Vision-Based Safe Human-Robot Collaboration with Uncertainty Guarantees

Jakob Thumm[1], Marian Frei[2], Tianle Ni[3], Matthias Althoff[4], and Marco Pavone[1]

***Abstract*— We propose a framework for vision-based human pose estimation and motion prediction that gives conformal prediction guarantees for certifiably safe human-robot collaboration. Our framework combines aleatoric uncertainty estimation with OOD detection for high probabilistic confidence. To integrate our pipeline in certifiable safety frameworks, we propose conformal prediction sets for human motion predictions with high, valid confidence. We evaluate our pipeline on recorded human motion data and a real-world human-robot collaboration setting.**

I. Introduction

Autonomous robots will become an integral part of our society, performing tedious and dangerous tasks in industry, households, and healthcare. To ensure safety in these human-centered environments, it is crucial to accurately perceive human poses, predict their motion, and control the robot to prevent critical collisions with the human. Hereby, we have to provide certifiable safety guarantees in all possible unseen situations.

Current safe human-robot collaboration (HRC) approaches can provably guarantee human safety if an accurate human pose measurement is available [1]–[4]. These approaches typically rely on a marker-based motion-tracking system for accurate pose estimation, which drastically limits their deployment potential. Previous works that estimate the human pose from more mobile sensor modalities, e.g., RGB-D cameras, usually assume a fixed maximal estimation error [5]–[9] or provide uncertainty estimates without conformal guarantees [10]–[18]. Most of these approaches can catastrophically fail if inputs are out of distribution (OOD), which is not captured in the predicted aleatoric uncertainty.

For predicting all possible states that the human can occupy in a given time interval, traditional safe HRC approaches [2]–[9] use simple motion models, e.g., that the human can move with up to $v_{\max} = 1.6\,\mathrm{m\,s^{-1}}$ in any direction as defined in ISO 13855:2010 [19]. These models tend to be highly conservative as they are not data-driven. State-of-the-art human motion prediction models [20]–[25] are less conservative but often lack (i) heteroscedastic aleatoric uncertainty estimates, (ii) end-to-end propagation of the pose estimation uncertainty, and (iii) conformal prediction guarantees.

To alleviate these shortcomings, we propose a framework for vision-based human pose estimation and motion prediction in Fig. 1. First, we estimate the two-dimensional (2D) human pose and its covariances in the two camera frames. Then, we perform an uncertainty-aware triangulation to retrieve the 3D pose with covariances. Given a history of human poses, we predict future 3D poses and their covariances. Based on the predicted covariances, we propose conformal prediction sets to over-approximate the uncertainty in the motion prediction with an $1-\epsilon$ confidence bound. To detect critical OOD inputs in the pose estimation and motion prediction, we use the gradient-based OOD detection method in [26]. By reusing past motion predictions as potential replacements for OOD inputs, we establish a smooth operation of our pipeline in unseen situations. The conformal prediction sets returned by our pipeline directly feed into the provably safe HRC approach SARA shield [4], which establishes certifiable safety guarantees at all times.

The main contribution of our work is a framework for safe HRC from vision-based inputs featuring

- 3D human motion prediction with end-to-end uncertainty propagation,
- conformal prediction sets for human poses, and
- a method to handle OOD inputs in a continuous prediction setting.

We evaluate our pipeline on the Human3.6M dataset [27] and in a real-world HRC setting with SARA shield [4]. In our experiments, we find that (i) our human pose estimation and motion prediction performs similarly to state-of-the-art models, (ii) the conformal prediction sets reduce the conservatism of model-based predictions by a factor of 11, and (iii) our proposed OOD handling reduces interruptions in real-world operation by $36.0\,\%$.

II. Preliminaries

For brevity, we use the colon notation $A : B$ to refer to the sequence $A, A+1, \dots, B$. We denote a point in the two-dimensional (2D) image space as $(u, v) \in \mathbb{R}^2$ and a point in the three-dimensional (3D) Euclidean space as $\boldsymbol{p} \in \mathbb{R}^3$. The human pose at timestep k is denoted by a tuple of 3D points $\mathcal{P}_k = (\boldsymbol{p}_k^1, \boldsymbol{p}_k^2, \dots, \boldsymbol{p}_k^J) \in \mathbb{R}^{J\times 3}$, and the motion of a human at the discrete timesteps $0, 1, \dots K$ is denoted by a sequence of poses $\mathcal{P}_{1:K} = (\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_K) \in \mathbb{R}^{K\times J\times 3}$. Finally, we refer to a sphere with center $\boldsymbol{p}_c$ and radius r as $\mathcal{B}(\boldsymbol{p}_c, r)$.

[1]Jakob Thumm and Marco Pavone are with the Department of Aeronautics and Astronautics, Stanford University. thumm@stanford.edu, pavone@stanford.edu

[2]Marian Frei is with the Chair of Imaging and Computer Vision, RWTH Aachen University. marian.frei@lfb.rwth-aachen.de

[3]Tianle Ni is with the School of Artificial Intelligence, Shanghai Jiao Tong University. tianle.ni@tum.de

[4]Matthias Althoff is with the Department of Computer Engineering, Technical University of Munich. althoff@tum.de

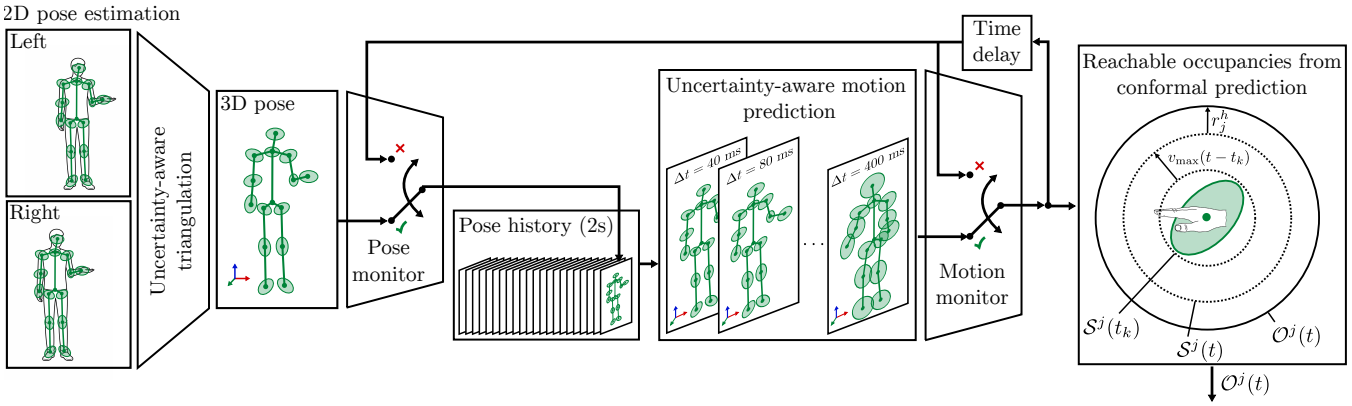

Fig. 1: Methodological overview of our pose estimation and motion prediction pipeline with conformal prediction sets.

A. Conformal Predictions

Let $\mathcal{Z} = (\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1:N}$ be a set of i.i.d. training tuples with inputs $\boldsymbol{x}_i \in \mathcal{X}$ and labels $\boldsymbol{y}_i \in \mathcal{Y}$, which is devided into a training set $\mathcal{Z}^{\text{train}}$ and a calibration set $\mathcal{Z}^{\text{cal}}$ with $\mathcal{Z}^{\text{train}} \cup \mathcal{Z}^{\text{cal}} = \mathcal{Z}$. Further, let a prediction model $f_\theta : \mathcal{X} \to \mathcal{Y}$ be trained on $\mathcal{Z}^{\text{train}}$. We denote a bounded set $\mathcal{S}_i \subseteq \mathcal{Y}$ as a conformal prediction set if $P(\boldsymbol{y}_i \in \mathcal{S}_i) \geq 1 - \epsilon$ with confidence level $1 - \epsilon$ [28, Sec. 2.3]. A non-conformity measure $A(\mathcal{Z}^{\text{train}}, z_i) \in \mathbb{R}$ is a function that scores how dissimilar a calibration sample z_i is from the training dataset [29].

B. Problem Statement

Given a history of N_I stereo image pairs $\{(\mathbf{I}_{k,1}, \mathbf{I}_{k,2})\}_{k=1:N_I}$ with $f_{\text{cam}} = \frac{1}{t_{k+1} - t_k}$ frames per second, our goal is to predict a sequence of N_P tuples of J conformal prediction sets $((\mathcal{S}_1^1, \mathcal{S}_1^2, \dots \mathcal{S}_1^J), \dots, (\mathcal{S}_{N_P}^1, \mathcal{S}_{N_P}^2, \dots \mathcal{S}_{N_P}^J))$ that include the future human joint positions with $1 - \epsilon$ confidence.

III. Human Pose Pipeline

To solve the problem in Sec. II-B, we propose the human pose pipeline depicted in Fig. 1.

A. Uncertainty-Aware 3D Pose Estimation

For our 3D pose estimation, we first adapt YOLO26 [30] to return the covariance matrix $\boldsymbol{C}_{i,\text{2D}}^j \in \mathbb{R}^{2\times 2}$ in addition to the 2D mean (u_i^j, v_i^j) of each human joint $j = 1, \dots, J$ in the two calibrated camera images $\mathbf{I}_1$ and $\mathbf{I}_2$. We extract the approximated covariance matrix from the scale of the Gaussian distribution returned by the regression model of the flow-based regression [11] used in YOLO26.

We then estimate the 3D joint positions from the two 2D poses by standard linear triangulation [31]. Following [32], we estimate the cross-covariance from residual correlations over the calibration set $\mathcal{Z}^{\text{cal}}$. We then obtain the 3D covariances through first-order covariance propagation [33] and add a small constant isotropic term to the covariances to account for systematic reconstruction errors, e.g., camera calibration imperfections.

B. Uncertainty-Aware 3D Motion Prediction

We refer to the tuple of covariance matrices of a human pose at timestep k as $\mathcal{C}_k = (\boldsymbol{C}_k^1, \boldsymbol{C}_k^2, \dots, \boldsymbol{C}_k^J) \in \mathbb{R}^{J\times 3\times 3}$. Given a sequence of past poses $\mathcal{P}_{1:K_I}$ and their covariances $\mathcal{C}_{1:K_I} = (\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_{K_I})$, we predict a sequence of future human 3D poses $\mathcal{P}_{K_I+1:K_I+K_P}$ and their heteroscedastic aleatoric covariance matrices $\mathcal{C}_{K_I+1:K_I+K_P}$.

1) Model Architecture: For our motion prediction model, we extend the discrete cosine transform (DCT) transformer model of [21] with uncertainty inputs and outputs. We first apply a DCT to the pose and covariance inputs $\mathcal{P}_{1:K_I}$ and $\mathcal{C}_{1:K_I}$ along the temporal axis to obtain a frequency representation whose coefficients are naturally ordered from low to high frequencies [20], [21]. We embed these coefficients into a higher-dimensional feature space and add learnable positional encodings to retain temporal information. The embedded uncertainty is scaled by a learnable factor and added to the pose features before the transformer blocks. A key architectural choice in our model is to process low- and high-frequency components separately, using dedicated multi-layer perceptron paths after the attention module. Low frequencies capture smooth global trends, whereas high frequencies capture rapid, fine-scale motion variations. After separate processing, both streams are recombined and mapped back to pose space using the inverse DCT. To predict $\mathcal{C}_{K_I+1:K_I+K_P}$, we follow the Cholesky factorization approach in [34], which yields stable training and avoids invalid covariance matrices.

2) Training Objective: We train the uncertainty head using a multivariate Gaussian negative log-likelihood (NLL) loss [34]

$$\mathcal{L}_{\text{NLL}} = \frac{1}{JK_P} \sum_{k=K_I+1}^{K_I+K_P} \sum_{j=1}^{J} \frac{1}{2} \log |\boldsymbol{C}_k^j| + \frac{1}{2} \boldsymbol{d}_k^{j\top} \boldsymbol{C}_k^{j-1} \boldsymbol{d}_k^j, \quad (1)$$

with residual $\boldsymbol{d}_k^j = \boldsymbol{p}_k^j - \hat{\boldsymbol{p}}_k^j$. To stabilize the mean prediction quality, we add the ℓ_1 pose loss

$$\mathcal{L}_{\text{pose}} = \frac{1}{JK_P} \sum_{k=K_I+1}^{K_I+K_P} \sum_{j=1}^{J} \|\boldsymbol{d}_k^j\|_1 \quad (2)$$

to the total loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NLL}} + \lambda \mathcal{L}_{\text{pose}}$.

3) Stable Multi-Phase Optimization: Direct end-to-end training with covariance prediction can be unstable. Therefore, we follow a staged procedure: (i) train a deterministic base predictor using $\mathcal{L}_{\text{pose}}$, (ii) add the uncertainty head and train it first with the backbone frozen (decoupled learning), and (iii) fine-tune the full model end-to-end with a gradually reduced λ to balance accuracy and calibration.

C. Conformal Prediction Sets

To construct the conformal prediction sets from our uncertainty estimates, we adapt the procedure in [28, Sec. 2.3]. Let a training tuple for the motion model be $\boldsymbol{z}_i = ((\mathcal{P}_{1:K_I}, \mathcal{C}_{1:K_I})_i, (\mathcal{P}_{K_I+1:K_I+K_P}, \mathcal{C}_{K_I+1:K_I+K_P})_i)$. We define the non-conformity measure for timestep k and joint j as

$$A_k^j(\boldsymbol{z}_i) = \frac{\|\boldsymbol{d}_{k,i}^j\|_2}{\sqrt{\lambda_{\max}(\boldsymbol{C}_{k,i}^j)}}, \tag{3}$$

where $\lambda_{\max}(\boldsymbol{C}_{k,i}^j)$ is the maximal eigenvalue of $\boldsymbol{C}_{k,i}^j$. Let α_k^j be the $1-\epsilon$ percentile of the scores $\alpha_{k,i}^j = A_k^j(\boldsymbol{z}_i)$ for all $\boldsymbol{z}_i \in \mathcal{Z}^{\text{cal}}$ such that $P(\alpha_{k,i}^j \leq \alpha_k^j) \geq 1-\epsilon$ [28, Sec. 2.3, Step 3 and 4].

Proposition 3.1: Given a predicted position $\hat{\boldsymbol{p}}_k^j$ and covariance $\boldsymbol{C}_k^j$, the sphere $\mathcal{S}^j(t_k) = \mathcal{B}\left(\hat{\boldsymbol{p}}_k^j, \alpha_k^j \sqrt{\lambda_{\max}(\boldsymbol{C}_k^j)}\right)$ is a conformal prediction set for joint j at time t_k.

Proof: The proof follows [28, Theo. 1]. ∎

Our sets only provide conformal predictions at predefined timesteps. To retrieve the conformal predictions at any time $t \geq t_{K_I}$, we find the maximum time t_k for all $k = K_I+1 : K_I+K_P$ for which $t_k \leq t$, set $\Delta t = t - t_k$, and extend the radius of $\mathcal{S}^j(t_k)$ by $\Delta t v_{\max}$ with $v_{\max} = 1.6\,\mathrm{m\,s^{-1}}$ as defined in [19]. To include the entire human body in the sets, we can use the SARA tool [35] to determine the full-body reachable occupancies $\mathcal{O}^j(t)$ from the conformal prediction sets at a given time point.

D. OOD Detection

Our conformal prediction in Sec. III-C assumes that samples are drawn from the training distribution, which may not hold when deploying robots in new environments. Therefore, we deploy the sketching Lanczos uncertainty (SLU) OOD detection method of [26] to detect OOD inputs in the pose estimation SLU_{2D} and motion prediction SLU_{mot}. We calibrate the two OOD thresholds τ_{2D} and τ_{mot} on the calibration datasets, so that $P(\text{SLU}(\boldsymbol{z}_i) \leq \tau) \geq 1 - \epsilon_{\text{OOD}}$. Since the SLU computation time scales linearly with the number of output parameters, we use reduced models that only predict the positions of the human head and hand for the OOD detection. For the motion prediction network, we further reduce the model output to only predict the timestep $K_I + K_P/2$. Lastly, we treat a missing human in the frame as OOD and leave the handling of humans entering and leaving the workspace to future work.

Algorithm 1 Uncertainty-Aware Human Pose Pipeline

Require: Stereo camera stream $\{(\mathbf{I}_{k,1}, \mathbf{I}_{k,2})\}_{k\geq 1}$, f_{2D}, f_{mot}, SLU_{2D}, SLU_{mot}, τ_{2D}, τ_{mot}, K_I, K_P, N_{req}

1: Initialize pose buffer $\mathcal{H} \leftarrow \varnothing$, pose validity buffer $\mathbf{v} \leftarrow \mathbf{0}_{K_I}$, and motion buffer $\mathcal{M}$ to NaN
2: **while** True **do**
3: Acquire stereo frames $(\mathbf{I}_{k,1}, \mathbf{I}_{k,2})$
4: Estimate 2D pose and covariances ▷ Sec. III-A
5: Compute 3D pose $\mathcal{P}_k$ and covariances $\mathcal{C}_k$ via stereo triangulation ▷ Sec. III-A
6: Compute $s_{\text{2D}} \leftarrow \text{SLU}_{\text{2D}}(\mathbf{I}_{k,1})$
7: **if** $s_{\text{2D}} \leq \tau_{\text{2D}}$ **then**
8: Append $(\mathcal{P}_k, \mathcal{C}_k)$ to $\mathcal{H}$; set $v_k \leftarrow 1$
9: **else**
10: Append $\mathcal{M}[0]$ to $\mathcal{H}$; set $v_k \leftarrow 0$
11: **end if**
12: **if** $|\mathcal{H}| < K_I$ **then continue**
13: **end if** ▷ Wait until pose buffer is full
14: Predict $(\mathcal{P}_{K_I+1:K_I+K_P}, \mathcal{C}_{K_I+1:K_I+K_P}) \leftarrow f_{\text{mot}}(\mathcal{H})$ ▷ Sec. III-B
15: Compute $s_{\text{mot}} \leftarrow \text{SLU}_{\text{mot}}(\mathcal{H})$
16: **if** $s_{\text{mot}} \leq \tau_{\text{mot}}$ **and** $\sum_{i=K_I-N_{\text{req}}+1}^{K_I} v_i = N_{\text{req}}$ **then**
17: $\mathcal{M} \leftarrow (\mathcal{P}_{K_I+1:K_I+K_P}, \mathcal{C}_{K_I+1:K_I+K_P})$ ▷ If input is ID and last N_{req} poses came from images, accept new prediction
18: **else**
19: $\mathcal{M}[0 : K_P-1] \leftarrow \mathcal{M}[1 : K_P]$; $\mathcal{M}[K_P] \leftarrow$ NaN ▷ Continue on prior prediction
20: **end if**
21: $\mathcal{O}_{1:K_P} \leftarrow \text{ConformalSets}(\mathcal{M})$ ▷ Sec. III-C
22: Publish reachable occupancies $\mathcal{O}_{1:K_P}$
23: **end while**

E. Handling OOD Events

As our motion prediction requires K_I valid input poses, our pipeline would not return any prediction in the timeframe $K_I f_{\text{cam}}$ after an invalid 2D pose. With, e.g., $K_I = 50$ and $\epsilon_{\text{OOD}} = 95\,\%$, this would induce a failed motion prediction in $92.3\,\%$ of timesteps. To circumvent this, we propose an algorithm that reuses previous motion predictions for OOD pose estimation.

We describe the total process of our human pose pipeline in Algorithm 1. At each timestep, SLU_{2D} scores the current camera image against the training distribution of f_{2D} (line 6). If the 2D OOD score of the first image is below the calibrated threshold τ_{2D}, the estimated pose $\mathcal{P}_k$ and covariance $\mathcal{C}_k$ are appended to the pose buffer $\mathcal{H}$ and the corresponding validity flag is set to $v_k = 1$ (line 8). If the pose is classified as OOD, we discard the estimated 3D pose and replace it by the first entry of the current motion prediction buffer $\mathcal{M}[0]$ and set the validity flag to $v_k = 0$ (line 10). This ensures the pose buffer contains a temporally consistent sequence of plausible poses during periods where the pose estimator is unreliable.

Once the pose buffer contains K_I entries (line 13), we start predicting human motions using f_{mot} (line 14) and the

motion OOD score with $\mathrm{SLU}_{\mathrm{mot}}$ (line 15). The motion buffer $\mathcal{M}$ is updated with the new prediction if (i) the motion input is in-distribution ($s_{\mathrm{mot}} \leq \tau_{\mathrm{mot}}$), and (ii) the last N_{req} entries of the pose buffer all originate from valid image observations (lines 16–17). The second condition prevents a continuous feedback loop in the motion prediction if the 2D pose input is OOD. If either condition fails, the motion buffer is shifted by one timestep to the left and the vacated slot at the end of the buffer is marked as invalid (line 19). This design ensures the pipeline degrades gracefully under OOD conditions and resumes normal operation only after N_{req} consecutive valid image-based poses have been observed.

IV. Experiments

In this section, we evaluate the performance of our pose pipeline on real-world data and real-world robot deployment.

A. Prediction Accuracy

We evaluate our motion prediction accuracy on the Human3.6M (H36M) benchmark [27], following the standard protocol used in prior work [20], [36], where we use subjects S1, S6, S7, S8, and S9 for training, S11 for validation, and S5 for testing. We report all results for $K_I = 50$ frames, $K_P = 10$ frames, $f_{\mathrm{cam}} = 25\,\mathrm{fps}$, $\epsilon_{\mathrm{OOD}} = 95\,\%$, and $J = 13$ joints on the test set. As the primary metric, we use the mean per joint position error (MPJPE) [20].

We compare against the common baselines HisRep [21], ST-DGCN [22], ST-Trans [24], and SiMLPe [36] in Table I. For a fair comparison, we evaluate the model performance on ground-truth 3D pose inputs. After training on ground-truth poses (stage 1), our model outperforms state-of-the-art models. However, after training on the estimated 3D poses and adding uncertainty prediction (final), our model performs slightly worse on the original task.

TABLE I: Motion prediction test results on H36M.

	↓ MPJPE (mm)			
Method	**80 ms**	**160 ms**	**320 ms**	**400 ms**
Repeating Last-Frame [36]	23.8	44.4	76.1	88.2
One FC [36]	14.0	33.2	68.0	81.5
HisRep [21]	10.4	22.6	47.1	58.3
ST-DGCN [22]	10.3	22.7	47.4	58.5
ST-Trans [24]	10.4	23.4	48.4	59.2
SiMLPe [36]	9.6	21.7	46.3	57.3
Ours (stage 1)	**8.7**	**16.7**	**41.1**	**54.5**
Ours (final)	18.4	28.1	53.1	67.2

B. Conformal Prediction Set Evaluation

We compare the validity of our conformal prediction sets against the constant velocity model of ISO 13855:2010 [19], which assumes a maximal velocity of $v_{\mathrm{max}} = 1.6\,\mathrm{m\,s^{-1}}$ for all human joints in any direction. We calibrate the conformal prediction sets on the H36M validation data using a confidence level of $1-\epsilon = 99\,\%$. We report the percentage of ground-truth joint positions within the predicted sets and the average set volume in Table II. In these experiments, our conformal prediction sets achieve a higher coverage ($98.25\,\%$) than ISO 13855:2010 ($97.93\,\%$) while reducing the mean set volume by a factor of 11 compared to ISO 13855:2010. The coverage is slightly lower than our calibration confidence, which indicates that the test data includes faster movements than the calibration dataset. Note that the assumption of $v_{\mathrm{max}} = 1.6\,\mathrm{m\,s^{-1}}$ defined in ISO 13855:2010 did not hold in our experiments.

TABLE II: Conformal prediction set test results on H36M.

Method	↑ Coverage (%)	↓ Volume (m^3)
ISO 13855:2010 [19]	97.93	0.191
Conformal prediction sets (ours)	**98.25**	**0.017**

C. Full Pipeline Evaluation

We evaluate the efficacy of our OOD handling mechanism described in Algorithm 1 by executing our full pipeline on the H36M test data with varying N_{req} values. Here, $N_{\mathrm{req}} = K_I = 50$ indicates that any OOD input in the 2D pose estimation would result in an invalid motion prediction, and $N_{\mathrm{req}} = 3$ is the recommended value used in our real-world experiments. Our results in Table III show that our OOD pipeline reduces the rate of invalid pose buffers $\sum_{i=K_I-N_{\mathrm{req}}+1}^{K_I} v_i < N_{\mathrm{req}}$ by $36.0\,\%$ while only increasing the average MPJPE by $2.6\,\%$. Therefore, our OOD handling significantly increases the rate of valid motion predictions while maintaining a high prediction accuracy.

TABLE III: Full pipeline evaluation results on H36M.

N_{req}	↓ $\mathcal{H}$ invalid [%]	↑ Motion valid [%]	↓ MPJPE [mm]
3 (ours)	**9.45**	**85.48**	53.56
10	11.72	82.10	53.13
50	14.75	75.29	**52.22**

D. Real-World Deployment

We integrated our human pose pipeline in SARA shield [3], [4] and tested it in a real-world HRC setting on a Franka Emika robot. In our real-world deployment, we use the Intel RealSense 435i camera for human perception and retrieve the depth information directly from its output. Hereby, we mark all 3D poses whose depth differs by more than $0.8\,\mathrm{m}$ from the median depth as OOD. A video of the deployment is available at `https://youtu.be/oeN8RgwpzhE`. In the speed and separation monitoring mode of SARA shield, the robot always came to a complete stop before the human operator could reach the robot.

V. Conclusion

We presented a vision-based framework for safe HRC with end-to-end uncertainty propagation, conformal prediction sets, and graceful OOD handling. Our conformal prediction sets significantly reduce the prediction volume over ISO 13855, while our OOD handling effectively reduces invalid motion predictions. Future work includes fusion with diverse sensor modalities, reasoning over critical safety situations, and a 3D pose estimation from RGB-D inputs.

VI. Acknowledgements

This work was supported by Toyota Motor Engineering & Manufacturing North America, Inc. and Volkswagen Group of America, Inc.

References

[1] S. Haddadin, S. Haddadin, A. Khoury, T. Rokahr, S. Parusel, R. Burgkart, A. Bicchi, and A. Albu-Schäffer, "On making robots understand safety: Embedding injury knowledge into control," *The International Journal of Robotics Research*, vol. 31, no. 13, pp. 1578–1602, 2012.

[2] D. Beckert, A. Pereira, and M. Althoff, "Online verification of multiple safety criteria for a robot trajectory," in *Proc. of the IEEE Conf. on Decision and Control (CDC)*, 2017, pp. 6454–6461.

[3] J. Thumm and M. Althoff, "Provably safe deep reinforcement learning for robotic manipulation in human environments," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2022, pp. 6344–6350.

[4] J. Thumm, J. Balletshofer, L. Maglanoc, L. Muschal, and M. Althoff, "A general safety framework for autonomous manipulation in human environments," *Accepted for Publication in the IEEE Transactions on Robotics*, 2026.

[5] A. M. Zanchettin, N. M. Ceriani, P. Rocco, H. Ding, and B. Matthias, "Safety in human-robot collaborative manufacturing environments: Metrics and control," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 882–893, 2016.

[6] M. J. Rosenstrauch, T. J. Pannen, and J. Krüger, "Human robot collaboration - using kinect V2 for ISO/ts 15066 speed and separation monitoring," *Procedia CIRP*, vol. 76, pp. 183–186, 2018.

[7] S. Kumar, S. Arora, and F. Sahin, "Speed and separation monitoring using on-robot time-of-flight laser-ranging sensor arrays," in *IEEE Int. Conf. on Automation Science and Engineering (CASE)*, 2019, pp. 1684–1691.

[8] B. Lacevic, A. R. S. E. M. Newishy, A. M. Zanchettin, and P. Rocco, "Enhanced performance of human-robot collaboration using braking surfaces and trajectory scaling," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2023, pp. 5942–5949.

[9] B. Lacevic, A. M. Zanchettin, and P. Rocco, "Safe human-robot collaboration via collision checking and explicit representation of danger zones," *IEEE Transactions on Automation Science and Engineering*, vol. 20, no. 2, pp. 846–861, 2023.

[10] N. B. Gundavarapu, D. Srivastava, R. Mitra, A. Sharma, and A. Jain, "Structured aleatoric uncertainty in human pose estimation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 50–53.

[11] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, and C. Lu, "Human pose regression with residual log-likelihood estimation," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 2021, pp. 11 025–11 034.

[12] L. Bramlage, M. Karg, and C. Curio, "Plausible uncertainties for human pose regression," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 2023, pp. 15 133–15 142.

[13] S. Schaefer, D. F. Henning, and S. Leutenegger, "Glopro: Globally-consistent uncertainty-aware 3D human pose estimation & tracking in the wild," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2023, pp. 3803–3810.

[14] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose++: Vision Transformer for generic body pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 1212–1230, 2024.

[15] Y. Ying, X. Huang, and W. Dong, "Multi-view active sensing for human–robot interaction via hierarchically connected tree," *Sensors and Actuators A: Physical*, vol. 378, 2024.

[16] T. Maeda, K. Takeshita, N. Ukita, and K. Tanaka, "Multimodal active measurement for human mesh recovery in close proximity," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9970–9977, 2024.

[17] V. Davoodnia, S. Ghorbani, M.-A. Carbonneau, A. Messier, and A. Etemad, "Upose3d: Uncertainty-aware 3D human pose estimation with cross-view and temporal cues," in *Proc. of the European Conf. on Computer Vision (ECCV)*, 2024, pp. 19–38.

[18] R. Khirodkar, T. Bagautdinov, J. Martinez, S. Zhaoen, A. James, P. Selednik, S. Anderson, and S. Saito, "Sapiens: Foundation for human vision models," in *Computer Vision – ECCV 2024*, 2025, pp. 206–228.

[19] ISO, "Safety of machinery - positioning of safeguards with respect to the approach speeds of parts of the human body," International Organization for Standardization, Tech. Rep. DIN EN ISO 13855:2010-10 ST N, 2010.

[20] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 9489–9497.

[21] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020, pp. 474–489.

[22] T. Ma, Y. Nie, C. Long, Q. Zhang, and G. Li, "Progressively generating better initial guesses towards next stages for high-quality human motion prediction," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6437–6446.

[23] Y. Zhang, K. Ding, J. Hui, S. Liu, W. Guo, and L. Wang, "Skeleton-rgb integrated highly similar human action prediction in human–robot collaborative assembly," *Robotics and Computer-Integrated Manufacturing*, vol. 86, 2024.

[24] S. Saadatnejad, M. Mirmohammadi, M. Daghyani, P. Saremi, Y. Z. Benisi, A. Alimohammadi, Z. Tehraninasab, T. Mordan, and A. Alahi, "Toward reliable human pose forecasting with uncertainty," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4447–4454, 2024.

[25] K. A. Eltouny, W. Liu, S. Tian, M. Zheng, and X. Liang, "De-tgn: Uncertainty-aware human motion forecasting using deep ensembles," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2192–2199, 2024.

[26] M. Miani, L. Beretta, and S. Hauberg, "Sketched lanczos uncertainty score: A low-memory summary of the fisher information," in *Proc. of the Int. Conf. on Neural Information Processing Systems (NeurIPS)*, 2024.

[27] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[28] S. Messoudi, S. Destercke, and S. Rousseau, "Ellipsoidal conformal inference for multi-target regression," in *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, 2022, pp. 294–306.

[29] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer International Publishing, 2022.

[30] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee, "Yolo26: Key architectural enhancements and performance benchmarking for real-time object detection," 2026. [Online]. Available: http://arxiv.org/abs/2509.25164

[31] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[32] N. J. Bryan, P. Smaragdis, and G. J. Mysore, "Clustering and synchronizing multi-camera video via landmark cross-correlation," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2389–2392.

[33] A. Lewandowski, D. F. Williams, P. D. Hale, J. C. Wang, and A. Dienstfrey, "Covariance-based vector-network-analyzer uncertainty analysis for time-and frequency-domain measurements," *IEEE Transactions on Microwave Theory and Techniques*, vol. 58, no. 7, pp. 1877–1886, 2010.

[34] R. L. Russell and C. Reale, "Multivariate uncertainty in deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7937–7943, 2021.

[35] S. Schepp, J. Thumm, S. B. Liu, and M. Althoff, "SaRA: A tool for safe human–robot coexistence and collaboration through reachability analysis," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2022, pp. 4312–4317.

[36] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda-Pineda, and F. Moreno-Noguer, "Back to mlp: A simple baseline for human motion prediction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4809–4819.