Computer Vision and Image Understanding 251 (2025) 104233

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

3D Pose Nowcasting: Forecast the future to improve the present

Alessandro Simoni [a], Francesco Marchetti [b,*], Guido Borghi [c], Federico Becattini [d], Lorenzo Seidenari [b], Roberto Vezzani [a], Alberto Del Bimbo [b]

[a] Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, 41100, Italy
[b] Dipartimento di Ingegneria dell'Informazione, University of Florence, Firenze, 50134, Italy
[c] Department of Education and Humanities, University of Modena and Reggio Emilia, Reggio Emilia, 42121, Italy
[d] Dipartimento di Ingegneria dell'informazione e scienze matematiche, University of Siena, Siena, 53100, Italy

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Human pose estimation
Robot pose estimation
Pose forecasting
Pose Nowcasting
3D pose

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 (Peshkin et al., 2001).

Non-invasive camera monitoring combined with specific computer vision algorithms, such as Robot and Human Pose Estimators (Zheng et al., 2023; Lee et al., 2020), are key and enabling technologies for safe interaction between humans and robots (Colgate et al., 2008). For instance, in the Industry 4.0 setting (Lasi et al., 2014), in which the same workplace is shared between workers and cobots (Kolbeinsson et al., 2019), the ability to detect poses and avoid collisions is fundamental for safety. Furthermore, recent investigations (Weiss et al., 2011, 2021) confirm that – rather than the complete removal of humans – future generations of manufacturing will support the coexistence of humans and cobots, stressing the urgency for new investigations related to physical and social coworker coordination (Dautenhahn and Saunders, 2011). 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 (Hasegawa et al., 2010), markers (Kalaitzakis et al., 2021) or access to the robot's proprietary software (Geravand et al., 2013), 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 (Browning and Collier, 1989), 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 nowcasting paradigm is based on jointly predicting present and future states given the past ones, motivated by the fact that future forecasting can aid the

* Corresponding author.
E-mail addresses: alessandro.simoni@unimore.it (A. Simoni), francesco.marchetti@unifi.it (F. Marchetti), guido.borghi@unimore.it (G. Borghi), federico.becattini@unisi.it (F. Becattini), lorenzo.seidenari@unifi.it (L. Seidenari), roberto.vezzani@unimore.it (R. Vezzani), alberto.delbimbo@unifi.it (A. Del Bimbo).

https://doi.org/10.1016/j.cviu.2024.104233
Received 30 May 2024; Received in revised form 28 October 2024; Accepted 11 November 2024
Available online 20 November 2024

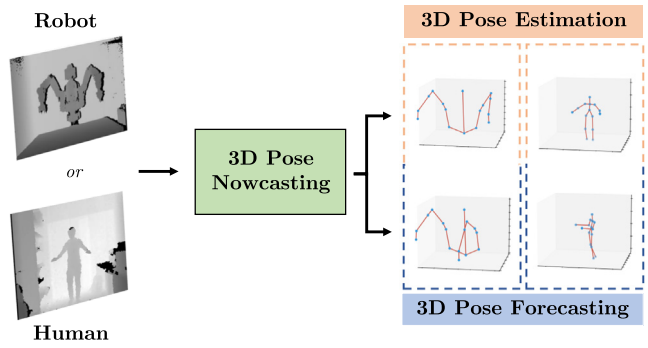

Fig. 1. Estimating current and future poses through 3D Pose Nowcasting, using depth images as input data, is a fundamental technology for safe interaction between workers and collaborative machines in indoor scenarios, such as the Industry 4.0 setting.

prediction of the present. No future data, other than the past states, is given as input to the model. The nowcasting paradigm distinguishes itself from the direct estimation and forecasting paradigms, which either focus on the present or the future alone, and can potentially bring benefits to both.

The proposed method for 3D Pose Nowcasting, outlined in Fig. 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 (Sarbolandi et al., 2015), usually common in indoor scenarios such as workplaces. Besides, depth acquisition devices nowadays are inexpensive, yet accurate (Zanuttigh et al., 2016). Moreover, in the Sim2Real (Höfer et al., 2021) setting, the use of depth reduces the domain gap between synthetic and real scenarios (Simoni et al., 2022), thus enabling the usage of large-scale datasets without the time-consuming collecting and labeling procedures required with real data. This application demonstrates that the method not only merges established techniques but also optimizes them for real-world scenarios, highlighting its practical value.

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 (Andriluka et al., 2014) (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) (Simoni et al., 2022) 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. By integrating depth maps with pose forecasting, the method addresses specific challenges in robot pose estimation that have not been effectively tackled by each technique in isolation.

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. (2014) proposed to use a random forest classifier to classify and then group depth maps pixels, obtaining skeleton joints. A similar approach is reported by Widmaier et al. (2016), 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 (Garrido-Jurado et al., 2014), 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 (Lepetit et al., 2009) 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 (Labbé et al., 2021; Noguchi et al., 2022) 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. (2012) 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, (Yub Jung et al., 2015) proposed to use a regression tree to predict the probability distribution of the direction of a specific joint. This approach is able to run up to 1000 fps.

Entering the deep learning-based field, some works introduce the use of NNs in combination with a single depth frame. In the work of Wang et al. (2018), 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, Garau et al. (2021) 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 by Zhang et al. (2020) is based on a point clouds proposal module followed by a 3D pose regression module. Similarly, the same authors (Zhang et al., 2021) 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 literature works have been developed originally for the hand pose estimation task (Moon et al., 2018; Xiong et al., 2019; Guo et al., 2017) and then adapted to tackle also the human pose estimation task.

Pose forecasting. Recently, Sampieri et al. (2022) 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 (Lee et al., 2017; Marchetti et al., 2020; Ivanovic and Pavone, 2019; Luc et al., 2018) and human behavior understanding (Pavllo et al., 2018; Toyer et al., 2017;

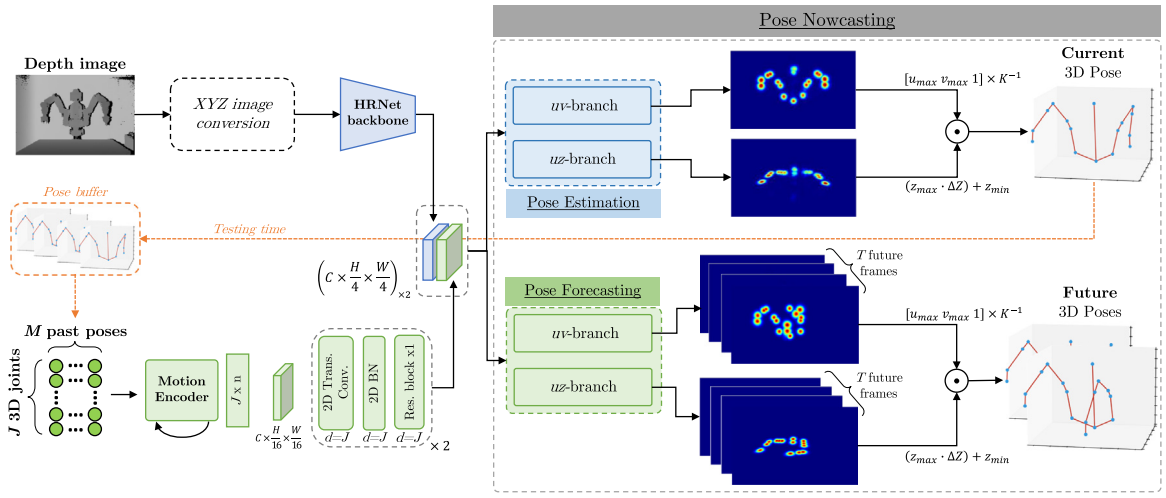

Fig. 2. Overview of the proposed 3D Pose Nowcasting framework. First, features related to the depth map and the past poses are extracted. These features are then concatenated and fed to two different branches, i.e. the Pose Estimation and Pose Forecasting ones. Finally, the framework outputs the current and the near-future 3D poses. For the sake of visualization, heatmaps are stacked channel-wise.

Chiu et al., 2019; Chao et al., 2017; Diller et al., 2022). 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 (Lee et al., 2017; Alahi et al., 2016). Interestingly, most approaches formulate the forecasting task as a multimodal prediction task, due to the intrinsic uncertainty of the problem (Vondrick et al., 2016; Lee et al., 2017; Salzmann et al., 2022; Guimard et al., 2022). 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 (Plizzari et al., 2021; Li and Li, 2021; Adeli et al., 2021; Sofianos et al., 2021). Unlike these methods, Mangalam et al. (2020) 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 Mangalam et al. (2020), 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 (Zuo et al., 2019) dataset consists of synthetic and real acquisitions of a single type of robotic arm, for a total of about 5k frames. DREAM (Lee et al., 2020) and WIM (Noguchi et al., 2022), 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 (Sampieri et al., 2022). 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 (Simoni et al., 2022), 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 (Höfer et al., 2021) scenario, i.e. the training and testing frames belong to two different domains: synthetic (generated through ROS and Gazebo (Koenig and Howard) 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.

With regard to the estimation of human poses, we adopt the ITOP dataset (Haque et al., 2016), which has been used as a benchmark by several prior works (Zhang et al., 2020; Garau et al., 2021; Zhang et al., 2021; Wang et al., 2018; Garau and Conci, 2023). 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 (Ionescu et al., 2013) dataset contains very low-quality depth images, acquired through the MESA Imaging SR4000 device. The NTU dataset (Trivedi et al., 2021), 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 Shotton et al. (2012), reducing their accuracy. The mRI dataset (An et al., 2022) 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 Fig. 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 (Simoni et al., 2022) 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 $t = 1, \dots, M$ poses $P_j^t = [X_j^t, Y_j^t, Z_j^t]$ with $j = 1, \dots, J$ 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 $I_{XYZ} = \pi(D \cdot K^{-1})$ 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

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

Table 1
Robot pose estimation results on SimBa. The proposed framework is tested by taking different combinations regarding the input (only Depth or Depth+M past poses) and the presence or absence of the auxiliary pose forecasting task (*Ours* or *Ours w/o forecasting*). Method marked with * uses a relative joint representation.

Input	Model	mAP (%) ↑ 2cm	4cm	6cm	8cm	10cm	ADD (cm) ↓
Depth	ResNet-18 (He et al., 2016)	0.57	9.40	19.99	27.06	44.44	12.20±4.12
2D joints	Martinez et al. (2017)*	13.70	26.96	37.98	48.40	58.33	10.03±3.53
Depth	Pavlakos et al. (2017)	3.35	18.15	42.24	61.60	86.15	7.11±0.65
Depth	Simoni et al. (2022)	6.33	53.75	79.75	93.90	98.12	4.41±1.09
Depth	*Ours w/o forecasting*	16.25	57.51	89.81	**99.26**	**99.81**	3.77±0.98
Depth + M past poses	*Ours w/o forecasting*	17.68	62.17	90.94	97.62	99.41	3.67±1.19
Depth	*Ours*	17.92	66.10	91.50	98.59	99.68	3.60±0.75
Depth + M past poses	*Ours*	**30.68**	**66.90**	**92.69**	98.02	98.38	**3.52**±1.30

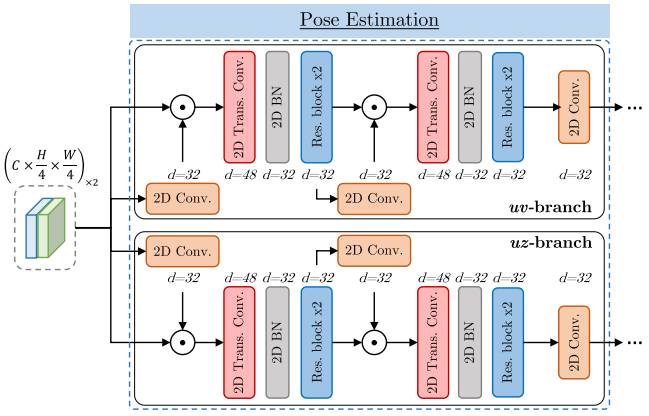

Fig. 3. Architecture of the Pose Estimation branch. The input is represented by the concatenation of features extracted from depth maps and past joints. Each uv/uz sub-branch generates the heatmap-based SPDH representation of 3D joint locations.

to have better generalization capabilities across different domains with respect to common depth images (Simoni et al., 2022). Being aware of the recent and significant advances in HPE (Dang et al., 2019), we exploit the well-known HRNet-32 architecture (Sun et al., 2019), 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: 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 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 Fig. 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, 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 $\mathcal{L} = \mathcal{L}_{PE} + \mathcal{L}_{PF}$ between generated and ground truth heatmaps:

$$\mathcal{L}_{PE} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \| H_j^t - \hat{H}_j^t \|_2 \tag{1}$$

$$\mathcal{L}_{PF} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \frac{1}{T} \sum_{k=1}^{t+T} \| H_j^{t+k} - \hat{H}_j^{t+k} \|_2 \tag{2}$$

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

4. Experimental validation

4.1. Datasets

SimBa (Simoni et al., 2022) 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

[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.

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 or without the bast branch, while tested in an autoregressive manner.

			mAP (%) ↑					
Input	Model	Horizon	**2**cm	**4**cm	**6**cm	**8**cm	**10**cm	ADD (cm) ↓
M past poses	*Linear*	t	0.31	6.02	15.81	25.78	41.23	16.89±5.73
M past poses	*Linear*	$t+0.5s$	0.42	5.54	15.34	25.40	40.58	17.54±6.20
M past poses	*Linear*	$t+1s$	0.29	4.78	14.76	23.44	38.08	19.25±6.20
M past poses	*Linear*	$t+1.5s$	0.32	4.34	14.11	22.76	36.72	19.75±6.17
M past poses	*Linear*	$t+2s$	0.37	3.98	13.72	21.84	35.96	20.04±6.10
M past poses	*Ours*	t	5.33	22.77	37.42	57.96	78.05	8.38±3.88
M past poses	*Ours*	$t+0.5s$	4.77	20.74	37.31	55.63	76.53	8.61±4.07
M past poses	*Ours*	$t+1s$	4.41	19.65	35.58	53.16	73.58	9.09±4.04
M past poses	*Ours*	$t+1.5s$	4.12	19.34	33.40	51.65	72.08	9.73±4.23
M past poses	*Ours*	$t+2s$	4.02	18.81	32.56	50.32	70.21	10.41±4.59
Depth	*Ours*	t	17.92	66.1	91.5	98.59	**99.68**	3.60±0.75
Depth	*Ours*	$t+0.5s$	19.14	65.05	90.92	98.28	**99.58**	3.71±0.79
Depth	*Ours*	$t+1s$	17.63	58.93	84.02	91.57	**93.94**	4.61±1.90
Depth	*Ours*	$t+1.5s$	18.19	53.56	72.91	80.79	83.29	5.87±3.09
Depth	*Ours*	$t+2s$	17.42	50.60	68.12	76.04	78.52	7.15±4.06
Depth + M past poses	*Ours*	t	**30.68**	**66.90**	**92.69**	**98.02**	98.38	**3.52**±1.30
Depth + M past poses	*Ours*	$t+0.5s$	**31.32**	**66.04**	**91.71**	**97.66**	98.33	**3.57**±1.33
Depth + M past poses	*Ours*	$t+1s$	**28.89**	**59.67**	**84.39**	**91.04**	92.65	**4.50**±2.25
Depth + M past poses	*Ours*	$t+1.5s$	**26.41**	**55.99**	**78.14**	**85.93**	**87.93**	**5.71**±3.48
Depth + M past poses	*Ours*	$t+2s$	**25.04**	**53.43**	**73.52**	**81.27**	**83.39**	**6.85**±4.38

Table 3
Nowcasting framework applied to state-of-the-art methods. We observe that by predicting future joint positions, all methods gain significant improvements in mAP and ADD.

		mAP (%) ↑					
Model	Type	**2**cm	**4**cm	**6**cm	**8**cm	**10**cm	ADD (cm) ↓
Pavlakos et al. (2017)	Original	3.35	18.15	42.24	61.60	86.15	7.11±0.65
Pavlakos et al. (2017)	Nowcasting	**5.33**	**28.05**	**53.42**	**74.94**	**88.07**	**6.29**±1.38
Simoni et al. (2022)	Original	6.33	53.75	79.75	93.90	**98.12**	4.41±1.09
Simoni et al. (2022)	Nowcasting	**14.53**	**56.47**	**84.31**	**94.03**	97.56	**4.11**±0.92

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 (Haque et al., 2016) 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 Sarbolandi et al., 2015), and action complexity, with several occlusions produced during movements.

The proposed system has been trained and tested on the SimBa dataset, 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 dataset, which has characteristics similar to the context of our interest, albeit applied to human poses.

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:

$$\text{mAP} = \frac{1}{|N|} \sum_{j \in N} \left(\|\mathbf{v}_j - \hat{\mathbf{v}}_j\|_2 < \delta \right) \tag{3}$$

where N is the number of skeleton joints, $\mathbf{v}_j$ is the predicted joint and $\hat{\mathbf{v}}_j$ is the ground truth. This metric is intended as the accuracy of the per-joint L2 distance, using different thresholds ($\delta = \{2, 4, 6, 8, 10\}$ centimeters and it improves the interpretability of the 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 cm, +20 cm] and [−30 cm, +30 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.

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

Table 4
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. Methods reports are as follows: RF (Shotton et al., 2012), IEF (Carreira et al., 2016), VI (Haque et al., 2016), RTW (Yub Jung et al., 2015), CMB (Wang et al., 2018), REN-9 × 6x6 (Guo et al., 2017), A2J (Xiong et al., 2019), V2V (Moon et al., 2018), DECA-D3 (Garau et al., 2021), WSM (Zhang et al., 2020), AdaPose (Zhang et al., 2021). We observe that our method improves when pose forecasting (PF) is used.

	mAP (%) at 10 cm ↑												
	RF	IEF	VI	RTW	CMB	REN-9 × 6x6	A2J	V2V*	DECA-D3	WSM	AdaPose	*Ours w/o PF*	*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>

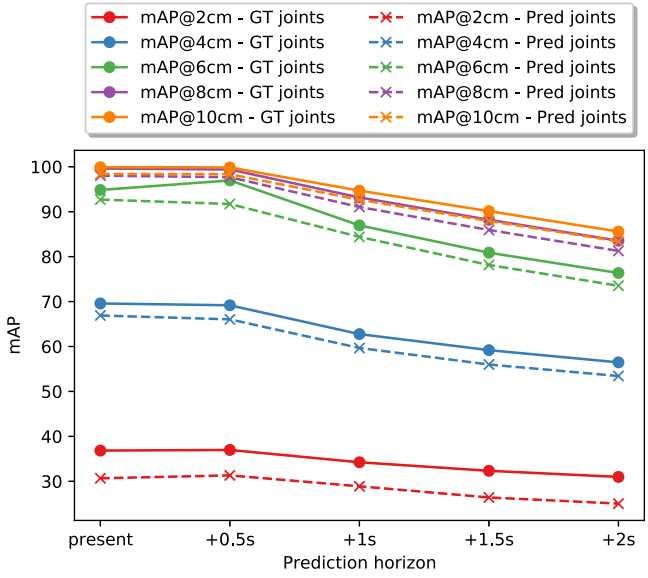

Fig. 4. Comparison on Simba in terms of mAP using ground truth and predicted 3D joints as input to pose forecasting branch.

experiments, when the model is optimized to forecast the future, past poses are fed at 10 Hz for a duration of 1s. In output instead, we sample poses at 2 Hz 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 different combinations of data input and auxiliary task. As input, we have utilized either depth or depth and past poses, whereas for auxiliary task we study the effect of including forecasting or not.

Following Simoni et al. (2022), we test the same competitors to predict the 3D poses reporting the results in Table 1. In particular, we train a ResNet-18 (He et al., 2016) to directly regress 3D coordinates from depth maps. We then evaluate the method proposed by Martinez et al. (2017), 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 of Pavlakos et al. (2017), 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 16 GB, considerably higher than all the other methods (approximately 9 times higher than ours, see Section 4.5). Finally, Simoni et al. (2022) 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 2 cm mark. Both the usage of past poses and the use of the forecasting loss, even when used individually, bring an improvement to our proposed approach.

Then, we show in Fig. 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 2 cm threshold. Finally, as shown in Table 2, it must be noted that at $1s$ ADD is roughly 1 cm 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. Similarly, we include a depth-only nowcasting baseline, that does not leverage any past information. This baseline performs better than the past-based model, but is still much worse than our proposed approach. In Fig. 5 (right) we show qualitative results for poses predicted by our model with and without the forecasting branch, highlighting its importance.

Nowcasting framework in SoTA methods. To verify the importance of applying the proposed framework in the Pose Estimation task, we have also applied the Nowcasting mechanism in other state-of-the-art models that receive the depth image as input. The models presented in Simoni et al. (2022) and Pavlakos et al. (2017) were taken into consideration, with the original results of the Robot Pose estimation reported in Table 1. We modified such methods to include a Nowcasting loss during training and we observed that this reflects in a gain in both mAP and ADD, proving that our proposed framework is general and can be easily plugged into other methods. The comparison between the original methods and their Nowcasting-based counterparts is reported in Table 3.

Results on ITOP. We show in Table 4 our results compared to the state-of-the-art. We show results for our approach both with and without pose forecasting (PF). Overall results for all methods on ITOP are generally worse than on SimBa, due to the fact human movements are

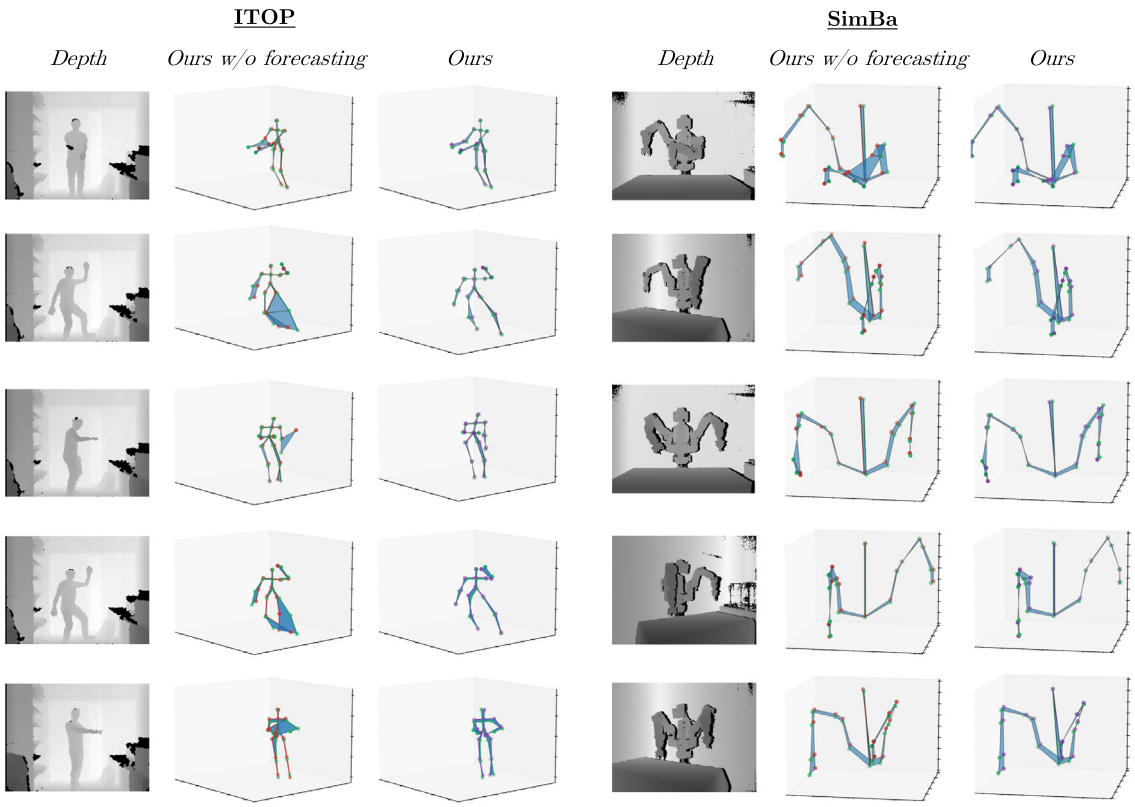

Fig. 5. Qualitative examples for both ITOP and SimBa datasets where it can be appreciated the improvement in the pose estimation using the proposed approach. Green joints represent the ground truth pose, whereas red and violet represent respectively the poses estimated by our method without future and our full method. Blue regions connect ground truth skeletons and predictions, highlighting errors.

Table 5
Results on human pose estimation and forecasting on ITOP side-view test set. The model takes as input both depth and past poses.

Horizon	mAP (%) ↑					
	2cm	**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

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 (Zhang et al., 2021), an approach expressively developed for the HPE task (differently from ours) which obtains a slightly higher mAP metric. In fact, the human context has unique characteristics, such as the low predictability of movements, which require human-specific solutions such as priors of joint velocities.

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 5 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 Fig. 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 1.8 GB. Since we feed to the architecture 1 s of 3D poses sampled at 10 Hz 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 90 ms.

4.6. Effect of noise in past poses

We have performed a quantitative analysis on what happens if the input of past pose estimations in some timesteps are noisy. We have given a Gaussian noise of 0.1 and 0.5 cm to the joints in a variable number of input past joints (between 1 and 10). The timesteps where to apply the noise are chosen randomly. This simulates a scenario where sensors or predictions are faulty, and we aim at quantifying the impact of different degrees of faultiness on the final results. In Fig. 6 the results of the mAP on SimBa are reported. As we can see from the trend of the results curve, there is a slight degradation as we increase the number of timesteps with noisy data. However, we notice that mAP degrades gracefully, maintaining a good level. Note that this experiment is carried out without retraining on noisy data.

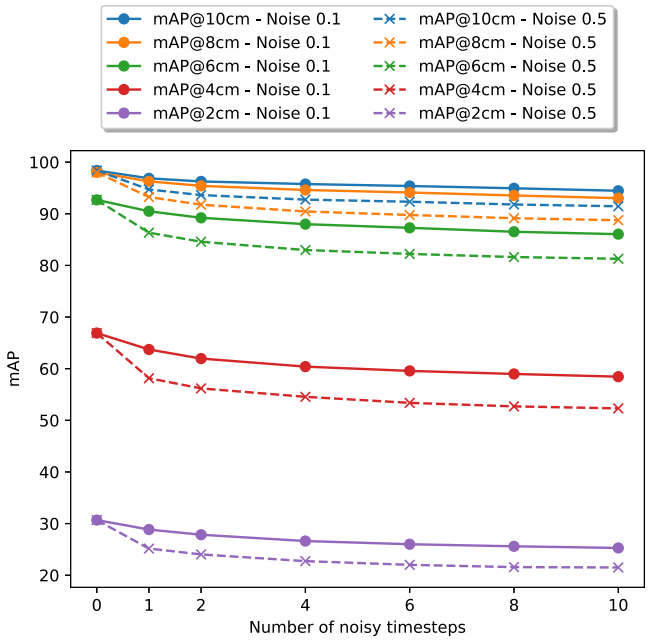

Fig. 6. Performance of the model with noise (cm) in past poses estimations.

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.

CRediT authorship contribution statement

Alessandro Simoni: Writing – original draft, Software, Methodology. **Francesco Marchetti:** Writing – review & editing, Writing – original draft, Software, Methodology. **Guido Borghi:** Writing – review & editing, Writing – original draft, Supervision. **Federico Becattini:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Lorenzo Seidenari:** Writing – review & editing, Writing – original draft, Supervision. **Roberto Vezzani:** Writing – review & editing, Writing – original draft, Supervision. **Alberto Del Bimbo:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the Piano per lo Sviluppo della Ricerca (PSR 2023) of the University of Siena - project FEATHER: Forecasting and Estimation of Actions and Trajectories for Human–robot intERactions. This work was partially supported by the European Commission under European Horizon 2020 Programme, grant number 951911—AI4Media. This work was partially supported by the Piano per lo Sviluppo della Ricerca (FAR 2022) of the University of Modena and Reggio Emilia -DIEF - project: AI platform with digital twins of interacting robots and people.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2024.104233.

Data availability

Data will be made available on request.

References

Adeli, V., Ehsanpour, M., Reid, I., Niebles, J.C., Savarese, S., Adeli, E., Rezatofighi, H., 2021. Tripod: Human trajectory and pose dynamics forecasting in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13390–13400.

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S., 2016. Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 961–971.

An, S., Li, Y., Ogras, U., 2022. Mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. Adv. Neural Inf. Process. Syst. 35, 27414–27426.

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014. 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. pp. 3686–3693.

Bohg, J., Romero, J., Herzog, A., Schaal, S., 2014. Robot arm pose estimation through pixel-wise part classification. In: Proc. of the IEEE International Conference on Robotics and Automation. pp. 3143–3150.

Browning, K., Collier, C., 1989. Nowcasting of precipitation systems. Rev. Geophys. 27 (3), 345–370.

Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J., 2016. Human pose estimation with iterative error feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4733–4742.

Chao, Y.-W., Yang, J., Price, B., Cohen, S., Deng, J., 2017. Forecasting human dynamics from static images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 548–556.

Chiu, H.-k., Adeli, E., Wang, B., Huang, D.-A., Niebles, J.C., 2019. Action-agnostic human pose forecasting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. IEEE, pp. 1423–1432.

Colgate, E., Bicchi, A., Peshkin, M.A., Colgate, J.E., 2008. Safety for physical human-robot interaction. In: Springer Handbook of Robotics. Springer, pp. 1335–1348.

Dang, Q., Yin, J., Wang, B., Zheng, W., 2019. Deep learning based 2d human pose estimation: A survey. Tsinghua Sci. Technol. 24 (6), 663–676.

Dautenhahn, K., Saunders, J., 2011. New frontiers in human robot interaction, vol. 2, John Benjamins Publishing.

Diller, C., Funkhouser, T., Dai, A., 2022. Forecasting characteristic 3D poses of human actions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15914–15923.

Garau, N., Bisagno, N., Bródka, P., Conci, N., 2021. DECA: Deep viewpoint-equivariant human pose estimation using capsule autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11677–11686.

Garau, N., Conci, N., 2023. Capsules as viewpoint learners for human pose estimation. arXiv preprint arXiv:2302.06194.

Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F.J., Marín-Jiménez, M.J., 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognit. 47 (6), 2280–2292.

Geravand, M., Flacco, F., De Luca, A., 2013. 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. IEEE, pp. 4000–4007.

Guimard, Q., Sassatelli, L., Marchetti, F., Becattini, F., Seidenari, L., Bimbo, A.D., 2022. Deep variational learning for multiple trajectory prediction of 360° head movements. In: Proceedings of the ACM Multimedia Systems Conference. pp. 12–26.

Guo, H., Wang, G., Chen, X., Zhang, C., 2017. Towards good practices for deep 3d hand pose estimation. arXiv preprint arXiv:1707.07248.

Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., Fei-Fei, L., 2016. Towards viewpoint invariant 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 160–177.

Hasegawa, H., Mizoguchi, Y., Tadakuma, K., Ming, A., Ishikawa, M., Shimojo, M., 2010. Development of intelligent robot hand using proximity, contact and slip sensing. In: Proceedings of the IEEE International Conference on Robotics and Automation. IEEE, pp. 777–784.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Höfer, S., Bekris, K., Handa, A., Gamboa, J.C., Mozifian, M., Golemo, F., Atkeson, C., Fox, D., Goldberg, K., Leonard, J., et al., 2021. Sim2Real in robotics and automation: Applications and challenges. IEEE Trans. Autom. Sci. Eng. 18 (2), 398–400.

Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. 36 (7), 1325–1339.

Ivanovic, B., Pavone, M., 2019. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2375–2384.

Kalaitzakis, M., Cain, B., Carroll, S., Ambrosi, A., Whitehead, C., Vitzilaios, N., 2021. Fiducial markers for pose estimation: Overview, applications and experimental comparison of the artag, apriltag, aruco and stag markers. J. Intell. Robot. Syst. 101, 1–26.

Koenig, N., Howard, A., 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, Vol. 3. IEEE, pp. 2149–2154.

Kolbeinsson, A., Lagerstedt, E., Lindblom, J., 2019. Foundation for a classification of collaboration levels for human-robot cooperation in manufacturing. Prod. Manuf. Res. 7 (1), 448–471.

Labbé, Y., Carpentier, J., Aubry, M., Sivic, J., 2021. Single-view robot pose and joint angle estimation via render & compare. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1654–1663.

Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., Hoffmann, M., 2014. Industry 4.0. Bus. Informat. Syst. Eng. 6 (4), 239–242.

Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H., Chandraker, M., 2017. Desire: Distant future prediction in dynamic scenes with interacting agents. In: Proc. of the IEEE/CVF CVPR. pp. 336–345.

Lee, T.E., Tremblay, J., To, T., Cheng, J., Mosier, T., Kroemer, O., Fox, D., Birchfield, S., 2020. Camera-to-robot pose estimation from a single image. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 9426–9432.

Lepetit, V., Moreno-Noguer, F., Fua, P., 2009. EPnP: Efficient perspective-n-point camera pose estimation. Int. J. Comput. Vis. 81 (2), 155–166.

Li, X., Li, D., 2021. GPFS: a graph-based human pose forecasting system for smart home with online learning. ACM Trans. Sensor Netw. 17 (3), 1–19.

Luc, P., Couprie, C., Lecun, Y., Verbeek, J., 2018. Predicting future instance segmentation by forecasting convolutional features. In: Proceedings of the European Conference on Computer Vision. pp. 584–599.

Mangalam, K., Adeli, E., Lee, K.-H., Gaidon, A., Niebles, J.C., 2020. Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2784–2793.

Marchetti, F., Becattini, F., Seidenari, L., Del Bimbo, A., 2020. Multiple trajectory prediction of moving agents with memory augmented networks. IEEE Trans. Pattern Anal. Mach. Intell..

Martinez, J., Hossain, R., Romero, J., Little, J.J., 2017. A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Moon, G., Chang, J.Y., Lee, K.M., 2018. 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. pp. 5079–5088.

Noguchi, A., Iqbal, U., Tremblay, J., Harada, T., Gallo, O., 2022. Watch it move: Unsupervised discovery of 3D joints for re-posing of articulated objects. In: Proc. of the IEEE/CVF Conference on CVPR. pp. 3677–3687.

Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K., 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7025–7034.

Pavllo, D., Grangier, D., Auli, M., 2018. Quaternet: A quaternion-based recurrent model for human motion. arXiv preprint arXiv:1805.06485.

Peshkin, M.A., Colgate, J.E., Wannasuphoprasit, W., Moore, C.A., Gillespie, R.B., Akella, P., 2001. Cobot architecture. IEEE Trans. Robot. 17 (4), 377–390.

Plizzari, C., Cannici, M., Matteucci, M., 2021. Spatial temporal transformer network for skeleton-based action recognition. In: Pattern Recognition. pp. 694–701.

Salzmann, T., Pavone, M., Ryll, M., 2022. Motron: Multimodal probabilistic human motion forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6457–6466.

Sampieri, A., di Melendugno, G.M.D., Avogaro, A., Cunico, F., Setti, F., Skenderi, G., Cristani, M., Galasso, F., 2022. Pose forecasting in industrial human-robot collaboration. In: Proc. of the European Conference on Computer Vision. pp. 51–69.

Sarbolandi, H., Lefloch, D., Kolb, A., 2015. Kinect range sensing: Structured-light versus time-of-flight kinect. Comput. Vision Image Understand. 139, 1–20.

Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., et al., 2012. Efficient human pose estimation from single depth images. IEEE Trans. Pattern Anal. Mach. Intell. 35 (12), 2821–2840.

Simoni, A., Pini, S., Borghi, G., Vezzani, R., 2022. Semi-perspective decoupled heatmaps for 3D robot pose estimation from depth maps. IEEE Robot. Autom. Lett. 7 (4), 11569–11576.

Sofianos, T., Sampieri, A., Franco, L., Galasso, F., 2021. Space-time-separable graph convolutional network for pose forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11209–11218.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5693–5703.

Toyer, S., Cherian, A., Han, T., Gould, S., 2017. Human pose forecasting via deep markov models. In: Proceedings of the International Conference on Digital Image Computing: Techniques and Applications. IEEE, pp. 1–8.

Trivedi, N., Thatipelli, A., Sarvadevabhatla, R.K., 2021. 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. pp. 1–9.

Vondrick, C., Pirsiavash, H., Torralba, A., 2016. Anticipating visual representations from unlabeled video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 98–106.

Wang, K., Lin, L., Ren, C., Zhang, W., Sun, W., 2018. Convolutional memory blocks for depth data representation learning.. In: Proceeding of the International Joint Conferences on Artificial Intelligence. pp. 2790–2797.

Weiss, A., Buchner, R., Tscheligi, M., Fischer, H., 2011. Exploring human-robot cooperation possibilities for semiconductor manufacturing. In: Proceedings of the International Conference on Collaboration Technologies and Systems. IEEE, pp. 173–177.

Weiss, A., Wortmeier, A.-K., Kubicek, B., 2021. Cobots in industry 4.0: A roadmap for future practice studies on human–robot collaboration. IEEE Trans. Hum.-Mach. Syst. 51 (4), 335–345.

Widmaier, F., Kappler, D., Schaal, S., Bohg, J., 2016. Robot arm pose estimation by pixel-wise regression of joint angles. In: Proc. of the International Conference on Robotics and Automation. pp. 616–623.

Xiong, F., Zhang, B., Xiao, Y., Cao, Z., Yu, T., Zhou, J.T., Yuan, J., 2019. 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. pp. 793–802.

Yub Jung, H., Lee, S., Seok Heo, Y., Dong Yun, I., 2015. Random tree walk toward instantaneous 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2467–2474.

Zanuttigh, P., Marin, G., Dal Mutto, C., Dominio, F., Minto, L., Cortelazzo, G.M., 2016. Time-of-flight and structured light depth cameras. Technol. Appl. 978–3.

Zhang, Z., Hu, L., Deng, X., Xia, S., 2020. Weakly supervised adversarial learning for 3D human pose estimation from point clouds. IEEE Trans. Vis. Comput. Graphics 26 (5), 1851–1859.

Zhang, Z., Hu, L., Deng, X., Xia, S., 2021. Sequential 3D human pose estimation using adaptive point cloud sampling strategy.. In: Proceeding of the International Joint Conferences on Artificial Intelligence. pp. 1330–1337.

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M., 2023. Deep learning-based human pose estimation: A survey. ACM Comput. Surv. 56 (1), 1–37.

Zuo, Y., Qiu, W., Xie, L., Zhong, F., Wang, Y., Yuille, A.L., 2019. Craves: Controlling robotic arm with a vision-based economic system. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4214–4223.