SEMI-SUPERVISED 3D HAND-OBJECT POSE ESTIMATION
VIA POSE DICTIONARY LEARNING

Zida Cheng, Siheng Chen✉, Ya Zhang✉

Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

3D hand-object pose estimation is an important issue to understand the interaction between human and environment. Current hand-object pose estimation methods require detailed 3D labels, which are expensive and labor-intensive. To tackle the problem of data collection, we propose a semi-supervised 3D hand-object pose estimation method with two key techniques: pose dictionary learning and an object-oriented coordinate system. The proposed pose dictionary learning module can distinguish infeasible poses by reconstruction error, enabling unlabeled data to provide supervision signals. The proposed object-oriented coordinate system can make 3D estimations equivariant to the camera perspective. Experiments are conducted on FPHA and HO-3D datasets. Our method reduces estimation error by 19.5% / 24.9% for hands/objects compared to straightforward use of labeled data on FPHA and outperforms several baseline methods. Extensive experiments also validate the robustness of the proposed method.

Index Terms— Hand-object pose estimation, semi-supervision, pose dictionary learning

1. INTRODUCTION

Estimating the pose of hands and manipulated objects is an important task to understand and recognize the behavior of human beings. The related techniques are highly valuable for many practical applications, e.g., AR/VR games and robotics [1, 2, 3, 4]. Here we specifically consider 3D hand-object pose estimation; that is, given the 2D poses of a pair of hand and object, we aim to estimate the corresponding 3D poses. Recent 3D hand-object pose estimation methods use deep neural networks (CNN [5, 6, 7], GCN [8], transformer [9, 10]) to jointly model hands and objects. However, these methods require detailed 3D labels, demanding expensive sensing equipment and a huge amount of manpower [11, 12, 13]. To tackle the problem of data collection, we propose to perform 3D hand-object pose estimation in a *semi-supervised* manner; that is, we only label a small subset of hand-object poses and use both labeled and unlabeled data to train a pose estimator.

There are two main challenges in semi-supervised 3D hand-object pose estimation task: (1) unlabeled data should be effectively leveraged to improve the estimation; and (2) the estimation result should be equivariant with respect to the camera perspective; that is, for multiple 2D poses obtained through photographing the same 3D pose from different camera perspectives, 3D estimation results should be consistent in terms of geometric shapes and only vary in terms of camera perspectives.

Fig. 1. Hand-Object pose estimation results obtained by the proposed method from the camera view (left) and another view (right). Blue lines are ground truth.

To address the first challenge, we introduce an internal self-supervision task and use labeled data to train a 3D grasping pose dictionary, which can well reconstruct feasible 3D grasping poses; in other words, this trained 3D grasping pose dictionary can use the reconstruction errors to distinguish feasible 3D grasping poses from infeasible ones, acting as a discriminator. In our system, even though we do not know the ground-truth 3D grasping poses for unlabeled data, we can use this 3D grasping pose dictionary to identify bad 3D estimations and supervise our pose estimation module to adjust. To address the second challenge, we propose an object-oriented cylindrical coordinate system, which uses the center of the object as the origin and synchronizes the object's orientation. In this coordinate system, the joints' positions are invariant to the camera perspective and an estimation model can focus on estimating grasping poses, excluding the influence of the camera's perspective. After the estimation, we can transform the result back to the standard camera-based Cartesian coordinate system, which makes the estimation equivariant to camera's perspective.

Integrating above designs, we propose a novel semi-supervised 3D hand-object pose estimation network, which includes two training phases. In the first phase, we use labeled data only to train a *pose dictionary learning* module, whose functionality is to decompose an input 3D grasping pose into a linear combination of 3D grasping pose atoms. The training process is to learn such a 3D grasping pose dictionary through self-reconstruction. Different from vanilla autoencoders [14, 15, 16, 17], we use the 3D grasping pose dictionary and the corresponding linear approximation to regularize the reconstruction process, whose benefits are to model the grasping poses more explicitly and constrain the variance of reconstruction process. In the second phase, we train our pose estimation module with the pose dictionary learning module fixed. The pose estimation module is implemented by a graph U-net [8] model and is trained with both labeled and unlabeled data, where labeled data provides direct supervision and unlabeled data provides supervision signals through the reconstruction error of the pose dictionary learning module. In both phases, the proposed object-oriented cylindrical coordinate is used in the reconstruction process. We conduct experiments on FPHA [11] and HO-3D [12] datasets and the experimental results show that our method improves estimation accuracy significantly and is of high robustness.

The main contributions of our method are as follows.

- We propose a semi-supervised 3D hand-object pose estimation network, which is trained by leveraging both labeled and unlabeled data. To our best knowledge, this is the first work to tackle 3D hand-object pose estimation task in a semi-supervised setting.

- We propose a pose dictionary learning module to perform an auxiliary self-reconstruction task, enabling unlabeled data to provide supervision signals.

- We propose an object-oriented cylindrical coordinate system to represent 3D poses, making 3D estimation results equivariant to the camera perspective.

- We conduct experiments on FPHA [11] and HO-3D [12] datasets. The proposed method reduces the estimation error by 19.5%/24.9% for hands/objects compared to straightforward use of labeled data on FPHA, and outperforms several baseline methods. Extensive experiments show that our method is robust to the pose dictionary size and weight of the reconstruction loss.

2. METHODOLOGY

2.1. Problem Formulation

The goal of 3D hand-object pose estimation is to estimate the 3D coordinates of hand joints and the object bounding box's 8 corners from the corresponding 2D coordinates. Mathematically, let $\mathcal{X} = \{\mathbf{X}_i\}_{i=1}^N$ be the 2D pose input, where $\mathbf{X}_i \in \mathbb{R}^{2 \times (m+8)}$ with m is the number of joints in a hand. Among them, only N_L samples $\mathcal{X}_L = \{\mathbf{X}_i\}_{i=1}^{N_L}$, have 3D annotations; denoted as $\mathcal{Y}_L = \{\mathbf{Y}_i\}_{i=1}^{N_L}$, where $\mathbf{Y}_i \in \mathbb{R}^{3 \times (m+8)}$. Let $\mathcal{P}_L = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N_L}$ be the annotated pairs of 2D-3D poses. The rest $N - N_L$ samples $\mathcal{X}_U = \{\mathbf{X}_i\}_{i=N_L+1}^N$ do not have 3D annotations. The task is to infer 3D poses from 2D poses through the limited annotated pairs $\mathcal{P}_L$ as well as a huge amount of unannotated 2D poses $\mathcal{X}_U$.

Here we tackle this task by proposing a neural-network-based estimation model; that is, given both labeled and unlabeled data for training, we aim to propose a pose estimation network to generate 3D estimations for unlabeled data. Different from many other recent works [5, 6, 7, 8, 9, 18, 19], here we consider a semi-supervised setting; that is, besides limited labeled data, we are allowed to involve a huge amount of unlabeled data to train the network. To handle this new setting, our system includes two learning modules: a pose estimation module, which estimates 3D poses from 2D poses, and a dictionary learning module, which enables the supervision from unlabeled data. The whole training procedure has two phases. In the first phase, we train the pose dictionary learning module based on labeled data. In the second phase, we fix the pose dictionary learning module and train the pose estimation module based on both labeled and unlabeled data. Note that the pose dictionary learning module is an auxiliary module that benefits the training of unlabeled data. During the inference time, we only need the pose estimation module to provide 3D estimations.

2.2. Object-Oriented Cylindrical Coordinate System

Before presenting our estimation model, we first introduce a synchronized coordinate system to represent each 3D grasping pose. The new coordinate system can make the estimation result equivariant to the camera perspective. We firstly consider transforming the camera coordinate xyz to an object-oriented Cartesian coordinate system $x'y'z'$, where the origin is the center of the object bounding box and the axes are parallel to the edges. Then the proposed object-oriented cylindrical coordinate (ρ, ϕ, z') is defined based on $x'y'z'$.

Fig. 2. Transformation from camera coordinate to the proposed object-oriented cylindrical coordinate.

In this coordinate system, ρ is the distance to the z' axis and ϕ is the polar angle with respect to x' axis. The whole transformation is illustrated in Fig. 2.

As the definition intrinsically lays objects at the center, we only consider the m hand joints in the final cylindrical coordinates. Another problem is the numerical representation of ϕ. For example, -0.99π and 0.99π are almost the same angle but their numerical values are quite different. Therefore, we use the sine and cosine value to represent angles. Finally, a hand joint is represented by 4 elements $(\rho, \cos(\phi), \sin(\phi), z')$. For a 3D pose $\mathbf{Y}$ in the camera coordinates, the whole transformation process is denoted as $\mathcal{T}(\cdot)$: $\mathbf{h} = \mathcal{T}(\mathbf{Y}) = vec(\mathcal{F}(\mathbf{Y})) \in \mathbb{R}^{4m}$, where $\mathcal{F}(\cdot)$ denotes the transformation from camera coordinate to the object-oriented cylindrical coordinate, and $vec(\cdot)$ means the flatten operation to form a column vector. This proposed object-oriented cylindrical coordinate is used for the input and output of pose dictionary learning module in all two training phases.

2.3. Phase I: Train Pose Dictionary Learning Module

We now present a pose dictionary learning module, which will be used to enable unlabeled data to provide supervision signals. The functionality of a pose dictionary learning module is to reconstruct a 3D grasping pose through a trainable pose dictionary. When this pose dictionary is well trained, the reconstruction error of an input pose can reflect the realistic level of the 3D grasping pose. Therefore, given unlabeled data, this pose dictionary can find unrealistic 3D estimation and provide supervision signals for the pose estimation module.

We train the pose dictionary learning module through self-reconstruction based on labeled data. As shown in Fig. 3, the pose dictionary learning module is comprised of a trainable pose dictionary $\mathbf{D} \in \mathbb{R}^{4m \times k}$ and a pose encoder $Enc(\cdot, \theta_e)$, where k is the number of pose atoms and θ_e is the parameters of pose encoder. Each column vector in the pose dictionary is a trainable atom that represents an elementary grasping pose. The pose encoder is based on a multilayer perceptron model with residual connections. In the labeled data, for a 3D pose $\mathbf{Y}$, we firstly transform it to the proposed object-oriented cylindrical coordinate: $\mathbf{h} = \mathcal{T}(\mathbf{Y})$. Then we input $\mathbf{h}$ to the pose encoder and get the coefficients $\mathbf{c}$: $\mathbf{c} = Enc(\mathbf{h}; \theta_e) \in \mathbb{R}^k$.

Note that the pose encoder $Enc(\cdot, \theta_e)$ includes a softmax operation so that the sum of all elements of $\mathbf{c}$ equals 1. To reconstruct a 3D pose, we can use a linear combination of pose atoms to approximate. Here $\mathbf{c}$ performs as the weight coefficients corresponding to the atoms; that is, $\widetilde{\mathbf{h}} = \mathbf{D}\mathbf{c} \in \mathbb{R}^{4m}$. The reconstruction loss is defined by the mean square error:

$$
\begin{aligned}
\mathcal{L}_{rec}(\mathcal{H}_L) &= \frac{1}{4m \cdot |\mathcal{H}_L|} \sum_{\mathbf{h} \in \mathcal{H}_L} \|\mathbf{D} \cdot Enc(\mathbf{h}; \theta_e) - \mathbf{h}\|^2 \\
&= \frac{1}{4m \cdot |\mathcal{H}_L|} \sum_{\mathbf{h} \in \mathcal{H}_L} \left\| \widetilde{\mathbf{h}} - \mathbf{h} \right\|^2,
\end{aligned}
\tag{1}
$$

where $\mathcal{H}_L = \{\mathcal{T}(\mathbf{Y}) | \mathbf{Y} \in \mathcal{Y}_L\}$.

Fig. 3. Train the pose dictionary learning module based on labeled data through an auxiliary self-reconstruction task.

Fig. 4. Train the pose estimation module with pose dictionary learning module fixed based on all data.

To ease the training of the pose dictionary $\mathbf{D}$, we perform k-means clustering on $\mathcal{H}_L$ and initialize $\mathbf{D}$ with those cluster centers. To ensure each pose atom is feasible, we put an additional loss on $\mathbf{D}$ to constrain its elements in valid intervals. Specifically, elements for sine and cosine values should lie in $[-1, 1]$ and ρ values should be no less than 0. Finally we define the valid dictionary loss $\mathcal{L}_{dict}$ by:

$$\mathcal{L}_{dict} = \frac{2}{3|\mathcal{D}_{sc}|} \sum_{d \in \mathcal{D}_{sc}} \mathcal{I}(d; -1, 1) + \frac{1}{3|\mathcal{D}_{\rho}|} \sum_{d \in \mathcal{D}_{\rho}} \mathcal{I}(d; 0, +\infty), \tag{2}$$

where $\mathcal{D}_{sc}$ is the set of sine and cosine elements in $\mathbf{D}$, $\mathcal{D}_{\rho}$ is the set of ρ elements, and the interval loss function $\mathcal{I}(\cdot)$ is defined as $\mathcal{I}(d; d_{min}, d_{max}) = \max(d_{min} - d, 0) + \max(d - d_{max}, 0)$.

The final loss for training the pose dictionary learning module is: $\mathcal{L}_{pdl} = \mathcal{L}_{rec}(\mathcal{H}_L) + \lambda_{dict}\mathcal{L}_{dict}$, where λ_{dict} is the weight for valid dictionary loss and we empirically set it to a large value to make sure pose atoms are valid.

To sum up, the pose dictionary learning module is comprised of a trainable pose dictionary and a pose encoder. It performs the reconstruction of 3D poses while the reconstruction process is regularized by the pose dictionary. By training the pose dictionary learning module on the limited annotated data, we squeeze information about feasible grasping poses into the pose dictionary. After the training is finished, the reconstruction error of a given grasping pose can reflect whether the input is realistic.

2.4. Phase II: Train Pose Estimation Module

The pose estimation module $\mathcal{E}(\cdot; \theta_p)$ is implemented based on an adaptive graph U-net proposed by [8] with parameters θ_p. It takes a 2D pose $\mathbf{X}$ as input and produces an estimated 3D pose, $\widehat{\mathbf{Y}} = \mathcal{E}(\mathbf{X}; \theta_p)$.

To train this network, we consider supervisions from two aspects; see an overall illustration in Fig. 4. First, based on labeled data, we consider a direct fully-supervised estimation loss (in the camera coordinate xyz):

$$\mathcal{L}_L = \frac{1}{3(m + 8) \cdot N_L} \sum_{(\mathbf{X}, \mathbf{Y}) \in \mathcal{P}_L} \|\mathcal{E}(\mathbf{X}; \theta_p) - \mathbf{Y}\|_F^2, \tag{3}$$

where $|\mathcal{P}_L| = N_L$ is the number of labeled data and $\|\cdot\|_F$ means Frobenius norm.

Second, based on both labeled and unlabeled data, we consider a self-supervision loss. We input all data to the pose estimation

module and obtain 3D estimations; we next represent those estimations in the proposed object-oriented cylindrical coordinate; finally, we input the transformed estimations to the fixed pose dictionary learning module. Recall the definition of reconstruction loss in the first training phase in Eq. (1). Simply changing $\mathcal{H}_L$ to $\widehat{\mathcal{H}}$, the reconstruction loss in the second training phase is: $\mathcal{L}_{rec}(\widehat{\mathcal{H}})$, where $\widehat{\mathcal{H}} = \{\mathcal{T}(\mathcal{E}(\mathbf{X}; \theta_p)) | \mathbf{X} \in \mathcal{X}\}$.

Here we freeze the parameters of the pose dictionary learning module and only update θ_p. When the reconstruction loss is large, it reflects that the estimation result from the pose estimation module is unrealistic and the well-trained pose dictionary learning module cannot represent such 3D poses. Therefore, minimizing the reconstruction loss pushes the the pose estimation network to produce realistic estimations. The reconstruction loss enables non-annotated data $\mathcal{X}_U$ to provide supervision signals.

The overall loss to train the pose estimation module is:

$$\mathcal{L} = \mathcal{L}_L + \lambda_r \mathcal{L}_{rec}(\widehat{\mathcal{H}}), \tag{4}$$

where λ_r balances the two terms.

3. EXPERIMENTS

3.1. Datasets and Implementation Details

First-Person Hand Action Dataset(FPHA) [11] contains first-person videos of hands manipulating various objects. A subset of it (21501 frames) include 3D annotations of 21 hand joints and 8 object bounding box corners. We remove the frames where the hand is not in contact with the object. Finally we get 10388 frames for training and 9761 for evaluation. **HO-3D** [12] dataset contains videos of hand-object interaction from third-person view. Here we use the subset for seen-object setting and remove frames without contact. As the frame rate is high, we take images every other frame in the training set. We finally use 9467 frames for training and 6120 for evaluation. For hand annotations, the evaluation set only provides the wrist joints so we only evaluate on wrist joints. HO-3D dataset is of low diversity, and some *different* video sequences are actually the *same* period of action captured by multiple cameras from different perspectives. Thus, the results of HO-3D is less convincing so we mainly demonstrate the results of FPHA.

For both datasets, we sample 5% of the frames as the annotated samples $\mathcal{X}_L$. As our method is based on single image, temporal information of video data can affect our evaluation of the model's performance. To reduce the affection, we divide the videos into subsequences of 5 frames each and randomly sample 5% of the subsequences as $\mathcal{X}_L$. The pose encoder is an 8-layer MLP model with (1024,256,256,1024,256,256,1024,k) hidden units in each layer. Two shortcut connections are added between output of (1st,4th) and (4th,7th) layers, similar to [20]. ReLU activation and batch normalization are used after every layer except the last. We set λ_{dict} and λ_r to 100 and the size k of pose dictionary is 30 by default. As semi-supervised task is sensitive to the amount of data, we do not perform pretraining on large scale synthetic dataset [6], which is different from [8, 9].

3.2. Results

Primary results: Table 1 reports the results on FPHA and HO-3D where Procrustes Aligned Mean Per Joint Position Error (MPJPE) is used as the metric. Fig. 5 reports the Percentage of Correct Keypoints (PCK) on FPHA. We compare with following methods: (1)

Method	FPHA		HO-3D	
	Hand	Object	Hand	Object
Fully supervision	8.77	14.31	54.94	32.17
5% supervision	14.00	25.19	59.09	36.12
BMC [21]	11.67	21.99	56.24	**33.24**
AE reconstructor	11.92	21.53	58.36	34.80
Ours	**11.27**	**18.91**	**56.09**	33.58

Table 1. MPJPE(mm) on FPHA and HO-3D. Best results in bold.

Fig. 5. PCK on FPHA

Method	Hand	Object
Ours	**11.27**	**18.91**
L2	11.35	20.33

Table 2. Impact of $\mathcal{L}_{dict}$

Fig. 6. Visualized examples of the learned pose dictionary. Left: our method; right: replacing $\mathcal{L}_{dict}$ with $L2$ regularization. Our method learns a more feasible pose dictionary.

(a) Impact of k (b) Impact of λ_r

Fig. 7. Impact of k and λ_r. Our method shows robustness to choice of the two hyperparameters.

training the pose estimation module with all samples labeled (fully supervision), which is the error lower bound; (2) training the pose estimation module only by the 5% labeled data straightforwardly (5% supervision), which is the error upper bound; (3) Biomechanical Constraints (BMC) [21] method transplanted to our task; (4) replacing the pose dictionary learning module with a vanilla autoencoder (AE reconstructor), which also performs the reconstruction task. Compared to straightforward use of the limited labeled data, our method significantly reduces the error. The proposed method outperforms BMC on FPHA and acquires comparative results on HO-3D. (HO-3D is less suitable for this task and its results is less convincing.) Our method surpasses AE reconstructor, validating the effectiveness of pose dictionary learning. As the annotated data $\mathcal{X}_L$ is selected from video sequences, we study the affection of temporal information. We create pseudo-labels for $\mathcal{X}_U$ by interpolation along time and train the pose estimation module by both original and pseudo-labels. Finally we get dramatically large error (>25mm/50mm) on hands/objects. It suggests that the affection of temporal information is eliminated.

Constraints on Pose Dictionary: Table 2 studies the impact of the constraints on pose dictionary $\mathcal{L}_{dict}$ on FPHA dataset. We replace $\mathcal{L}_{dict}$ with simple $L2$ regularization, which slightly increases the errors. Therefore, we believe that constraining elements of **D** in valid intervals can give feasible pose atoms and model the grasping poses better. In Fig. 6, we visualize several atom vectors from learned pose dictionary **D** with $\mathcal{L}_{dict}$ or $L2$ regularization. The results show that our method learns a more feasible pose dictionary.

Robustness to hyperparameters: We study the robustness of the proposed method with respect to important hyperparameters on FPHA dataset. k, the number of pose atoms in **D**, makes direct and significant impact on the reconstruction process in the pose dictionary learning module. The pose dictionary helps model the grasping poses more explicitly, which can be seen as regularization effect. The regularization is strong when k is small. If k is too large, the pose dictionary learning module gains high variance and may not be able to constrain the estimation result effectively. Fig. 7 (a) reports the impact of k. We set k in the moderate interval [10,60]. Our method has achieved low estimation errors in this interval, showing robustness to the choice of k. Another important hyperparameter is λ_r, which controls the weight of reconstruction loss in the sec-

ond training phase. As shown in Fig. 7 (b), our method performs well when $\lambda_r \in [10, 100]$. The errors increase when the weight is too large. We believe this is because the pose dictionary learning module is not a perfect reconstructor. An overlarge weight for reconstruction loss forces the pose estimation module to approach an imperfect target too much.

Ratio of labeled data: The ratio of labeled data is an important factor and we study its impact on estimation accuracy. Fig. 8 reports the MPJPE of all points. The proposed method (in red) is compared with the straightforward use of labeled data (in blue) on various ratios. The result shows our method can acquire stable accuracy gain.

Fig. 8. MPJPE of all points (hands and objects) as a function of the labeling ratio on FPHA. Lower bound of y axis is the error of fully supervision. Our method (in red) acquires stable accuracy gain over the straightforward use of labeled data (in blue).

4. CONCLUSION

For hand-object pose estimation task, detailed 3D labels are expensive and labor-intensive. To tackle the data collection problem, we propose a semi-supervised method based on pose dictionary learning. The proposed pose dictionary learning module performs an auxiliary reconstruction task. It enables unlabeled data to provide supervision signals for the pose estimation module by the reconstruction error. We propose to use an object-oriented coordinate system to make the estimation equivariant to the camera perspective. We show that the proposed method improves estimation accuracy significantly and is of high robustness.

5. REFERENCES

[1] S. M. Hussain.S, L. Liu, W. Xu, and C. Lu, "Fpha-afford: A domain-specific benchmark dataset for occluded object affordance estimation in human-object-robot interaction," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1416–1420.

[2] Ammar Ahmad, Cyrille Migniot, and Albert Dipanda, "Hand pose estimation and tracking in real and virtual interaction:a review," *Image and Vision Computing*, vol. 89, pp. 35 – 49, 2019.

[3] Min-Yu Wu, Pai-Wen Ting, Ya-Hui Tang, En-Te Chou, and Li-Chen Fu, "Hand pose estimation in object-interaction based on deep learning for virtual reality applications," *Journal of Visual Communication and Image Representation*, vol. 70, pp. 102802, 2020.

[4] B. Kang, K. Tan, N. Jiang, H. Tai, D. Tretter, and T. Nguyen, "Hand segmentation for hand-object interaction from depth map," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2017, pp. 259–263.

[5] Bugra Tekin, Federica Bogo, and Marc Pollefeys, "H+o: Unified egocentric recognition of 3d hand-object poses and interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid, "Learning joint reconstruction of hands and manipulated objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[7] D. Goudie and A. Galata, "3d hand-object pose estimation from depth with convolutional neural networks," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, 2017, pp. 406–413.

[8] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J. Crandall, "Hope-net: A graph-based model for hand-object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[9] Lin Huang, Jianchao Tan, Jingjing Meng, Ji Liu, and Junsong Yuan, "Hot-net: Non-autoregressive transformer for 3d hand-object pose estimation," in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA, 2020, MM '20, p. 3136–3145, Association for Computing Machinery.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] Brahmbhatt Samarth, Tang Chengcheng, D. Twigg Christopher, C. Kemp Charles, and Hays James, "Contactpose: A dataset of grasps with object contact and hand pose," in *European Conference on Computer Vision (ECCV)*, August 2020.

[14] Manassés Ribeiro, André Eugênio Lazzaretti, and Heitor Silvério Lopes, "A study of deep convolutional auto-encoders for anomaly detection in videos," *Pattern Recognition Letters*, vol. 105, pp. 13–22, 2018.

[15] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár, "Lossy image compression with compressive autoencoders," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017, OpenReview.net.

[16] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Artificial Neural Networks and Machine Learning – ICANN 2011*, Timo Honkela, Włodzisław Duch, Mark Girolami, and Samuel Kaski, Eds., Berlin, Heidelberg, 2011, pp. 52–59, Springer Berlin Heidelberg.

[17] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[18] Chiho Choi, Sang Ho Yoon, Chin-Ning Chen, and Karthik Ramani, "Robust hand pose estimation during the interaction with an unknown object," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2017.

[19] M. Oberweger, P. Wohlhart, and V. Lepetit, "Generalized feedback loop for joint hand-object pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 1898–1912, 2020.

[20] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi, "C3dpo: Canonical 3d pose networks for non-rigid structure from motion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[21] Spurr Adrian, Iqbal Umar, Molchanov Pavlo, Hilliges Otmar, and Kautertz Jan, "Weakly supervised 3d hand pose estimation via biomechanical constraints," in *European Conference on Computer Vision (ECCV)*, August 2020.