint estimation of 3D hand and object poses. In the following sections, we continue revisiting both key concepts.

3.1.1 Autoregressive Decoding. As an essential piece of the Transformer framework, the autoregressive decoding mechanism allows each token generation to access previously generated results. Specifically, given a source sentence $X = \{x_1, ..., x_N\}$, the decoding strategy factors the distribution of output sequence $Y = \{y_1, ..., y_T\}$ into a series of conditional probabilities with a left-to-right structure:

$$p_{ART}\left(Y|X;\theta\right) = \prod_{t=1}^{T} p\left(y_t|y_{1:t-1}, x_{1:N};\theta\right), \tag{1}$$

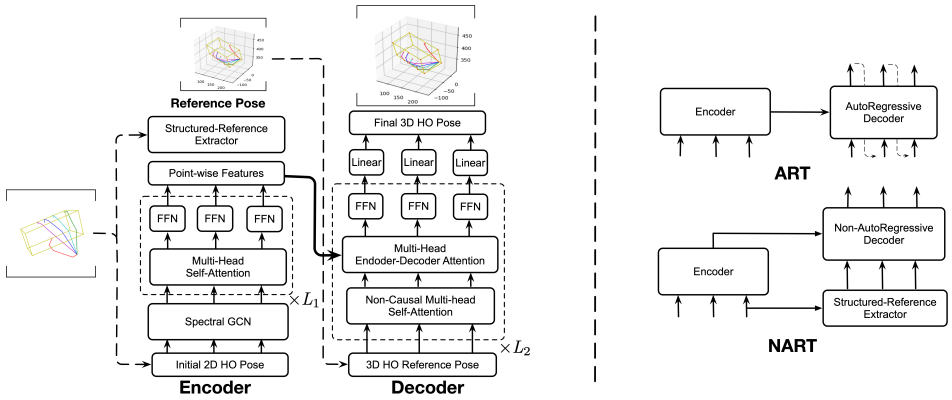

Figure 3: Left: Overview of our NART module. The encoder computes enhanced point-wise features given 2D HO pose while the structured-reference extractor feeds a reference 3D HO pose to the decoder. Then, the decoder explores necessary 3D pose patterns and use this information to attend over the encoder output. Our final estimation is thus based on the HO pose patterns and selective encoder output. Right: Simplified diagram to represent the ART model and our NART model. Typical ART model utilizes both previous output information and the encoder output, our model instead uses the encoder output and more reasonable structured output patterns. L_1 and L_2 is the number of layer for encoder and decoder, respectively.

where N and T denote the sequence length, and θ is the model parameters. According to Eq. 1, the autoregressive modeling strategy helps provide structured output patterns for each token generation, but it also suffers from heavy inference latency since t^{th} token y_t generation relies on previously generated output $y_{1:t-1}$. Moreover, since hand and object 3D keypoints are highly intercorrelated with each other, the biased sequential modeling might not work well in terms of learning the inter-point relationships. Hence, to improve the inference speed while enforcing more reasonable structured output patterns into the decoding process, we propose a non-autoregressive decoding mechanism.

3.1.2 Scaled Dot-Product Attention. Attention mechanism is a function that maps a query and a set of key-value pairs to an output vector. Specifically, the output is obtained using weighted average of the input values without regard to their distance. The attention weights measure the compatibility of given query with the input keys. Formally, we first assume the input is composed of queries and keys with dimension d_k and values with dimension d_v, then we can compute the scaled dot-product attention [50] as follows:

$$Attention\left(Q,K,V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where the set of queries, keys, and values are packed into matrices Q, K, and V, respectively. To further model the input information from different subspaces, the attention function can be equipped with multi-head [50]:

$$\begin{cases} MultiHead\left(Q,K,V\right) = Concat\left(head_1,...,head_h\right)W^O, \\ head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right); \end{cases} \tag{3}$$

where linear transformations $W_i^Q \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^K \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^V \in \mathbb{R}^{d_{model}\times d_v}$, $W^O \in \mathbb{R}^{hd_v\times d_{model}}$ are parameter matrices. h is the number of subspaces and $d_k = d_v = d_{model}/h = 16$ in our implementation. We depend on the self-attention as well as encoder-decoder attention mechanisms in our work.

3.2 2D Pose Initialization

We adopt ResNet [22] as backbone network to encode input HOI frame into a 1D context feature vector. We further pass the 1D feature vector through a fully-connected (FC) layer for regressing the 2D hand-object pose. We then concatenate the image context features with each 2D keypoint location, yielding point-wise features. In this manner, we encode both posture and image features into the point-wise representations, which are then fed into our non-autoregressive Transformer module for further processing. We adopt the mean squared error between the estimated 2D keypoint coordinates and the ground truth as loss, which is denoted as $\mathcal{L}_1$.

3.3 Non-Autoregressive Transformer

As the core module of our HOT-Net, it fully models the structural correlations among hand joints and object corners in order to generate physically valid 3D hand-object pose configuration space. Representative Non-AutoRegressive Transformer (NART) models [17, 18, 44, 54], instead of using previously generated tokens as decoder input, explicitly feed a modified copy of input tokens to decoder. For example, in [17], it feeds a copy of source sentence guided by fertilities indicating times each input token is copied. This parallel decoding strategy can achieve drastic inference speedup but come at the cost of inferior performance compared with ART models due to the lack of information from output sequence. The decoding process can be given as:

$$p_{NART}\left(Y|X;\theta\right) = \prod_{t=1}^{T} p\left(y_t|x'_{1:N}, x_{1:N};\theta\right), \tag{4}$$

where x' is a simple copy of x with minor modification, as shown in [17, 44, 54]. This motivates us to propose a decoding mechanism that can run in parallel while being able to exploit the structured

output patterns. Thus, we design a novel non-autoregressive structured learning framework for 3D hand-object pose estimation.

Taking the point-wise representations output from the 2D pose initialization module as input to our NART module, we pass it through our NART encoder and a proposed structured-reference extractor as shown in Fig. 3. The encoder captures the inherent dependencies among the 2D keypoints along with the image features to enhance each point representation prepared for further decoding. The structured-reference extractor, on the other hand, outputs a reference 3D hand-object pose supervised by the ground truth 3D hand and object poses. The goal is to utilize the reference pose as our NART decoder input and provide necessary 3D pose-related structural information.

Specifically, feeding the reference pose to the decoder, we first utilize a non-causal self-attention layer [17] to capture its inherent dependencies. Then, using the captured dependencies as queries, we adopt the encoder-decoder attention mechanism to attend over the enhanced point-wise features output from our NART encoder. This further helps our model find out informative features towards each 3D keypoint localization. Finally, the attention-weighted information from encoder output and the equivalent 3D target pose patterns are combined together to estimation each 3D hand joint or object corner location. The decoding process can be formulated as:

$$p\left(Y|X;\theta\right) = \prod_{m=1}^{M} p\left(y_m|{y'}_{1:M}, x_{1:M};\theta\right), \tag{5}$$

where M is the total number of hand joints and object corners, ${y'}_{1:M}$ denotes the reference pose, $x_{1:M}$ is the extracted 2D pose with image features, and $y_{1:M}$ is the output 3D keypoints. In this manner, our model can generate all 3D points parallelly using the reference pose as decoder input. We apply smooth L1 loss [15] for each frame between our final 3D hand-object pose estimation and the ground truth 3D pose. We denote it as $\mathcal{L}_2$.

3.3.1 Encoder. Similar to the encoder in ART model, our NART encoder also captures the long-range dependencies from the input data, which is the point-wise representations output from the 2D pose initialization module in our case. The encoder is composed of a simple spectral Graph Convolutional Network (GCN) [3, 8, 28] and multi-head self-attention layers. Since each point's features contain 2D keypoint coordinates and image context features, their inherent dependencies can serve as important hints for 3D pose estimation. We first adopt a three-layer spectral GCN similar to [5, 9, 28] for modeling the local dependencies among each 2D keypoint and its connected neighbors via aggregating the adjacent features. Then, to further model the long-range dependencies among hand joints and object corners without regard to their distance or connection, we employ self-attention layers. The encoder outputs enhanced point-wise features with more structural information embedded.

3.3.2 Structured-Reference Extractor. As mentioned earlier, our NART module relies on the structured-reference extractor to provide a reference 3D hand-object pose to the decoder. In this manner, the decoder can use the reference pose patterns, which serves as an equivalent 3D target pose patters, to help constrain the output pose configuration space and generate each 3D keypoint in parallel.

Specifically, as shown in Fig. 2 and Fig. 3, we feed the point-wise features output from the 2D pose initialization module into the structured-reference extractor to infer the reference 3D hand-object pose. We again utilize a multi-layer spectral GCN [5, 9, 28] with encoder-decoder framework as the structured-reference extractor. The GCN allows each node to enhance its representation based on its correlations with adjacent nodes, which helps encode pose structural information into the resulting reference pose.

We apply an intermediate supervision to encourage the reference pose to include more information regarding the ground truth 3D hand-object pose. The loss term $\mathcal{L}_3$ is defined as the smooth L1 loss between the reference pose and the 3D ground truth pose.

3.4 Structural Constraints for 3D Hand Pose

As the third module for our HOT-Net, we expect it to further improve the geometric validity of resulting 3D hand poses in the pose configuration space learnt from our NART module. Due to the highly articulated structure and severe occlusion under HOI scenarios, deep learning-based 3D hand pose estimation is an ill-posed problem. Thus, sometimes we can still observe some joints are put into positions where there is no evidence of presence of hand points, or being physically infeasible. Thus, besides purely relying on our NART module, we include additional constraints to help improve the 3D hand pose estimation accuracy and the kinematical feasibility of the hand skeleton.

The first type of constraint, which helps maintain the skeletal relation between the resulting hand joints, constrain two bone-related properties given a 3D hand pose. The second type of constraint requires hand pose itself to cooperate with hand shape, in the goal to reconstruct the hand mesh. Previous researches usually ignore the correlations between articulated pose and other visual factors for pose estimation. However, we can see that some factors need to help with each other to form another visual modality. For instance, the hand pose along with the hand shape feature should generate a complete 3D hand mesh. Besides relying on ground truth 3D coordinates as supervision, it further imposes another form of constraints on the resulting articulated pose space by asking hand pose to form this natural relation with hand shape.

3.4.1 Bone-Related Biological Constraints. The bone-related biological constraints explicitly maintain a geometrical relation between different joint locations. The classic 3D hand pose loss only constrains the joint locations while ignoring the structural relations between adjacent joints. Specifically, following [23], we apply one bone unit-direction loss to penalize the deviation in the direction of bones and also one bone length loss to restrict the distance in bone length. Both of which are given below:

$$\begin{cases} \mathcal{L}_{bl} = \frac{1}{(J-1)} \sum\limits_{i,j} smooth_{L_1}\left(\|b_{i,j}\|_2, \|\hat{b}_{i,j}\|_2\right), \\ \mathcal{L}_{bd} = \frac{1}{(J-1)} \sum\limits_{i,j} smooth_{L_1}\left(b_{i,j}/\|b_{i,j}\|_2, \hat{b}_{i,j}/\|\hat{b}_{i,j}\|_2\right); \end{cases} \tag{6}$$

where J is the number of hand joints, $b_{i,j} = y_i - y_j$ is the bone vector between hand joint y_i and y_j, and $\hat{b}_{i,j}$ is the ground truth bone vector. The total loss is given as:

$$\mathcal{L}_4 = \mathcal{L}_{bl} + \mathcal{L}_{bd}. \tag{7}$$

3.4.2 Cooperative Pose Constraints via Mesh Reconstruction. To generate hand mesh based on the cooperation between hand pose and shape information, we adopt the MANO model [40], a parametric deformable 3D hand model. The hand mesh are deformed and posed by the input shape $\beta \in \mathbb{R}^{10}$ and pose $\theta \in \mathbb{R}^{Jx3}$ parameters.

Specifically, to obtain the shape parameters β as well as the pose parameters, we pass the image context features generated from ResNet and the 3D hand pose output from NART module through three FC layers, respectively. In terms of the supervision, we apply a root-relative 3D hand pose smooth L1 loss $\mathcal{L}_J$, one shape regularizer $\mathcal{L}_\beta$ to enforce the hand shape to be close to the average shape in the MANO model, which is $\beta = 0 \in \mathbb{R}^{10}$, and also one pose reularizer $\mathcal{L}_\theta$. Both regurlarization losses adopt the mean squared error. The total loss is given below:

$$\mathcal{L}_5 = \mathcal{L}_J + \mathcal{L}_\theta + 2\mathcal{L}_\beta. \tag{8}$$

3.5 Training

We first pre-train the whole model on the ObMan [21] synthetic dataset and then fine-tune on both FP-HO [10] and HO-3D [20] datasets. The total loss for training is given below:

$$\mathcal{L} = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3 + \lambda_4\mathcal{L}_4 + \lambda_5\mathcal{L}_5 \tag{9}$$

where $\lambda_1 = 0.1$, $\lambda_2 = 1$, $\lambda_3 = 0.1$, $\lambda_4 = 0.1$, and $\lambda_5 = 0.5$ are the weight coefficients to balance different loss functions and are set by cross validation.

4 EXPERIMENTS

4.1 Datasets

4.1.1 First-Person Hand Action Benchmark (FPHAB) [10]. It is a recently published large-scale video collection covering a variety of hand-object interactions including 1175 videos with 45 types of activities and 6 subjects in egocentric viewpoint. Visible magnetic sensors are strapped on the human hands in order to automatically annotate the 3D hand joints. A subset of this dataset contains 3D mesh and 6D object pose annotations for 4 objects (juice bottle, liquid soap, milk, and salt). There are 10 different action categories involved in this subset, and we denote the subset as FP-HO.

4.1.2 HO-3D [20]. This dataset is also a recently published hand-object interaction dataset, which contains sequences with hands interacting with objects in third-person viewpoint. It is the first markerless hand-object dataset of color images with 77k annotated frames, corresponding depth maps, 65 sequences, 10 persons, and 10 objects. For the training set, it includes 66k frames with 3D pose annotations for both hands and objects. In the testing set with 11k frames, hands are only annotated with 3D location of the wrist.

4.1.3 ObMan [21]. This is a large-scale synthetic hand-object interaction dataset with 141k training frames, 6.4k validation frames, and 6.2k testing frames. Each image is generated via rendering given 3D hand meshs and 8 everyday object models from ShapeNet [6].

4.2 Evaluation Metrics

Following [45], we adopt mean 3D Euclidean distance error (in mm) for evaluating both 3D hand pose and 6D object pose. We also utilize the percentage of correct keypoint estimates (3D PCK) to measure accuracy on 3D hand pose estimation. For the accuracy on 6D object pose estimation, we employ the percentage of correct poses (PCP). Here, an object pose is correct if the 2D projection error of model vertices is less than a certain threshold.

Table 1: Comparison with state-of-the-art method H+O [45] on FP-HO [10]. The mean 3D distances (mm) is used as metric (Lower is better).

Model	Abs. HP	Abs. OP
H+O [45]	15.81	24.89
HOT-Net	**15.18**	**21.37**

Table 2: The AUC scores (Higher is better) for both FP-HO [10] and HO-3D [20] on 3D PCK curve and PCP curve.

Dataset	AUC on PCK	AUC on PCP
FP-HO	0.829	0.595
HO-3D	0.819	0.567

4.3 Implementation Details

Given HOI RGB frame, We adopt ResNet-50 [22] as the backbone network. The NART encoder is composed of 3 spectral GCN [5, 9, 28] layers following a standard Transformer encoder [50] with layer number L_1 as 6. For the structured-reference extractor, we still use spectral GCN with encoder-decoder structure, similar to [5, 9]. For NART decoder, we rely on the standard Transformer decoder with some modifications: the Masked Multi-Head Attention layer is replaced with a Non-Causal Multi-Head Attention layer [17] and the final SoftMax layer is changed to 1 FC layer. Layer number L_2 is set as 6. The header h is 8 and the d_{model} is 128. We do not use Position Encoding module for both encoder and decoder. For the MANO layer, we mainly follow the setting in [21]. We use Adam [27] as optimizer with batch size of 256. Instead of training the whole model together, we first pre-train the ResNet on 2D pose estimation and the structured-reference extractor. The learning rate is 0.001 for both modules with a shrink factor of 0.5 every 200 epochs for ResNet and 0.9 every 500 epochs for the latter. Then, we optimize the whole network with NART module and MANO layer for 1000 epochs. The learning rate starts from 0.001 with a shrink factor of 0.1 every 500 epochs. All experiments were conducted on single NVIDIA TITAN Xp GPU using PyTorch framework. We use official train/test splits.

4.4 Comparisons with the State-of-the-Arts

In this section, we report our model's performance and compare with the state-of-the-art methods on FP-HO. We do not compare with [9] since it does not provide mean 3D distance for both poses separately and its PCP curve is for 2D object pose. Note that in this section and Section 4.5, we use Abs. to denote 3D pose in camera coordinate system (c.s.) while Rel. represents root-relative 3D pose. HP, OP, MP denote 3D hand pose, object pose, and MANO pose.

As shown in Table 1, in terms of the hand and object mean 3D distance error, HOT-Net is superior to the state-of-the-art model H+O [45], especially for the object pose. For the 3D PCK metric of

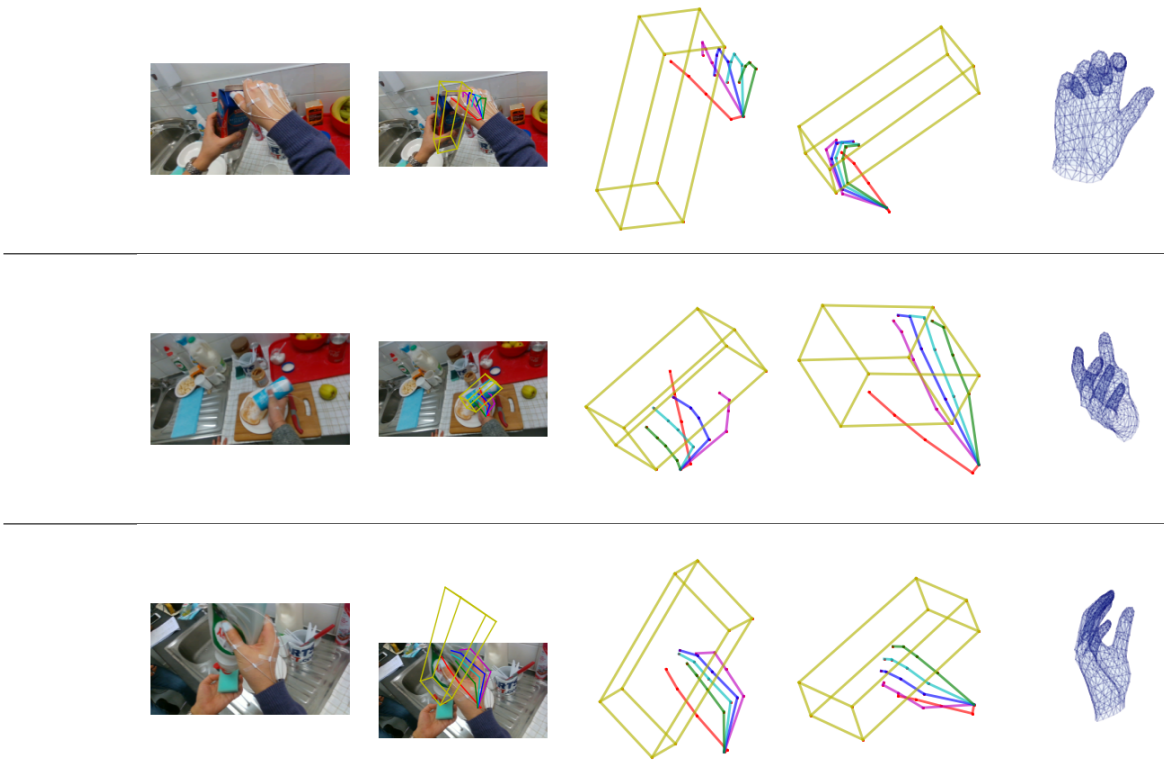

Figure 4: Qualitative results for FP-HO [10]. Our model can handle cases with different objects and severe mutual occlusions.

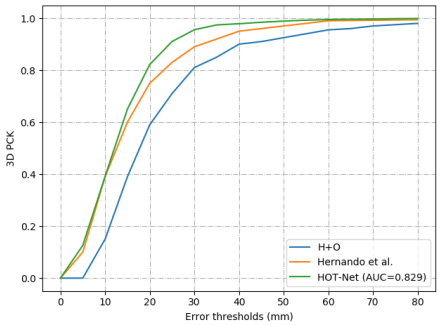

Figure 5: Comparison of hand pose estimation with H+O [45] and Hernando et al. [10] on 3D PCK curve.

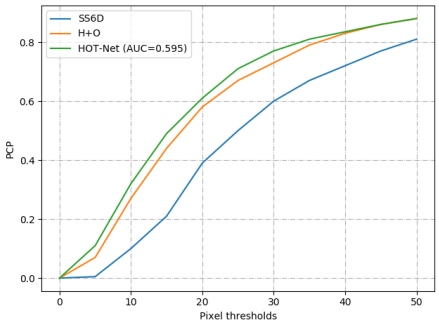

Figure 6: Comparison of object pose estmation with SS6D [46] and H+O [45] on PCP curve.

hand pose, we compare with H+O as well as another depth-based methods [10]. As can be seen in Fig. 5, our method outperforms H+O by a certain margin. and also achieves better performance than the depth-based model especially in the range of error thresholds from 15 mm to 60 mm. More importantly, our method directly operate on input full image without the need of hand bounding box, in contrast to [10]. In terms of PCP curve in Fig. 6 for object pose, we compare with H+O and another object pose estimation technique, SS6D [46]. While our method is slightly better than H+O, the object-only estimation approach [46] is inferior to both H+O and our HOT-Net. This indicates the effectiveness to model correlations between hand and object for both 3D poses estimation.

We also report our Area Under Curve (AUC) scores for both FP-HO and HO-3D on the PCK curve for hand pose and the PCP curve for object pose. Note that the AUC on PCK for HO-3D are measured in terms of the hand wrist since we do not have the complete hand pose annotations for the testing dataset. The PCP for HO-3D is based on 2D projection of object corners. Since HO-3D is a less

Table 3: Ablation study for the effectiveness of different structured modeling techniques on FP-HO [10].

Framework	Abs. HP	Rel. HP	Abs. OP	Rel. MP
H+O[45]	15.81	-	24.89	-
GCN	18.20	12.45	25.13	23.52
ART	20.75	15.26	29.48	27.16
HOT-Net	**15.18**	**10.41**	**21.37**	**21.94**

Table 4: Ablation study for the effectiveness of modeling correlations between hand and object on FP-HO [10].

Framework	Abs. HP	Rel. HP	Abs. OP	Rel. MP
H+O[45]	15.81	-	24.89	-
Hand-Only	16.84	12.18	-	23.12
Object-Only	-	-	27.59	-
HOT-Net	**15.18**	**10.41**	**21.37**	**21.94**

Table 5: Ablation study for the effectiveness of different structural constraints for hand pose on FP-HO [10].

Framework	Abs. HP	Rel. HP	Abs. OP	Rel. MP
H+O[45]	15.81	-	24.89	-
wo Mesh_Rec	15.28	10.76	21.39	-
wo Bone_Len	15.31	10.82	21.51	22.20
wo Bone_Dir	15.30	10.92	21.66	21.95
wo Bone_Loss	**15.16**	11.05	21.43	22.37
HOT-Net	15.18	**10.41**	**21.37**	**21.94**

constrained dataset with more object categories, it is more difficult to get better results. Some qualitative results for FP-HO are given in Fig. 4, which shows our model's ability to handle different poses, object categories, and severe occlusions.

4.5 Ablation Study

We use FP-HO to analyze the impacts of several components of HOT-Net and evaluate the results using the mean 3D distance error (in mm) metric (Lower is better). We also keep the results of H+O [45] in each table for analysis.

4.5.1 Effectiveness of Non-Autoregressive Structured Decoding. To show the effectiveness of proposed non-autoregressive decoding mechanism, we compare with two other learning frameworks. One is GCN-based model and the other one is autoregressive framework. For the comparison with GCN, we simply replace the NART module with another spectral GCN module [5, 9, 28] for direct 3D pose estimation. Similarly, for the autoregressive framework, we replace the core NART module with a typical ART model [50]. The bone-related constraints and mesh reconstruction are kept for both models. The results are given in Table 3. We can observe that both modules are inferior to HOT-Net and do not work well on both hand and object poses estimation. Spectral GCNs mainly rely on the edge connections to model correlations among adjacent nodes, this mechanism is not friendly towards modeling long-range dependencies. Thus, it might not be able to fully understand the correlations among hand joints and object corners. Attention can be a solution since it captures input dependencies without regard to their distance. As for the ART, it helps verify our point that, the sequential modeling is not suited towards the joint pose estimation problem. More importantly, due to the sequential nature of ART model, it runs much slower compared with the other methods.

4.5.2 Effectiveness of Modeling Correlations among Hand Joints and Object Corners. We perform this set of ablation study to understand the impact of modeling dependencies between hand joints and object corners. Hand-only and Object-only mean that we only model 3D hand pose or object pose using HOT-Net. According to the results on Table 4, HOT-Net and H+O both outperform the hand-only or object-only methods. This demonstrates the significance to use the help from object for hand pose estimation, and vice versa. It is worthy noting that the help from hand for object pose estimation leads to a large improvements based on our object pose results.

4.5.3 Effectiveness of structural Constraints on Hand Pose. To examine the impact of imposing structural constraints for hand pose, including bone length loss, bone direction loss, and mesh reconstruction loss, we conduct 4 sets of experiments as shown in Table 5. The experiments from top to bottom correspond to HOT-Net without mesh reconstruction, without bone length loss, without bone direction loss, and without both bone-related losses. The results in Table 5 suggest that bone-related constraints and cooperative constraints can improve the root-relative hand poses but do not have much effect on the hand poses in camera c.s.. It is also worthy noting that when we remove both bone losses, the relative hand pose becomes worse while the absolute hand pose actually is slightly better than our final results reported in the last row. It should not be hard to understand since the extra constraints mainly focus on the relative hand pose itself without paying much attention to the root location and it is also equivalent to asking learning model to put more efforts in learning the relative representation, which might lead to worse root location estimation.

5 CONCLUSION

In this paper, we propose to relate the joint 3D hand-object pose estimation problem with structured output learning mechanism commonly used in the sequence transduction tasks from NLP field. Besides modeling dependencies from extracted 2D hand-object pose, the proposed NART framework imposes structured output pose patterns into the decoding process to help each 3D keypoint localization. Moreover, we further optimize the articulated hand structure via a common physical constraint and a novel cooperative constraint. The latter encourages the hand pose to cooperate with hand shape information in order to generate full hand mesh. Our method outperforms the state-of-the-art methods and runs in real-time speed. The proposed attention-based non-autoregressive structured learning framework and the cooperative constraint can be further extended to other structured output prediction tasks, including human pose estimation, multi-hands interaction scenarios.

ACKNOWLEDGMENTS

This work is supported in part by a gift grant from Kwai and start-up funds from UB.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. 2012. Motion capture of hands in action using discriminative salient points. In *Proceedings of the European Conference on Computer Vision*.

[3] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203* (2013).

[4] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. 2018. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision*.

[5] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE international conference on computer vision*.

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

[7] Chiho Choi, Sang Ho Yoon, Chin-Ning Chen, and Karthik Ramani. 2017. Robust hand pose estimation during the interaction with an unknown object. In *Proceedings of the IEEE International Conference on Computer Vision*.

[8] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*. 3844–3852.

[9] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David Crandall. 2020. HOPE-Net: A Graph-based Model for Hand-Object Pose Estimation. *arXiv preprint arXiv:2004.00060* (2020).

[10] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. 2018. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[11] Liuhao Ge, Yujun Cai, Junwu Weng, and Junsong Yuan. 2018. Hand PointNet: 3D Hand Pose Estimation using Point Sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[12] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 2017. 3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[13] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 2019. 3D Hand Shape and Pose Estimation from a Single RGB Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[14] Liuhao Ge, Zhou Ren, and Junsong Yuan. 2018. Point-to-Point Regression PointNet for 3D Hand Pose Estimation. In *Proceedings of the European Conference on Computer Vision*.

[15] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*.

[16] Duncan Goudie and Aphrodite Galata. 2017. 3D hand-object pose estimation from depth with convolutional neural networks. In *IEEE International Conference on Automatic Face & Gesture Recognition*.

[17] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281* (2017).

[18] Junliang Guo, Xu Tan, Di He, Tao Qin, Linli Xu, and Tie-Yan Liu. 2019. Non-autoregressive neural machine translation with enhanced decoder input. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[19] Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. 2009. Tracking a hand manipulating an object. In *Proceedings of the IEEE International Conference on Computer Vision*.

[20] Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. 2019. Honnotate: A method for 3d annotation of hand and objects poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 3. 6.

[21] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. 2019. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[23] Yiming He, Wei Hu, Siyuan Yang, Xiaochao Qu, Pengfei Wan, and Zongming Guo. 2019. 3D Hand Pose Estimation in the Wild via Graph Refinement under Adversarial Learning. *arXiv*.

[24] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. 2018. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision*.

[25] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. 2017. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. In *Proceedings of the IEEE international conference on computer vision*.

[26] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*.

[27] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[28] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[29] Jameel Malik, Ahmed Elhayek, and Didier Stricker. 2018. Structure-aware 3D hand pose regression from a single depth image. In *International Conference on Virtual Reality and Augmented Reality*. Springer, 3–17.

[30] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. Ganerated Hands for Real-Time 3D Hand Tracking from Monocular RGB. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[31] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2017. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

[32] Markus Oberweger and Vincent Lepetit. 2017. DeepPrior++: Improving Fast and Accurate 3D Hand Pose Estimation. In *Proceedings of the IEEE international conference on computer vision Workshops*.

[33] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. 2015. Training a feedback loop for hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[34] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. 2019. Generalized feedback loop for joint hand-object pose estimation. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[35] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. 2011. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *Proceedings of the IEEE international conference on computer vision*. IEEE.

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

[37] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. 2019. PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[38] Mahdi Rad and Vincent Lepetit. 2017. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects Without Using Depth. In *Proceedings of the IEEE international conference on computer vision*.

[39] Javier Romero, Hedvig Kjellström, and Danica Kragic. 2010. Hands in action: real-time 3D reconstruction of hands in interaction with objects. In *IEEE International Conference on Robotics and Automation*.

[40] Javier Romero, Dimitrios Tzionas, and Michael J Black. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics* 36, 6 (2017), 245.

[41] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[42] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. 2018. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[43] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. 2016. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *Proceedings of the European Conference on Computer Vision*.

[44] Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, and Zhihong Deng. 2019. Fast Structured Decoding for Sequence Models. In *Advances in Neural Information Processing Systems*. 3011–3020.

[45] Bugra Tekin, Federica Bogo, and Marc Pollefeys. 2019. H+ O: Unified egocentric recognition of 3D hand-object poses and interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[46] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. 2018. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 292–301.

[47] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. 2014. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics* 33, 5 (2014), 169.

[48] Aggeliki Tsoli and Antonis A Argyros. 2018. Joint 3D tracking of a deformable object in interaction with a hand. In *Proceedings of the European Conference on Computer Vision*.

[49] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. 2016. Capturing hands in action using discriminative salient

points and physics simulation. *International Journal of Computer Vision* 118, 2 (2016), 172–193.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

[51] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. 2018. Dense 3d regression for hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[52] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 2019. 6-PACK: Category-level 6D Pose Tracker with Anchor-Based Keypoints. *arXiv preprint arXiv:1910.10750* (2019).

[53] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. 2019. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[54] Yiren Wang, Fei Tian, Di He, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2019. Non-autoregressive machine translation with auxiliary regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[55] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. 2017. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199* (2017).

[56] Linlin Yang and Angela Yao. 2019. Disentangling Latent Hands for Image Synthesis and Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[57] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. 2019. End-to-end hand mesh recovery from a monocular RGB image. In *Proceedings of the IEEE international conference on computer vision*.

[58] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. 2016. Deep kinematic pose regression. In *Proceedings of the European Conference on Computer Vision*.

[59] Xingyi Zhou, Qingfu Wan, Wei Zhang, Xiangyang Xue, and Yichen Wei. 2016. Model-based Deep Hand Pose Estimation. In *arXiv preprint arXiv:1606.06854*.

[60] Christian Zimmermann and Thomas Brox. 2017. Learning to Estimate 3D Hand Pose from Single RGB Images. In *Proceedings of the IEEE international conference on computer vision*.