3D Pose Nowcasting: Forecast the Future to Improve the Present

Alessandro Simoni*, Roberto Vezzani
University of Modena and Reggio Emilia

{alessandro.simoni, roberto.vezzani}@unimore.it

Francesco Marchetti*, Lorenzo Seidenari, Alberto Del Bimbo
University of Florence

{francesco.marchetti, lorenzo.seidenari,alberto.delbimbo}@unifi.it

Federico Becattini
University of Siena

{federico.becattini}@unisi.it

Guido Borghi
University of Bologna

{guido.borghi}@unibo.it

Abstract

Technologies to enable safe and effective collaboration and coexistence between humans and robots have gained significant importance in the last few years. A critical component useful for realizing this collaborative paradigm is the understanding of human and robot 3D poses using non-invasive systems. Therefore, in this paper, we propose a novel vision-based system leveraging depth data to accurately establish the 3D locations of skeleton joints. Specifically, we introduce the concept of Pose Nowcasting, denoting the capability of the proposed system to enhance its current pose estimation accuracy by jointly learning to forecast future poses. The experimental evaluation is conducted on two different datasets, providing accurate and real-time performance and confirming the validity of the proposed method on both the robotic and human scenarios.

1. Introduction

We are increasingly approaching an era in which humans and robots will share different spaces and moments of the day, both in social and working scenarios [43].

Non-invasive camera monitoring combined with specific computer vision algorithms, such as Robot and Human Pose Estimators [32, 64], are key and enabling technologies for safe interaction between humans and robots [10]. For instance, in the Industry 4.0 setting [30], in which the same workplace is shared between workers and cobots [28], the ability to detect poses and avoid collisions is fundamental for safety. Furthermore, recent investigations [56, 57] confirm that – rather than the complete removal of humans – future generations of manufacturing will support the co-
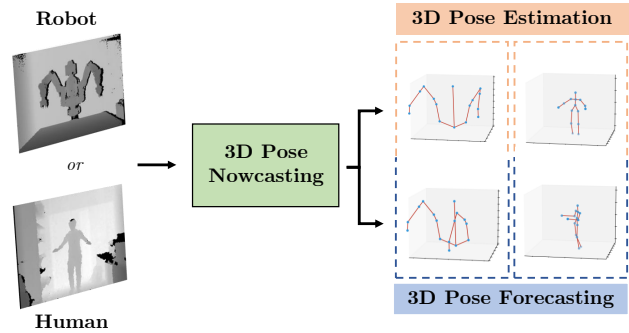

Figure 1. Estimating current and future poses through 3D Pose Nowcasting, using depth images as input data, is a fundamental technology for safe interaction between workers and collaborative machines in indoor scenarios, such as the Industry 4.0 setting.

existence of humans and cobots, stressing the urgency for new investigations related to physical and social coworker coordination [12]. Another possible application setting is represented by home automation, in which robots can autonomously perform actions but also interact with humans.

In both cases, technologies based on non-invasive sensors that are agnostic with respect to the state of the robot's encoders, are highly desirable. A variety of collision detection systems, especially for the industrial environment, has been proposed but, unfortunately, they often require the use of specific sensors [21], markers [26] or access to the robot's proprietary software [17], which is not always possible.

Therefore, in this paper, we propose a vision-based system able to accurately estimate the 3D poses by learning to forecast the near future as an auxiliary task. In particular, we show how the knowledge about the future at training

time improves the model's performance in the present.

Given the similarities with the weather forecasting [6], we refer to this novel paradigm as **3D Pose Nowcasting**, characterized by the following elements: i) the forecasting regards a brief time span (around a few seconds); ii) we are not required to access specific physical models or additional sensors other than the input data (in our case, depth images); iii) forecasting, in addition to enhancing present estimation, is important to raise alarms about imminent and unexpected events (*e.g.* collisions, hazards).

The proposed method for 3D Pose Nowcasting, outlined in Figure 1, is based on addressing the task from two different research fields, *i.e.* 3D Pose Estimation (PE) and 3D Pose Forecasting (PF), jointly learned during training. In particular, the model is trained end-to-end to estimate the 3D pose at the current timestep and the 3D poses at the next future timesteps.

Our approach is based on depth data enabling the development of a vision-based system robust to varying or absent environmental light sources [47], usually common in indoor scenarios such as workplaces. Besides, depth acquisition devices nowadays are inexpensive, yet accurate [61]. Moreover, in the Sim2Real [23] setting, the use of depth reduces the domain gap between synthetic and real scenarios [49], thus enabling the usage of large-scale datasets without the time-consuming collecting and labeling procedures required with real data.

From an architectural point of view, PE and PF are tackled through two double-branch CNNs, each specialized in estimating and forecasting joints in 3D world coordinates. The first branch is composed of a backbone originally developed for Human Pose Estimation [4] (HPE), while the second one is obtained by exploiting a motion encoder based on a recurrent neural network, that processes a sequence of past joint locations. The 3D world-coordinate locations of each joint are given in output in real-time, leveraging the recent Semi-Perspective Decoupled Heatmaps (SPDH) [49] as an intermediate representation of poses. To train the model, a double loss is used to optimize both the current pose and the future poses. This is justified by the fact that we want the forecasting loss to influence and improve the estimate at the current timestep.

Summarizing, the main contributions of our paper are:

- We introduce the novel paradigm of 3D Pose Nowcasting, a combination of 3D Pose Estimation and 3D Pose Forecasting in a joint optimization framework. By learning to predict the future, our model improves its pose estimation accuracy in the present.

- We demonstrate the robustness of our approach in the Sim2Real scenario, enabling effective exploitation of synthetic data at training time, and also domain transfer capabilities from synthetic to real.

- We obtain state-of-the-art performance in estimating the current robot's pose, also providing reliable future predictions. In addition, we show that 3D Pose Nowcasting can be easily exploited for estimating human body joints.

2. Related Work

Robot Pose Estimation from Depth. Only a limited amount of research addresses the task of pose estimation from depth data. Bohg *et al.* [5] proposed to use a random forest classifier to classify and then group depth maps pixels, obtaining skeleton joints. A similar approach is reported in [58], in which joint angles are directly regressed without any segmentation prior. However, these methods are unable to infer real-world 3D poses, limiting their estimates to joint angles. The large majority of literature works for robot pose estimation are developed for the RGB domain. In general, there are two main approaches: hand-eye calibration-based and rendering-based. In the former, methods are based on fiducial markers (*e.g.* ArUco [16]) placed on the robot's end effector, tracked through multiple cameras. Then, a 3D-2D correspondence problem is solved by relying on forward kinematics or the PnP [33] approach. Unfortunately, these methods are invasive since they require the physical application of markers on the robot, which is not always feasible or practicable. Differently, rendering-based methods [29, 40] use the render&compare paradigm, where an optimization algorithm iteratively refines the pose projected to the image with respect to the camera.

Human Pose Estimation from Depth. Shotton et al. [48] introduced a pioneering approach based on a random forest classifier to classify pixels enabling the segmentation of the human body. The 3D joint candidates are then identified through a weighted density estimator. Using similar features, in [60] the authors proposed to use a regression tree to predict the probability distribution of the direction of a specific joint. Entering the deep learning-based field, some works introduce the use of NNs in combination with a single depth frame. In [55], a specific memory module referred to as Convolutional Memory Block is introduced, merging the power of CNNs and a memory mechanism used to handle depth data. More recently, [14] introduced a capsule autoencoder network based on fast Variational Bayes capsule routing, focusing on improving viewpoint generalization both on intensity and depth data. Other works are based on point clouds sampled from depth data. In particular, the method described in [62] is based on a point clouds proposal module followed by a 3D pose regression module. Similarly, the same authors in [63] introduced a sequential pose estimation module based on a window of different frames, improving the general performance at the cost of increasing computational complexity. Finally, some
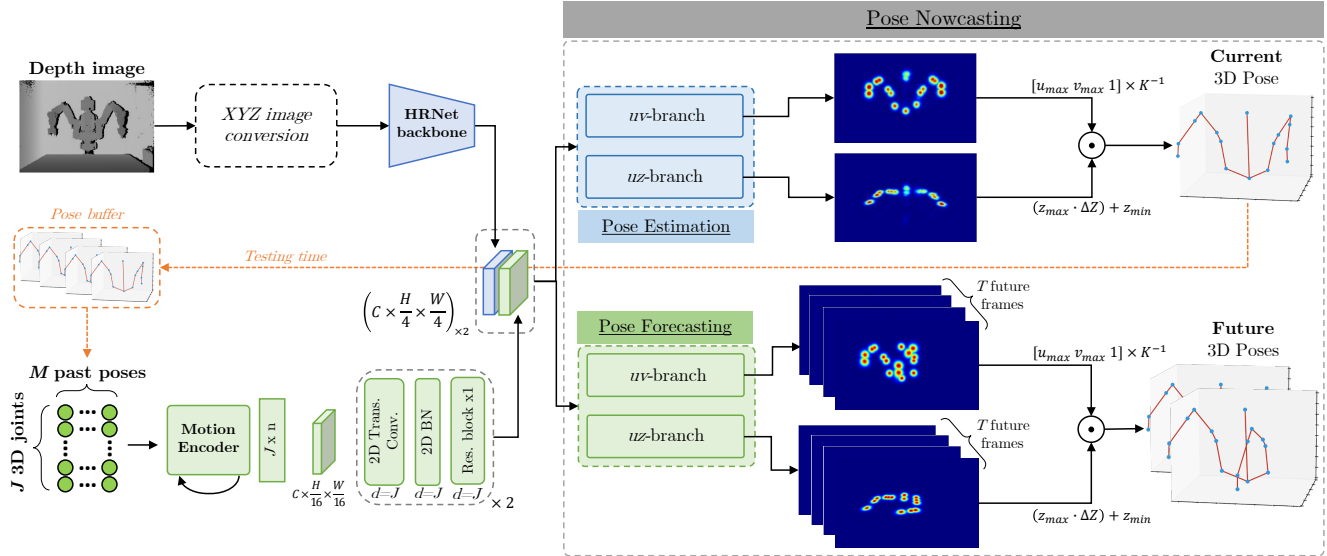
Figure 2. Overview of the proposed 3D Pose Nowcasting framework. First, features related to the depth map and the past poses are extracted. These features are then concatenated and fed to two different branches, *i.e.* the Pose Estimation and Pose Forecasting ones. Finally, the framework outputs the current and the near-future 3D poses. For the sake of visualization, heatmaps are stacked channel-wise.

literature works have been developed originally for the hand pose estimation task [19, 39, 59] and then adapted to tackle also the human pose estimation task.

Pose Forecasting. Recently, Sampieri *et al.* [46] proposed a graph convolutional neural network to jointly model robot arms and human operators from RGB images. Their goal is to anticipate human-robot collisions. In this work, we follow this research direction and we leverage a trajectory forecasting architecture to improve the current 3D robot pose estimate while also providing information about the future locations of robots and humans. From a general point of view, a large crop of literature has addressed motion forecasting tasks, especially in automotive [25, 31, 35, 37] and human behavior understanding [8, 9, 13, 42, 52]. The task can be framed as an encoder-decoder problem, where past motion is projected into a latent state and then decoded into a plausible future [2, 31]. Interestingly, most approaches formulate the forecasting task as a multimodal prediction task, due to the intrinsic uncertainty of the problem [18, 31, 45, 54]. More recently, several works have addressed the task of forecasting human poses. Compared to the automotive setting, this is a much more complex scenario, since body joints can move erratically and the position of the whole skeleton must be predicted at every timestep. Here, graph-based representations play an important role, since body joints can be naturally represented as connected nodes [1, 34, 44, 50]. Unlike these methods, Mangalam *et al.* [36] fused 3D skeletons, camera ego-motion and monocular depth estimates to forecast body poses. In a similar way, we propose a depth-based approach for pose

estimation and forecasting. Differently from [36], we focus on robot poses and, instead of observing a full sequence of depth and joints, we blend the current depth with an encoding of autoregressively generated past joints.

Depth-based datasets for Pose Estimation and Forecasting. We observe a substantial lack of datasets that can be used for robot pose estimation and forecasting starting from depth data. Recently, four different datasets have been introduced in the literature, but totally based on RGB data. Released in 2019, the CRAVES [65] dataset consists of synthetic and real acquisitions of a single type of robotic arm, for a total of about 5k frames. DREAM [32] and WIM [40], introduced in 2020 and 2022, contain 350k and 140k intensity frames, respectively, depicting different types of robots. One of the most recent datasets is referred to as CHICO [46]. Expressively introduced for collision detection in human-robot interaction, it collects more than 1 million frames acquired with multiple RGB cameras [1]. Therefore, the only dataset exploitable to test our method is the recent SimBa [49], consisting of more than 370k frames depicting the Rethink Baxter robot performing pick-and-place operations in random locations. This dataset has been acquired in the Sim2Real [23] scenario, *i.e.* the training and testing frames belong to two different domains: synthetic (generated through ROS and Gazebo [27] simulator) and real (acquired through the time-of-flight Microsoft Kinect v2 depth device). SimBa is suitable for our task due to the presence of video sequences, collected at 30 fps.

---

[1]This dataset presents corrupted 3D joint annotations on images not yet fixed by the authors, making it impossible for us to adopt it.

With regard to the estimation of human poses, we adopt the ITOP dataset [20], which has been used as a benchmark by several prior works [14, 15, 55, 62, 63]. Also in this case, we observe a substantial lack of depth-based datasets in the literature, suitable for our method, for different motivations. Human3.6M [24] dataset contains very low-quality depth images, acquired through the MESA Imaging SR4000 device. The NTU dataset [53], originally developed for the human action recognition task, contains good quality depth data, but unfortunately, the human pose annotations are automatically provided through the method described in [48], reducing their accuracy. The mRI dataset [3] appears to be an interesting dataset but depth data have yet to be released, at the time of writing.

3. Proposed Method

An overview of the proposed framework is depicted in Figure 2. It is organized in an encoder-decoder fashion that is split into two input branches and two output branches. The encoder extracts visual and temporal embeddings, while the decoder consists of the *Pose Nowcasting* block, which is made of two SPDH [49] branches dedicated to pose estimation and pose forecasting.

From a formal point of view, the encoder can be viewed as a single frame 2D depth input branch $\Pi(\cdot)$ and a temporal 3D joint recurrent input branch $\Gamma(\cdot)$. For a depth image D and a sequence of $t = 1, ..., M$ poses $P_j^t = [X_j^t, Y_j^t, Z_j^t]$ with $j = 1, ..., J$ 3D joints, two same-size feature maps $\Pi(D)$ and $\Gamma(\mathbf{P})$ are computed and concatenated. The output branches of the nowcasting decoder then independently generate current and future pose predictions.

3.1. Depth and Past Pose Input Processing

As mentioned, the first input branch is responsible for extracting the features related to the current pose. In this case, the input is represented by a depth image that is converted into an XYZ image, formally defined as follows:

$$I_{XYZ} = \pi(D \cdot K^{-1}) \qquad (1)$$

where π is the projection in the 3D space, D is the matrix of distances used to create the depth image and K is the projection matrix. This kind of depth representation has been proved to have better generalization capabilities across different domains with respect to common depth images [49]. Being aware of the recent and significant advances in HPE [11], we exploit the well-known HRNet-32 architecture [51], specifically the randomly initialized first four stages without the last convolution, as the backbone to extract pose-related features. These features are then concatenated with the ones extracted through the other branch, described as follows.

The second input branch incorporates temporal information obtained from previously estimated 3D joint positions:
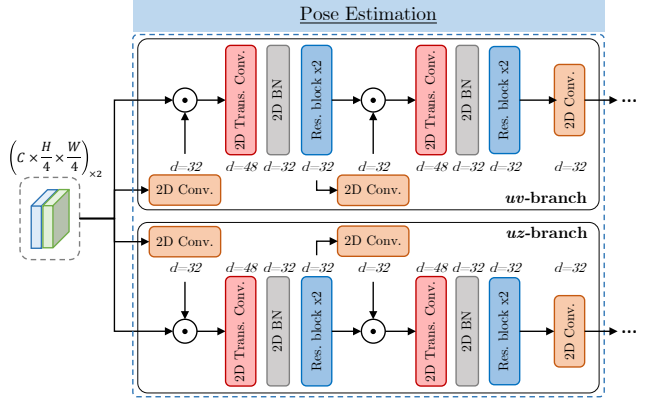

Figure 3. Architecture of the Pose Estimation branch. The input is represented by the concatenation of features extracted from depth maps and past joints. Each *uv/uz* sub-branch generates the heatmap-based SPDH [49] representation of 3D joint locations.

this information becomes available as soon as a buffer of poses of length M is filled by storing the outputs of the pose estimation block. This branch uses a motion encoder, implemented as a GRU[2], to process higher dimensional embeddings of each pose P_j^t. Its output is organized into a $C \times \frac{H}{16} \times \frac{W}{16}$ shaped feature map, which is then processed with two layers of residual transposed convolutions with BatchNorm. This architecture is both responsible for processing temporal information stored in previously estimated joints and for adapting the 3D representation to a 2D map that can be fused with the feature map extracted by $\Pi(\cdot)$ from depth images.

3.2. Pose Estimation and Forecasting Branches

Our framework is completed by the nowcasting block with two output branches jointly solving pose estimation and forecasting. Both branches exploit the same SPDH [49] representation, in which the 3D space is decomposed into two bi-dimensional spaces where skeleton joint locations are expressed through heatmaps. In particular, the uv space corresponds to the camera image plane (the front view of the acquired scene), while the uz space contains the quantized values of the depth dimension, *i.e.* a sort of birds-eye view of the scene with discretized information about the distance of the joints.

In the *pose estimation branch*, the SPDH representation is obtained through the architecture detailed in Figure 3, consisting of two residual transposed convolution layers followed by a BatchNorm and ReLU activation function. The estimated pose is represented by a set of $J \times 2$ heatmaps,

---

[2]Potentially any kind of recurrent architecture such as LSTMs or Transformers could be used. Since our focus is on Nowcasting, we adopt GRUs as commonly done in the trajectory forecasting literature, leaving the investigation of different architectures to future research.

Input	Model	mAP (%) ↑					ADD (cm) ↓
		2cm	4cm	6cm	8cm	10cm	
Depth	ResNet-18 [22]	0.57	9.40	19.99	27.06	44.44	$12.20_{\pm 4.12}$
2D joints	Martinez *et al.* [38] *	13.70	26.96	37.98	48.40	58.33	$10.03_{\pm 3.53}$
Depth	Pavlakos *et al.* [41]	3.35	18.15	42.24	61.60	86.15	$7.11_{\pm 0.65}$
Depth	Simoni *et al.* [49]	6.33	53.75	79.75	93.90	98.12	$4.41_{\pm 1.09}$
Depth	*Ours w/o forecasting*	16.25	57.51	89.81	**99.26**	**99.81**	$3.77_{\pm 0.98}$
Depth + M past poses	*Ours*	**30.68**	**66.90**	**92.69**	98.02	98.38	$\mathbf{3.52}_{\pm 1.30}$

Table 1. Robot pose estimation results on SimBa. The proposed framework is tested by taking as input a single depth image ("*Ours w/o forecasting*") or a depth image with the previously predicted 3D joints ("*Ours*"). Method marked with * uses a relative joint representation.

one pair for each joint in the uv and uz spaces.

In the *pose forecasting branch*, we adopt a lighter architecture to deal with the multiple SPDH representations that aim to model the near-future joint locations. In particular, we use two 2D convolutional layers, with a size of 32, interspersed with a BatchNorm and ReLU activation function. The forecasted poses are represented as $T \times (J \times 2)$ future heatmaps, where T is the forecasting horizon.

For both output branches, final predictions are obtained as follows: we compute the argmax of the uv heatmaps and we multiply the resulting values (u_{max}, v_{max}) with the inverse of the camera intrinsics K^{-1} to obtain the final 3D coordinates. Differently, with uz heatmaps, we transform the result of the argmax operation into a continuous value in the metric space multiplying it with the quantization step (ΔZ) computed in the defined depth range (z_{min}, z_{max}).

3.3. Losses

To train the model, we directly optimize the uv/uz heatmaps, before they are converted into 3D coordinates. The system is trained end-to-end optimizing the Mean Squared Error (MSE) loss function $\mathcal{L}$ between generated and ground truth heatmaps:

$$\mathcal{L}_{PE} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} ||H_j^t - \widehat{H}_j^t||_2 \qquad (2)$$

$$\mathcal{L}_{PF} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \frac{1}{T} \sum_{k=1}^{t+T} ||H_j^{t+k} - \widehat{H}_j^{t+k}||_2 \qquad (3)$$

$$\mathcal{L} = \mathcal{L}_{PE} + \mathcal{L}_{PF} \qquad (4)$$

where $\mathcal{L}_{PE}$ is the pose estimation loss between the estimated pose $\widehat{H}_j^t$ and the ground truth H_j^t at the current timestep t; $\mathcal{L}_{PF}$ is the auxiliary pose forecasting loss between the sequence of $k = 1, ..., T$ generated future poses H_j^{t+k} and their corresponding ground truths $\widehat{H}_j^{t+k}$; and $\mathcal{J}$ is the set of skeleton joints in both the *uv* and *uz* views. Note that $\widehat{H}_j^t$ is generated by the pose estimation branch whether $\widehat{H}_j^{t+k}$ are generated by the pose forecasting branch.

4. Experimental Validation

4.1. Datasets

SimBa [49] is a recent dataset specifically acquired for the robot pose estimation task from depth data. It presents unique features such as the presence of synthetic and real depth data, acquired with Gazebo and the Microsoft Kinect v2 sensor. Both domains consist of several sequences of random pick-and-place operations, acquired through randomly placed cameras (left, right and center). The acquired depth data leverages the Time-of-Flight technology and has a spatial resolution of 510×424. This dataset has challenges due to different domains for training and testing (Sim2Real scenario) and different positions of the acquisition devices.

ITOP [20] consists of 20 subjects performing 15 different complex actions, for a total of 50k frames (40k training and 10k testing, as reported in the original paper). Two Structured Light (SL) depth sensors (Asus Xtion Pro) are used to acquire data, one placed in front of the subject, and one placed on the top: in this paper, we focus on the side view, in which human joints are not fully occluded by the head and shoulders of the subject. Annotations consist of 2D and 3D joint coordinates, manually refined to lie inside the body to address human pose estimation from depth data. Unfortunately, not all annotations are valid, thus limiting the length of temporally consistent sequences. The challenges of this dataset are related to the limited quality of depth data, in terms of spatial resolution (320×240), depth accuracy (SL technology [47]), and action complexity, with several occlusions produced during movements.

The proposed system has been trained and tested on the SimBa dataset [49], specifically created for the estimation of robotic joints from depth images. In addition, we demonstrate the generalization capabilities of our approach by testing the system on the ITOP [20] dataset, which has characteristics similar to the context of our interest, albeit applied to human poses.

			mAP (%) ↑					
Input	Model	Horizon	2cm	4cm	6cm	8cm	10cm	ADD (cm) ↓
M past poses	*Linear*	t	0.31	6.02	15.81	25.78	41.23	$16.89_{\pm 5.73}$
M past poses	*Linear*	$t + 0.5s$	0.42	5.54	15.34	25.40	40.58	$17.54_{\pm 6.20}$
M past poses	*Linear*	$t + 1s$	0.29	4.78	14.76	23.44	38.08	$19.25_{\pm 6.20}$
M past poses	*Linear*	$t + 1.5s$	0.32	4.34	14.11	22.76	36.72	$19.75_{\pm 6.17}$
M past poses	*Linear*	$t + 2s$	0.37	3.98	13.72	21.84	35.96	$20.04_{\pm 6.10}$
M past poses	*Ours*	t	5.33	22.77	37.42	57.96	78.05	$8.38_{\pm 3.88}$
M past poses	*Ours*	$t + 0.5s$	4.77	20.74	37.31	55.63	76.53	$8.61_{\pm 4.07}$
M past poses	*Ours*	$t + 1s$	4.41	19.65	35.58	53.16	73.58	$9.09_{\pm 4.04}$
M past poses	*Ours*	$t + 1.5s$	4.12	19.34	33.40	51.65	72.08	$9.73_{\pm 4.23}$
M past poses	*Ours*	$t + 2s$	4.02	18.81	32.56	50.32	70.21	$10.41_{\pm 4.59}$
Depth + M past poses	*Ours*	t	**30.68**	**66.90**	**92.69**	**98.02**	**98.38**	$\mathbf{3.52_{\pm 1.30}}$
Depth + M past poses	*Ours*	$t + 0.5s$	**31.32**	**66.04**	**91.71**	**97.66**	**98.33**	$\mathbf{3.57_{\pm 1.33}}$
Depth + M past poses	*Ours*	$t + 1s$	**28.89**	**59.67**	**84.39**	**91.04**	**92.65**	$\mathbf{4.50_{\pm 2.25}}$
Depth + M past poses	*Ours*	$t + 1.5s$	**26.41**	**55.99**	**78.14**	**85.93**	**87.93**	$\mathbf{5.71_{\pm 3.48}}$
Depth + M past poses	*Ours*	$t + 2s$	**25.04**	**53.43**	**73.52**	**81.27**	**83.39**	$\mathbf{6.85_{\pm 4.38}}$

Table 2. Results on both robot pose estimation and forecasting on SimBa. The proposed method is compared to a linear model and our model without the depth-based input branch, while tested in an autoregressive manner.

4.2. Metrics

For the 3D pose estimation and forecasting tasks, we exploit standard literature metrics, *i.e.* Average Distance metric (ADD) and mean Average Precision (mAP). The first, that is the L_2 distance expressed in centimeters of all 3D robot joints to their ground truth positions, conveys the error related to the translation and rotation in the 3D world (the lower the better). The second metric is defined as:

$$\text{mAP} = \frac{1}{|N|} \sum_{j \in N} \left(\|\mathbf{v}_j - \widehat{\mathbf{v}}_j\|_2 < \delta \right) \quad (5)$$

where N is the number of skeleton joints, $\mathbf{v}_j$ is the predicted joint and $\widehat{\mathbf{v}}_j$ is the ground truth. This metric is intended as the accuracy of the ADD using different thresholds ($\delta = \{2, 4, 6, 8, 10\}$ centimeters in our experiments and it improves the interpretability of results.

4.3. Training

The proposed model is trained for 30 epochs by exploiting the MSE loss for the heatmaps produced by both the branches for the current and future poses. We use the Adam optimizer, with an initial learning rate of 10^{-3}, a decay factor of 10^{-1} at 50% and 75% of the training procedure and a batch size of 16. In all experiments, we use the original dataset splits to train and test the model.

During the training on both datasets, we apply data augmentation on the point clouds computed from the input depth maps. Specifically, 3D points are randomly translated with a maximum range of $[-20\text{cm}, +20\text{cm}]$ and
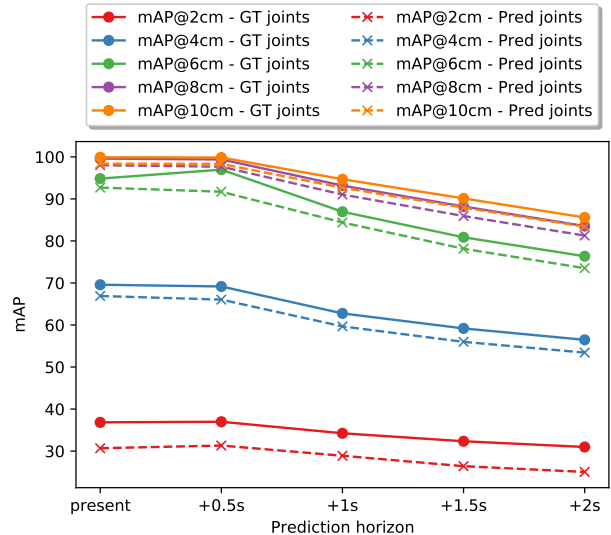

Figure 4. Comparison on Simba in terms of mAP using ground truth and predicted 3D joints as input to pose forecasting branch.

$[-30\text{cm}, +30\text{cm}]$ for XY and Z axes, respectively. Moreover, the points are rotated with a range of $[-5°, +5°]$ for the XZ axes. In terms of visual appearance, we introduce a pepper noise on about 15% of the pixels and a random dropout, consisting in setting with the value 0 several small portions of the input image: in this manner, we simulate the presence of depth noise, usually found in real-world depth sensors, and the presence of non-reflecting surfaces (on which the depth value is not valid) in the acquired scene.

	mAP (%) at 10cm ↑												
Joint	RF [48]	IEF [7]	VI [20]	RTW [60]	CMB [55]	REN-9x6x6 [19]	A2J [59]	V2V* [39]	DECA-D3 [15]	WSM [62]	AdaPose [63]	*Ours w/o forecasting*	*Ours*
Head	63.8	96.2	98.1	97.8	97.7	98.7	98.5	98.3	93.9	98.1	98.4	98.9	98.6
Neck	86.4	85.2	97.5	95.8	98.5	99.4	99.2	99.1	97.9	99.5	98.7	99.0	99.4
Shoulders	83.3	77.2	96.5	94.1	75.9	96.1	96.2	97.2	95.2	94.7	95.4	97.5	97.6
Elbows	73.2	45.4	73.3	77.9	62.7	74.7	78.9	80.4	84.5	82.8	90.7	84.4	84.4
Hands	51.3	30.9	68.7	70.5	84.4	55.2	68.3	67.3	56.5	69.1	82.1	76.8	77.4
Torso	65.0	84.7	85.6	93.8	96.0	98.7	98.5	98.7	99.0	99.7	99.7	98.7	98.8
Hips	50.8	83.5	72.0	80.3	87.9	91.8	90.8	93.2	97.4	95.7	96.4	87.6	90.4
Knees	65.7	81.8	69.0	68.8	84.4	89.0	90.7	91.8	94.6	91.0	94.4	86.8	89.7
Feet	61.3	80.9	60.8	68.4	83.8	81.1	86.9	87.6	92.0	89.9	92.8	75.3	88.0
Upper body	70.7	61.0	84.0	84.8	80.6	–	–	–	83.0	–	–	90.3	90.4
Lower body	59.3	82.1	67.3	72.5	86.5	–	–	–	95.3	–	–	85.5	90.7
Total body	65.8	71.0	77.4	80.5	83.4	84.9	88.0	88.7	88.7	89.6	**93.4**	88.0	<u>90.6</u>

Table 3. Per-joint results on human pose estimation on ITOP side-view test set. The best result is reported in bold, while the second best is underlined. As shown, the proposed framework achieves a significant accuracy on the total body, even though not expressively developed for the HPE task. Method marked with * uses 10 models ensemble.

4.4. Results

We report results on SimBa and ITOP, both with our full pipeline and with a baseline not leveraging the nowcasting paradigm. In all experiments, when the model is optimized to forecast the future, past poses are fed at 10Hz for a duration of 1s. In output instead, we sample poses at 2Hz with a temporal horizon of 2s maximum.

Results on SimBa. Table 1 shows results on the SimBa dataset, reporting mean Average Precision (mAP) using different thresholds ($\delta = \{2, 4, 6, 8, 10\}$ cm) as well as ADD. We report results using only the depth image (*Ours w/o forecasting*) and with the additional input of past predicted 3D joints (*Ours*). Following [49], we test the same competitors to predict the 3D poses reporting the results in Table 1. In particular, we train a ResNet-18 [22] to directly regress 3D coordinates from depth maps. We then evaluate the method proposed in [38], a sequence of MLPs trained to estimate 3D joint coordinates relying on their 2D positions. This approach only provides relative joint locations with respect to a specific root (the robot base). The third competitor, is based on the volumetric heatmap approach [41], a representation for encoding 3D locations in a sampled 3D volume. This approach, in addition to a limited accuracy, leads to a significant video memory occupation of about 16GB, considerably higher than all the other methods (approximately 9 times higher than ours, see Section 4.5). Finally, [49] uses the SPDH representation with a standard CNN. Even without the use of the GRU input our approach yields the state of the art on SimBa. Interestingly, when exploiting past joints' locations with a recurrent network and adding the pose forecasting branch, results are improved further especially at low spatial thresholds, almost doubling mAP at the 2cm mark.

	mAP (%) ↑					
Horizon	**2**cm	**4**cm	**6**cm	**8**cm	**10**cm	**ADD** (cm) ↓
t	10.19	38.76	64.32	79.12	86.57	6.49
$t + 0.5s$	1.94	9.61	21.48	33.91	44.75	17.66
$t + 1s$	1.20	6.72	16.39	27.78	38.56	18.94

Table 4. Results on human pose estimation and forecasting on ITOP side-view test set. The model takes as input both depth and past poses.

Then, we show in Figure 4 the results for 3D Pose Forecasting by comparing mAP at different future timestamps. As an upper bound, we report results relying on ground truth past joints' locations. Interestingly, even when autoregressively feeding back estimated joints as input, the performance drop is limited with a maximum difference of 6% for the 2cm threshold. Finally, as shown in Table 2, it must be noted that at $1s$ ADD is roughly 1cm higher than the ADD at the current timestep prediction, making the approach suitable for collision detection. Table 2 also shows a comparison between a simple baseline made of a linear regressor trained with SGD and our model with only the encoder-decoder for the forecasting branch. In the latter, the HRNet backbone extracting information from depth images is not used. In both configurations, we obtain much worse results, indicating the non-triviality of the task. In Figure 5 (right) we show qualitative results for poses predicted by our model with and without the forecasting branch, highlighting its importance.

Results on ITOP. We show in Table 3 our results compared to the state-of-the-art. Overall results for all methods on ITOP are generally worse than on SimBa, due to the
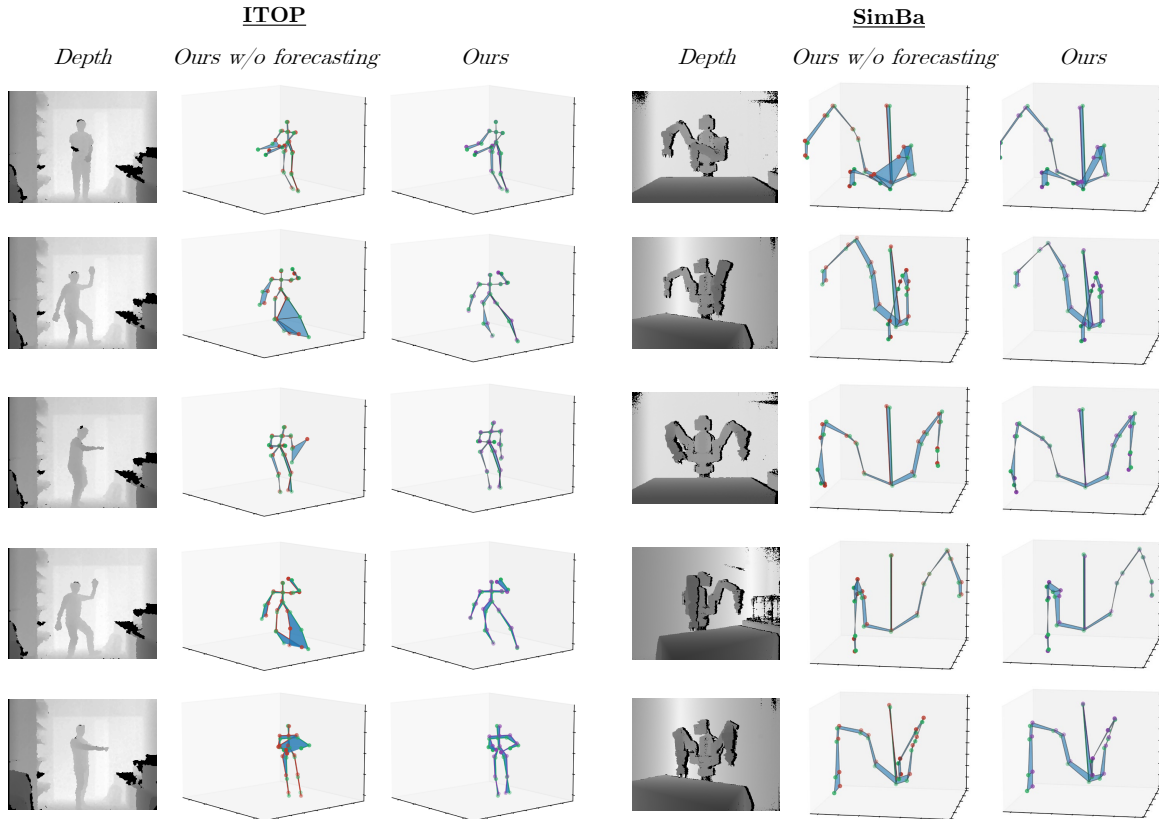
Figure 5. Qualitative examples for both ITOP and SimBa datasets where it can be appreciated the improvement in the pose estimation using the proposed approach. Green joints represent the ground truth pose, whereas red and violet represent respectively the poses estimated by our method without future and our full method. Blue regions connect ground truth skeletons and predictions, highlighting errors.

fact human movements are more erratic and complex with respect to robot arm motion. Moreover, training is made more challenging by the presence of invalid joints, *i.e.* joints without any manual annotation in the dataset. Nonetheless, on average considering the total body, our approach using a single depth frame is on par with most competing methods. Adding the supervision on future timesteps we rank above all methods except for AdaPose [63], an approach expressively developed for the HPE task (differently from ours) which obtains a slightly higher mAP metric.

Furthermore, it is interesting to notice which joints benefit the most from nowcasting, *i.e.* adding the forecasting branch. In general, the lower body registers a considerable improvement between the two variants of our approach. Hips and knees report a gain of approximately $+3\%$ mAP, whereas feet even $+13\%$ mAP. Given that feet demonstrate greater dynamism in comparison to other body joints, they manifest behavior that is comparatively less erratic than, for instance, hands, wherein the advantageous outcome is less apparent.

In Table 4 we show the performance of the framework addressing the forecasting task, which is more challenging in the presence of wide movements performed by humans. These results can be a useful baseline reference for future works that address the forecasting task on ITOP. In Figure 5 (left) we show qualitative results on ITOP, comparing the model with the present-only baseline.

4.5. Execution Time Analysis

Our model must be deployable in a work environment, thus must be efficient for safety applications, *e.g.* avoiding collisions and hazards. We measured inference time on an Intel i7 (2.90 GHz) CPU and Nvidia Titan XP GPU. The pose estimation branch alone runs at 20 FPS. Adding the forecasting branch, observing autoregressively generated poses and estimating future ones, the overall inference time is around 11 FPS with a video memory occupation of about 1.8GB. Since we feed to the architecture 1 second of 3D poses sampled at 10Hz and estimated by the model itself, we can run the whole framework in real-time without delays. The reaction time after observing the present frame before estimating the current and future poses is 90ms.

5. Conclusion and Future Work

We introduced the paradigm of 3D Pose Nowcasting, using depth data. The proposed framework jointly optimizes pose estimation and forecasting, exploiting two branches and the SPDH intermediate representation. We obtain state-of-the-art results in predicting current and near-future robot poses. The framework is also able to work with humans, achieving performance comparable with the current literature competitors on ITOP. In future work, we plan to adopt Domain Adaptation techniques to reduce the Sim2Real shift, and the use of recent transformer-based architectures to model the input sequences. Finally, we highlight the lack of depth-based datasets regarding human-machine interaction in social and working scenarios. This kind of data could lead to the realization of real-world collision detection and anticipation systems.

References

[1] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13390–13400, 2021. 3

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016. 3

[3] Sizhe An, Yin Li, and Umit Ogras. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in Neural Information Processing Systems*, 35:27414–27426, 2022. 4

[4] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2

[5] Jeannette Bohg, Javier Romero, Alexander Herzog, and Stefan Schaal. Robot arm pose estimation through pixel-wise part classification. In *Proc. of the IEEE International Conference on Robotics and Automation*, pages 3143–3150, 2014. 2

[6] KA Browning and CiG Collier. Nowcasting of precipitation systems. *Reviews of Geophysics*, 27(3):345–370, 1989. 2

[7] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4742, 2016. 7

[8] Yu-Wei Chao, Jimei Yang, Brian Price, Scott Cohen, and Jia Deng. Forecasting human dynamics from static images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 548–556, 2017. 3

[9] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1423–1432. IEEE, 2019. 3

[10] Ed Colgate, Antonio Bicchi, Michael Aaron Peshkin, and James Edward Colgate. Safety for physical human-robot interaction. In *Springer Handbook of Robotics*, pages 1335–1348. Springer, 2008. 1

[11] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019. 4

[12] Kerstin Dautenhahn and Joe Saunders. *New frontiers in human robot interaction*, volume 2. John Benjamins Publishing, 2011. 1

[13] Christian Diller, Thomas Funkhouser, and Angela Dai. Forecasting characteristic 3d poses of human actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15914–15923, 2022. 3

[14] Nicola Garau, Niccolo Bisagno, Piotr Bródka, and Nicola Conci. Deca: Deep viewpoint-equivariant human pose estimation using capsule autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11677–11686, 2021. 2, 4

[15] Nicola Garau and Nicola Conci. Capsules as viewpoint learners for human pose estimation. *arXiv preprint arXiv:2302.06194*, 2023. 4, 7

[16] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014. 2

[17] Milad Geravand, Fabrizio Flacco, and Alessandro De Luca. Human-robot physical interaction and collaboration using an industrial robot with a closed control architecture. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 4000–4007. IEEE, 2013. 1

[18] Quentin Guimard, Lucile Sassatelli, Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Deep variational learning for multiple trajectory prediction of 360° head movements. In *Proceedings of the ACM Multimedia Systems Conference*, pages 12–26, 2022. 3

[19] Hengkai Guo, Guijin Wang, Xinghao Chen, and Cairong Zhang. Towards good practices for deep 3d hand pose estimation. *arXiv preprint arXiv:1707.07248*, 2017. 3, 7

[20] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Li Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 160–177. Springer, 2016. 4, 5, 7

[21] Hiroaki Hasegawa, Yoshitomo Mizoguchi, Kenjiro Tadakuma, Aiguo Ming, Masatoshi Ishikawa, and Makoto Shimojo. Development of intelligent robot hand using proximity, contact and slip sensing. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 777–784. IEEE, 2010. 1

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5, 7

[23] Sebastian Höfer, Kostas Bekris, Ankur Handa, Juan Camilo Gamboa, Melissa Mozifian, Florian Golemo, Chris Atkeson, Dieter Fox, Ken Goldberg, John Leonard, et al. Sim2real in robotics and automation: Applications and challenges. *IEEE Transactions on Automation Science and Engineering*, 18(2):398–400, 2021. 2, 3

[24] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 4

[25] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2375–2384, 2019. 3

[26] Michail Kalaitzakis, Brennan Cain, Sabrina Carroll, Anand Ambrosi, Camden Whitehead, and Nikolaos Vitzilaios. Fiducial markers for pose estimation: Overview, applications and experimental comparison of the artag, apriltag, aruco and stag markers. *Journal of Intelligent & Robotic Systems*, 101:1–26, 2021. 1

[27] Nathan Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 2149–2154. IEEE. 3

[28] Ari Kolbeinsson, Erik Lagerstedt, and Jessica Lindblom. Foundation for a classification of collaboration levels for human-robot cooperation in manufacturing. *Production & Manufacturing Research*, 7(1):448–471, 2019. 1

[29] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Single-view robot pose and joint angle estimation via render & compare. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1654–1663, 2021. 2

[30] Heiner Lasi, Peter Fettke, Hans-Georg Kemper, Thomas Feld, and Michael Hoffmann. Industry 4.0. *Business & information systems engineering*, 6(4):239–242, 2014. 1

[31] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proc. of the IEEE/CVF CVPR*, pages 336–345, 2017. 3

[32] Timothy E Lee, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Oliver Kroemer, Dieter Fox, and Stan Birchfield. Camera-to-robot pose estimation from a single image. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 9426–9432, 2020. 1, 3

[33] Vincent Lepetit, Francesc Moreno-Noguer, and P Fua. Epnp: Efficient perspective-n-point camera pose estimation. *International Journal of Computer Vision*, 81(2):155–166, 2009. 2

[34] Xin Li and Dawei Li. Gpfs: a graph-based human pose forecasting system for smart home with online learning. *ACM Transactions on Sensor Networks*, 17(3):1–19, 2021. 3

[35] Pauline Luc, Camille Couprie, Yann Lecun, and Jakob Verbeek. Predicting future instance segmentation by forecasting convolutional features. In *Proceedings of the European Conference on Computer Vision*, pages 584–599, 2018. 3

[36] Karttikeya Mangalam, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2784–2793, 2020. 3

[37] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Multiple trajectory prediction of moving agents with memory augmented networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[38] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 5, 7

[39] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5079–5088, 2018. 3, 7

[40] Atsuhiro Noguchi, Umar Iqbal, Jonathan Tremblay, Tatsuya Harada, and Orazio Gallo. Watch it move: Unsupervised discovery of 3d joints for re-posing of articulated objects. In *Proc. of the IEEE/CVF Conference on CVPR*, pages 3677–3687, 2022. 2, 3

[41] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7025–7034, 2017. 5, 7

[42] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018. 3

[43] Michael A Peshkin, J Edward Colgate, Wit Wannasuphoprasit, Carl A Moore, R Brent Gillespie, and Prasad Akella. Cobot architecture. *IEEE Transactions on Robotics*, 17(4):377–390, 2001. 1

[44] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In *Pattern Recognition*, pages 694–701, 2021. 3

[45] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6457–6466, 2022. 3

[46] Alessio Sampieri, Guido Maria D'Amely di Melendugno, Andrea Avogaro, Federico Cunico, Francesco Setti, Geri Skenderi, Marco Cristani, and Fabio Galasso. Pose forecasting in industrial human-robot collaboration. In *Proc. of the European Conference on Computer Vision*, pages 51–69, 2022. 3

[47] Hamed Sarbolandi, Damien Lefloch, and Andreas Kolb. Kinect range sensing: Structured-light versus time-of-flight kinect. *Computer Vision Image Understanding*, 139:1–20, 2015. 2, 5

[48] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2821–2840, 2012. 2, 4, 7

[49] Alessandro Simoni, Stefano Pini, Guido Borghi, and Roberto Vezzani. Semi-perspective decoupled heatmaps for 3d robot pose estimation from depth maps. *IEEE Robotics and Automation Letters*, 7(4):11569–11576, 2022. 2, 3, 4, 5, 7

[50] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11209–11218, 2021. 3

[51] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 4

[52] Sam Toyer, Anoop Cherian, Tengda Han, and Stephen Gould. Human pose forecasting via deep markov models. In *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications*, pages 1–8. IEEE, 2017. 3

[53] Neel Trivedi, Anirudh Thatipelli, and Ravi Kiran Sarvadevabhatla. Ntu-x: an enhanced large-scale dataset for improving pose-based recognition of subtle human actions. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2021. 4

[54] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 98–106, 2016. 3

[55] Keze Wang, Liang Lin, Chuangjie Ren, Wei Zhang, and Wenxiu Sun. Convolutional memory blocks for depth data representation learning. In *Proceeding of the International Joint Conferences on Artificial Intelligence*, pages 2790–2797, 2018. 2, 4, 7

[56] Astrid Weiss, Roland Buchner, Manfred Tscheligi, and Hanspeter Fischer. Exploring human-robot cooperation possibilities for semiconductor manufacturing. In *Proceedings of the International Conference on Collaboration Technologies and Systems*, pages 173–177. IEEE, 2011. 1

[57] Astrid Weiss, Ann-Kathrin Wortmeier, and Bettina Kubicek. Cobots in industry 4.0: A roadmap for future practice studies on human–robot collaboration. *IEEE Transactions on Human-Machine Systems*, 51(4):335–345, 2021. 1

[58] Felix Widmaier, Daniel Kappler, Stefan Schaal, and Jeannette Bohg. Robot arm pose estimation by pixel-wise regression of joint angles. In *Proc. of the International Conference on Robotics and Automation*, pages 616–623, 2016. 2

[59] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 793–802, 2019. 3, 7

[60] Ho Yub Jung, Soochahn Lee, Yong Seok Heo, and Il Dong Yun. Random tree walk toward instantaneous 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2467–2474, 2015. 2, 7

[61] Pietro Zanuttigh, Giulio Marin, Carlo Dal Mutto, Fabio Dominio, Ludovico Minto, and Guido Maria Cortelazzo. Time-of-flight and structured light depth cameras. *Technology and Applications*, pages 978–3, 2016. 2

[62] Zihao Zhang, Lei Hu, Xiaoming Deng, and Shihong Xia. Weakly supervised adversarial learning for 3d human pose estimation from point clouds. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1851–1859, 2020. 2, 4, 7

[63] Zihao Zhang, Lei Hu, Xiaoming Deng, and Shihong Xia. Sequential 3d human pose estimation using adaptive point cloud sampling strategy. In *Proceeding of the International Joint Conferences on Artificial Intelligence*, pages 1330–1337, 2021. 2, 4, 7, 8

[64] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023. 1

[65] Yiming Zuo, Weichao Qiu, Lingxi Xie, Fangwei Zhong, Yizhou Wang, and Alan L Yuille. Craves: Controlling robotic arm with a vision-based economic system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4214–4223, 2019. 3