IROS KYOTO 2022
IEEE/RSJ International Conference on Intelligent Robots and Systems

Semi-Perspective Decoupled Heatmaps for 3D Robot Pose Estimation from Depth Maps

Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani

alessandro.simoni@unimore.it, s.pini@unimore.it, guido.borghi@unibo.it, roberto.vezzani@unimore.it

University of Modena and Reggio Emilia, Italy

UNIMORE AImage Lab AIRI AI Research & Innovation Center

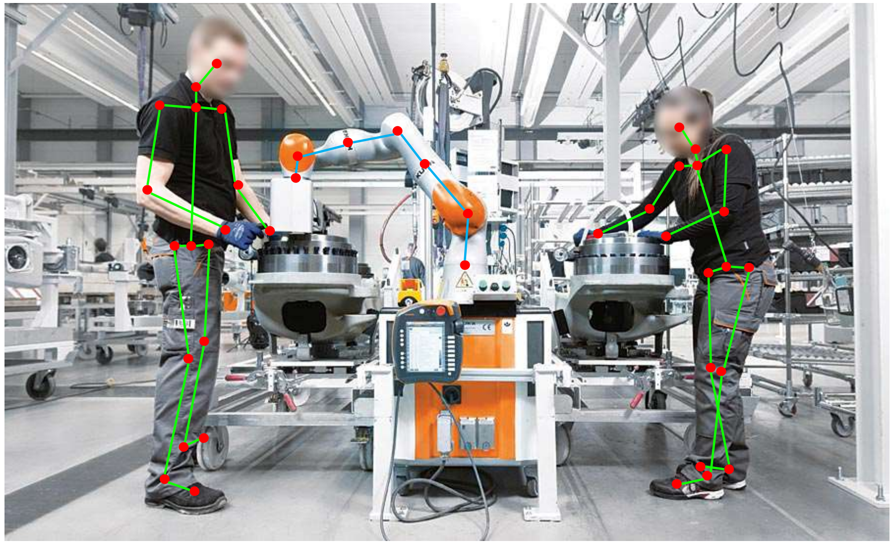
Why RPE?

SHARED WORKSPACE

ROBOTS + HUMANS

→ SURVEILLANCE SYSTEM WITH EXTERNAL CAMERAS

3D POSE OF HUMANS AND ROBOTS



- Analysis of interactions
- Anomaly detection
- Trajectory prediction

Our setting

ROBOTICS

NO ACCESS TO ENCODER DATA

DISABLED OR REVOKED BY THIRD PARTIES

COMPUTER GRAPHICS & SIMULATORS

TRAINING DATA GENERATION

SYNTHETIC TO REAL

TRAINING ON SYNTH AND TEST ON REAL

RGB-D OR DEPTH ONLY CAMERA DEVICES

PRECISE 3D SCENE INFORMATION

Our approach – Data Acquisition

Data Acquisition

Depth device

Collaborative robot

SimBa Dataset

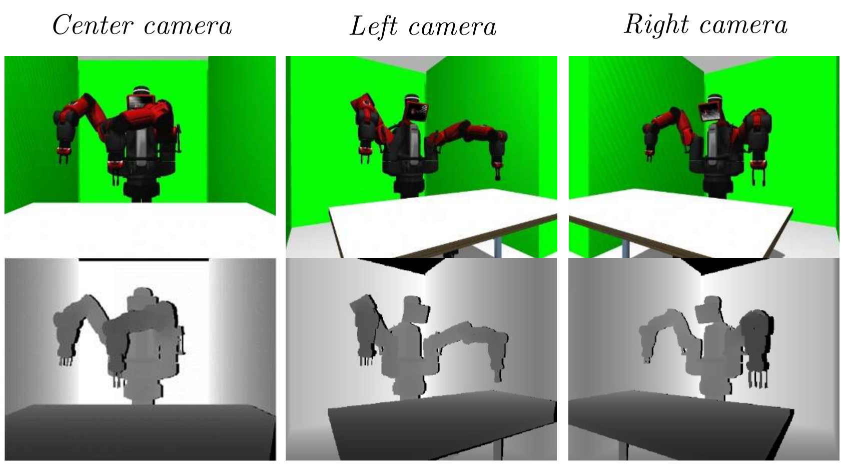
Rethink Baxter

SimBa Dataset

SYNTHETIC

ROS + Gazebo

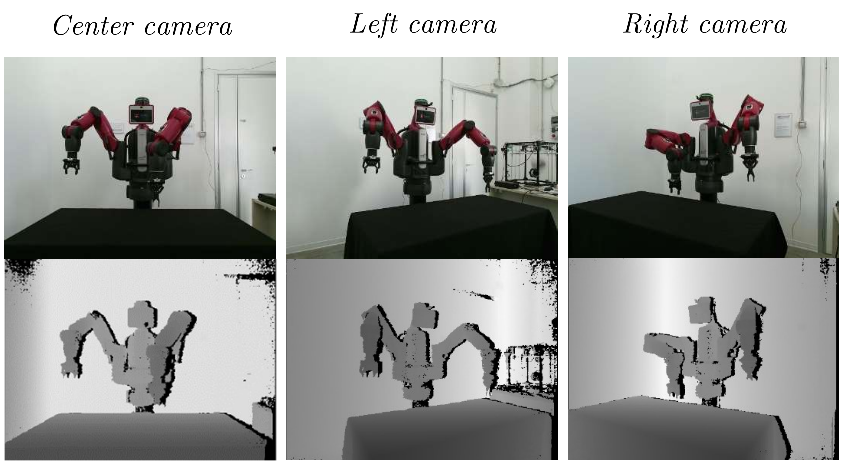
- Over 350k RGB-D images
- Pick-n-place locations
- 16 robot joints
- Camera positions



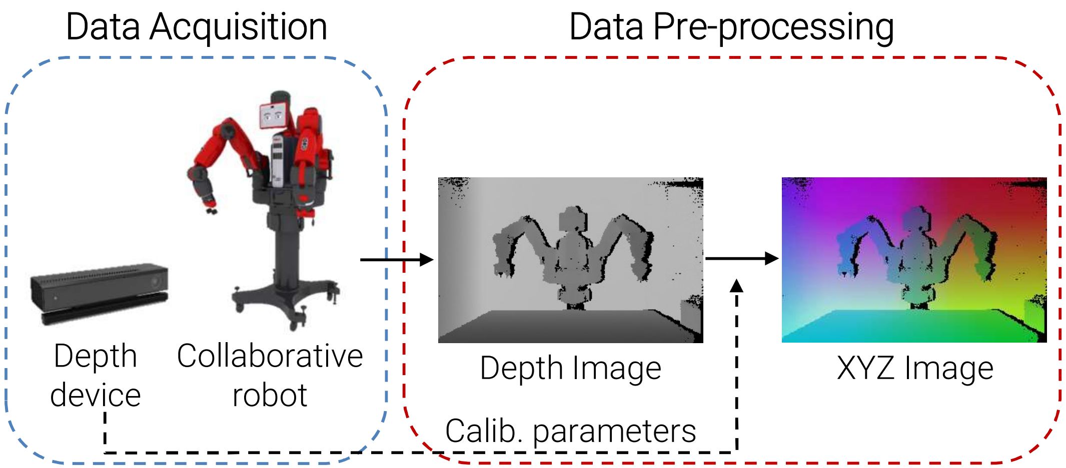
REAL

ROS + Microsoft Kinect One

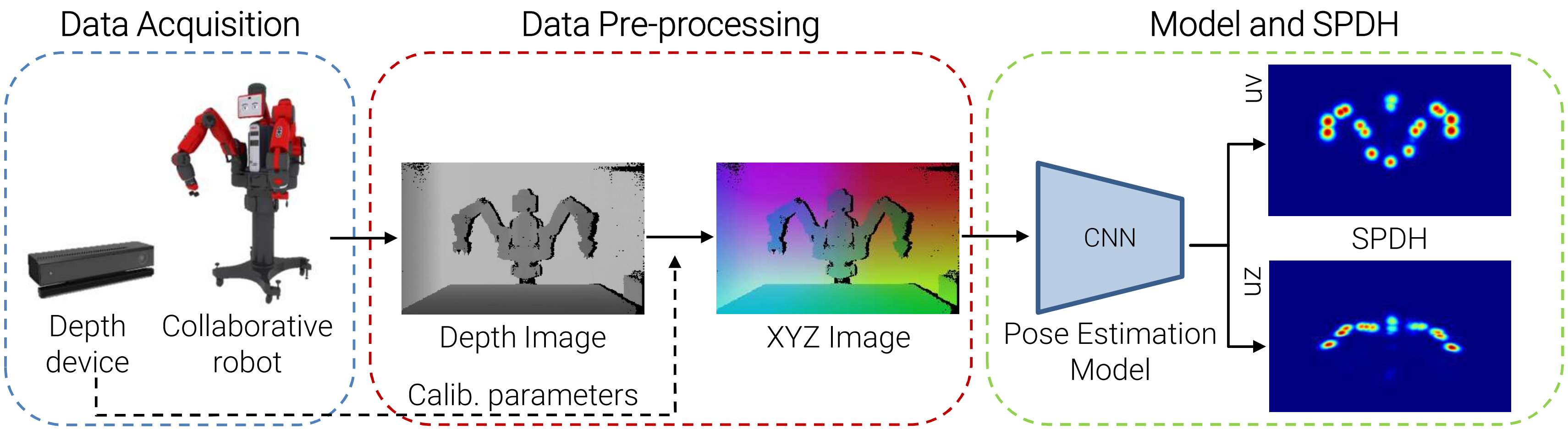
- Over 20k RGB-D images
- Camera positions
- 16 robot joints
- 20 pick-n-place sequences



Our approach – Data Pre-processing



Our approach - SPDH



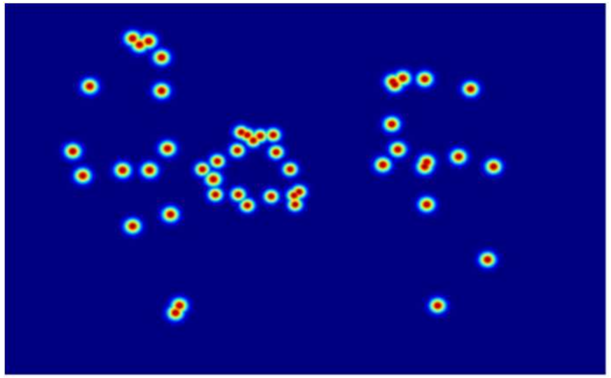
Interpretability

Drawn inspiration from Human Pose Estimation domain

↓

HEATMAPS INTERPRETABILITY

2D Pose Estimation
CNN Model

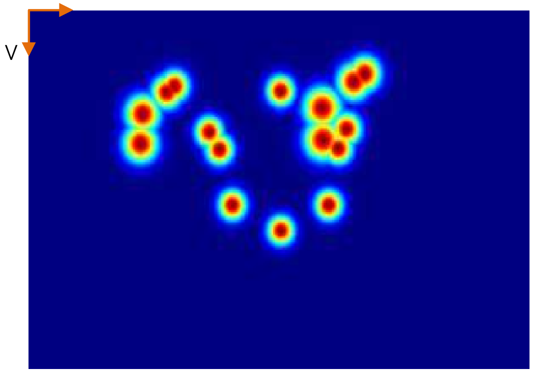


Idea

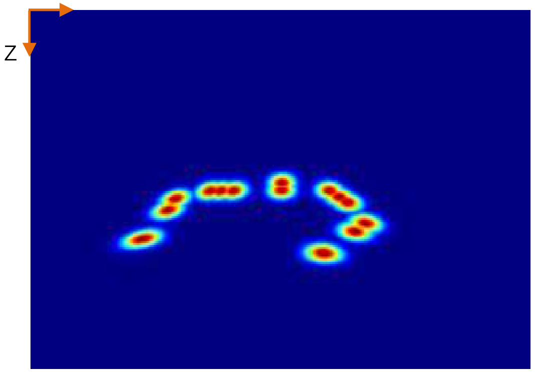
Find alternative representation for 3D pose of articulated objects

SEMI-PERSPECTIVE DECOUPLED HEATMAPS

UV SPACE

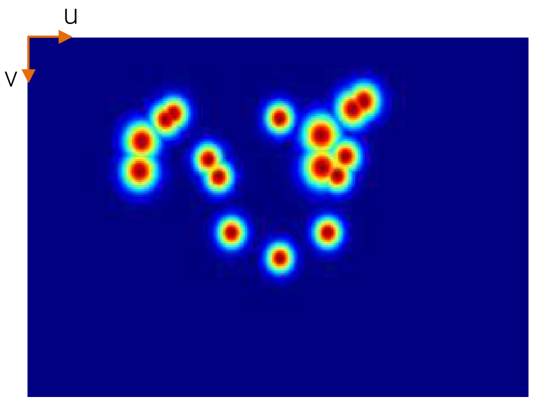


UZ SPACE



SPDH Computation

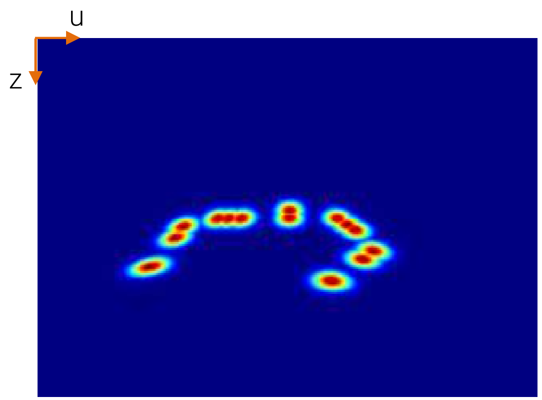
UV SPACE



Fixed sigma

$\sigma^m = 50$

UZ SPACE



$$\sigma_j = \frac{\sigma^m \cdot f}{Z_j}$$

near joints = bigger σ
far joints = smaller σ

$$\mathcal{H}_j^{uv}(p) = \mathcal{N}(p - p_j, \sigma_j)$$
$$= \frac{1}{2\pi\sigma_j} e^{-[(p^x - p_j^x)^2 + (p^y - p_j^y)^2]/(2\sigma_j^2)}$$

$$\bar{Z} = \{\bar{Z}_i \in Z;\ \bar{Z}_{min} \leq \bar{Z}_i \leq \bar{Z}_{max}\}$$

subspace of Z

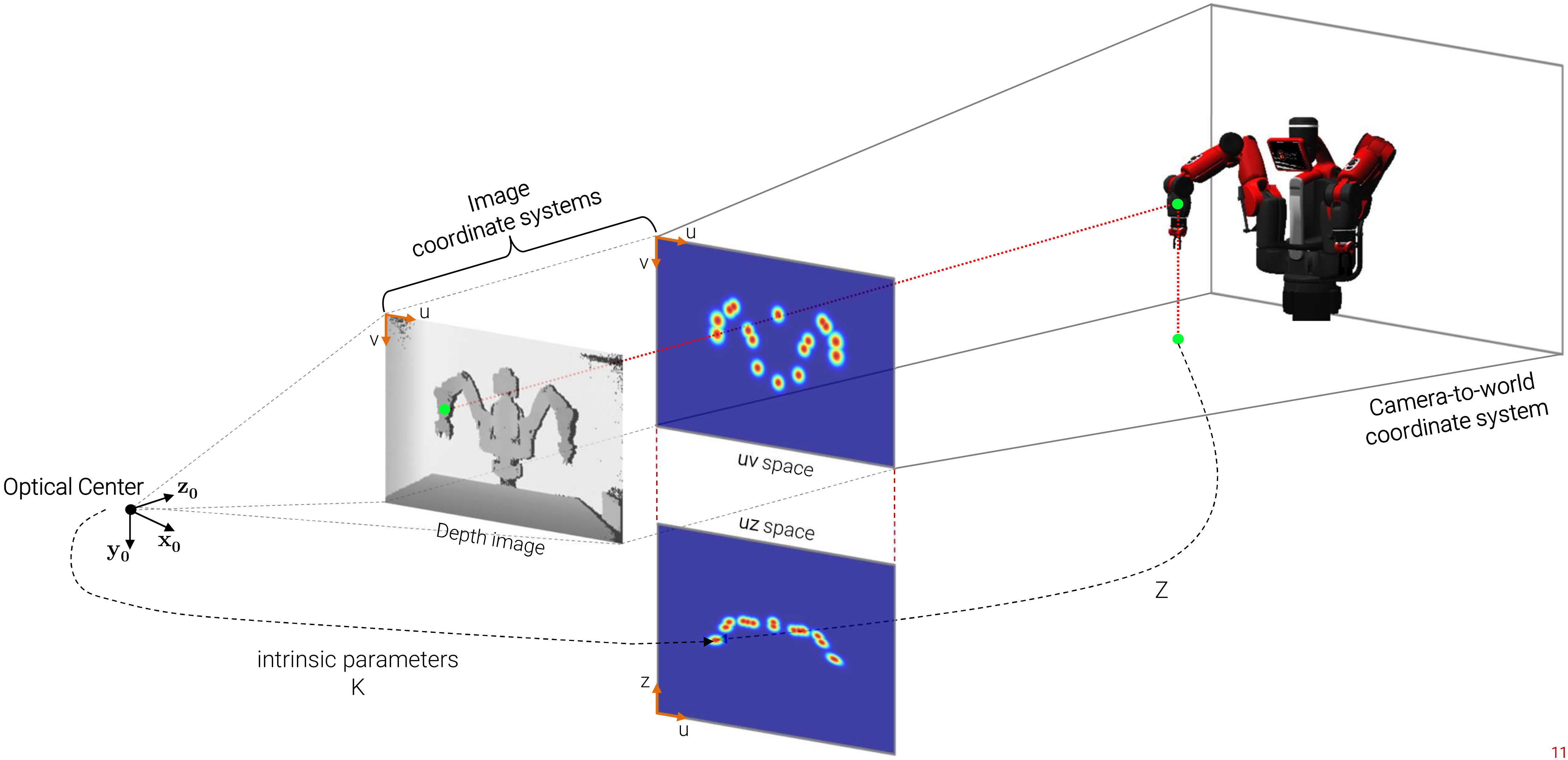
$$z = \frac{\bar{Z}_{max} - \bar{Z}_{min}}{\Delta Z}$$

subspace discretization (1 pixel = ΔZ)

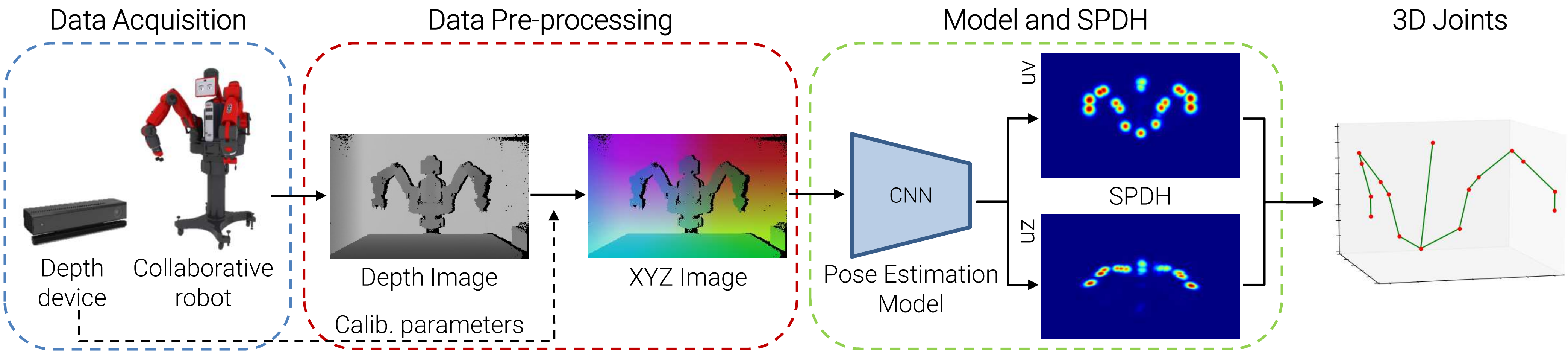
$$P(p) = \left((p^x - c) \cdot \frac{\bar{Z}^y}{f},\ \bar{Z}^y\right)$$

$$\mathcal{H}_j^{uz}(p) = \mathcal{N}(P(p) - P_j, \sigma^m)$$
$$= \frac{1}{2\pi\sigma^m} e^{-d(p)/(2\sigma^{m\,2})}$$

SPDH Visualization

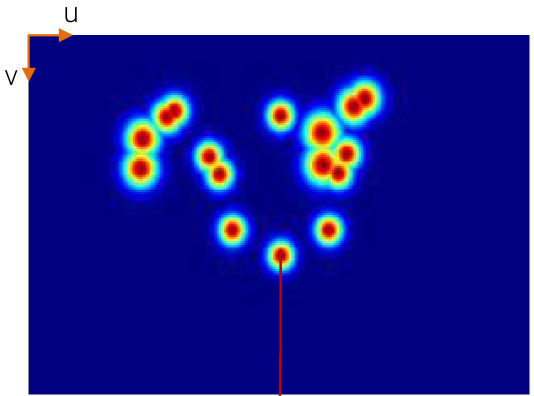


Our approach – 3D Joints

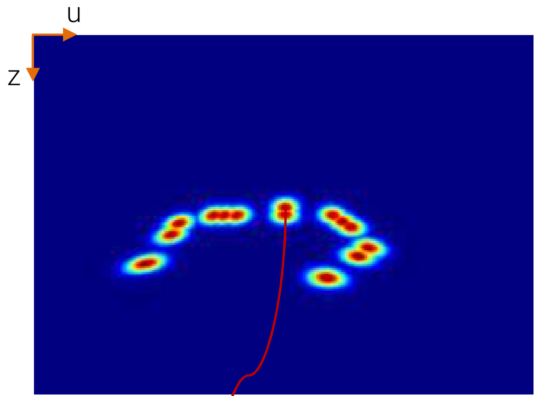


From SPDH to 3D pose

UV SPACE



UZ SPACE



uv coordinates of the gaussian peak

$$\hat{p}_j$$

z coordinate of the gaussian peak

$$\hat{Z}_j = (z \cdot \Delta Z) + \bar{Z}_{\min}$$

$$\boxed{\hat{P}_j = (K^{-1} \cdot \hat{p}_j) \cdot \hat{Z}_j}$$

Quantitative Results

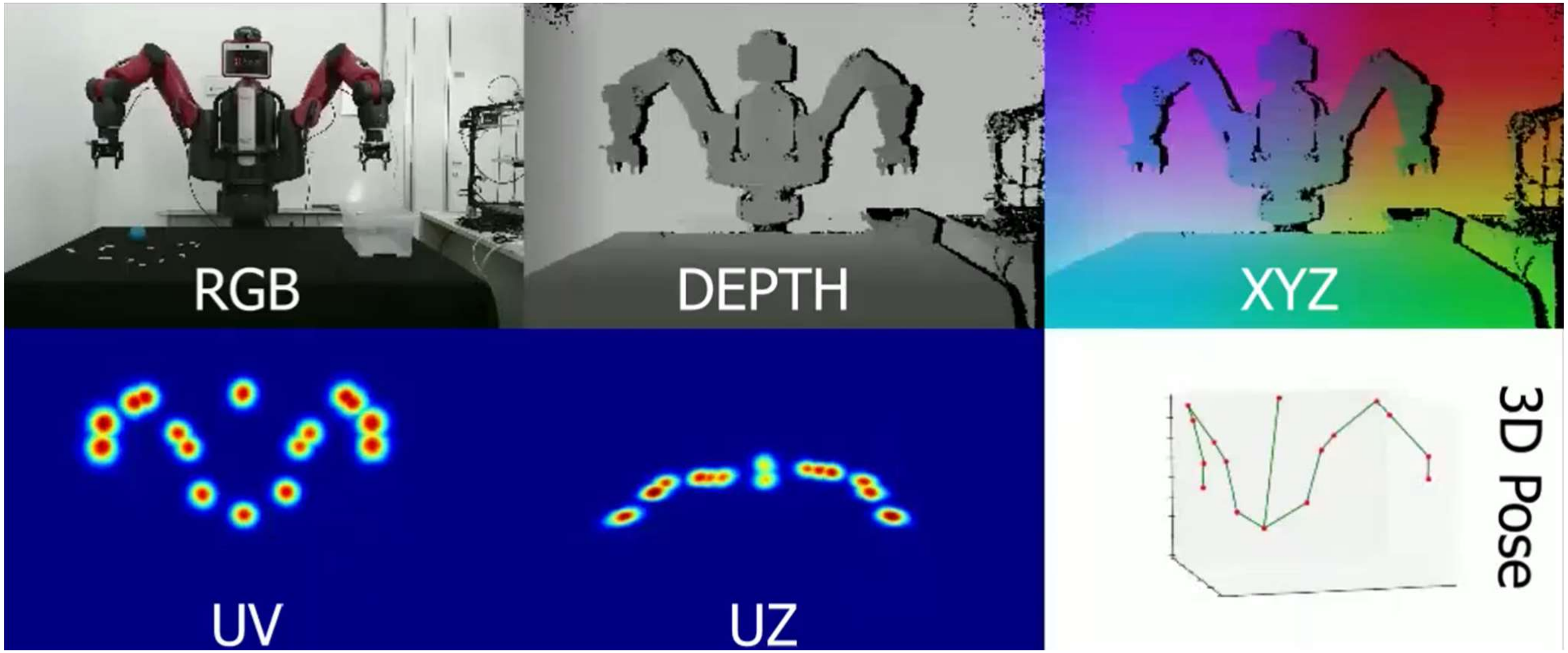
		mAP (%) ↑				**ADD** (cm) ↓	
Approach	**Network**	**40**mm	**60**mm	**80**mm	**100**mm	$L1$	$L2$
2D to 3D (depth)	Stacked Hourglass (1 HG) [1]	8.98	31.21	49.12	66.11	15.63 ±6.62	11.59 ±5.32
2D to 3D (depth)	Stacked Hourglass (2 HG) [1]	10.13	31.94	50.54	67.14	14.88 ±6.10	11.06 ±5.04
2D to 3D (depth)	FPM (MobileNet) [2]	9.83	29.09	49.13	66.70	16.25 ±6.66	11.66 ±5.38
2D to 3D (depth)	FPM (SqueezeNet) [2]	10.84	32.87	51.58	67.87	15.12 ±6.11	11.22 ±5.07
2D to 3D (depth)	HRNet-32 [3]	12.52	33.23	49.57	67.18	14.51 ±5.59	10.86 ±4.64
2D to 3D (depth)	HRNet-48 [3]	12.15	32.55	50.83	67.99	14.62 ±5.78	10.99 ±4.81
3D regression	ResNet-18 [4]	9.40	19.99	27.06	44.44	17.10 ±5.43	12.20 ±4.12
2D to 3D lifting	Martinez et al. [5] *	26.96	37.98	48.40	58.33	14.01 ±4.84	10.03 ±3.53
Vol. heatmaps	Pavlakos et al. [6]	18.15	42.24	61.60	86.15	10.35 ±1.07	7.11 ±0.65
SPDH (ours)	HRNet-32 [3]	**53.75**	**79.75**	**93.90**	**98.12**	**6.62** ±1.53	**4.41** ±1.09

* relative joint positions

1. Newell et al., "Stacked hourglass networks for human pose estimation". In ECCV 2016.
2. Martìnez-Gonzàlez et al., "Efficient convolutional neural networks for depth-based multi-person pose estimation". In IEEE Trans. Circuits Syst. Video Technol. 2019.
3. Sun et al., "Deep high-resolution representation learning for human pose estimation". In CVPR 2016.
4. He et al., "A simple yet effective baseline for 3d human pose estimation". In CVPR 2016.
5. Martinez et al., "Single-view robot pose ´ and joint angle estimation via render & compare". In ICCV 2016.
6. Pavlakos et al., "Coarse-to-fine volumetric prediction for single-image 3D human pose". In CVPR 2017.

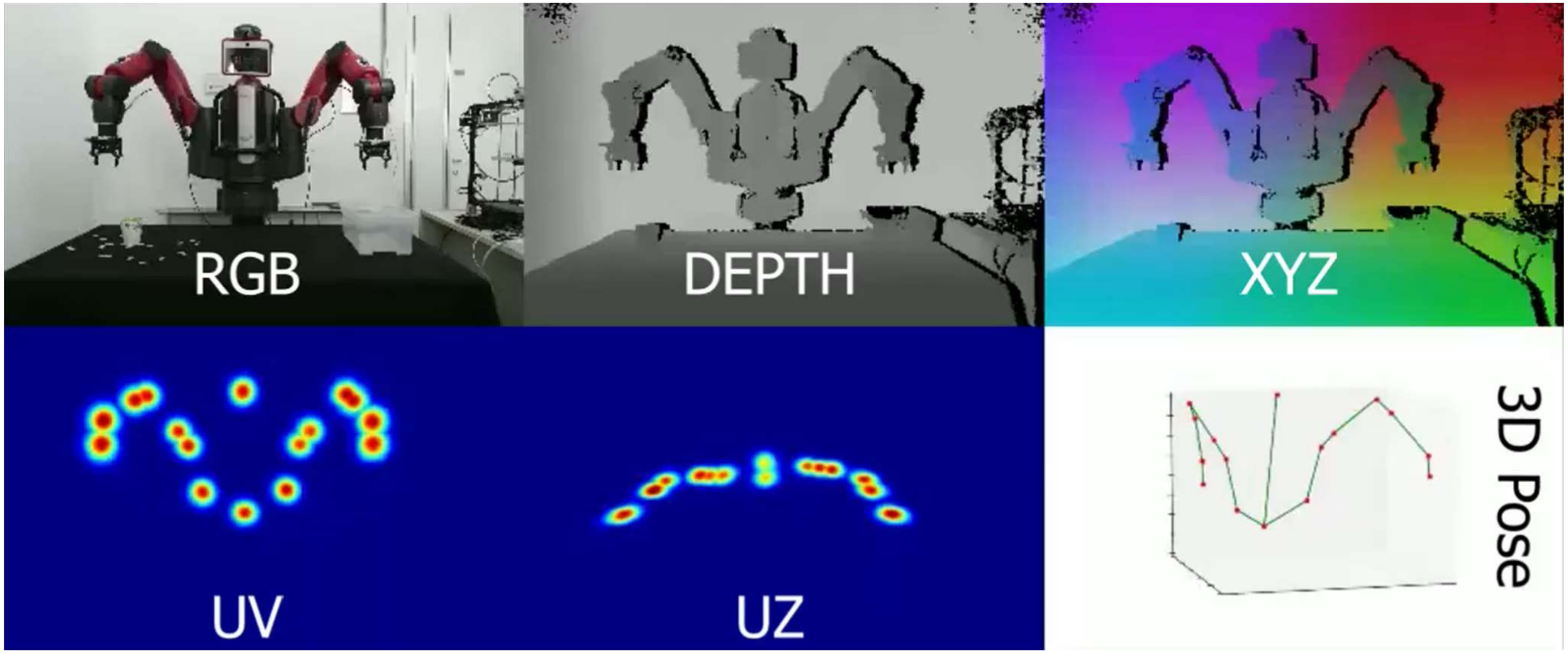
Qualitative Results

Ball pick-n-place real sequence



Qualitative Results

Cup pick-n-place real sequence



Conclusion

CONTRIBUTIONS

- Depth maps to reduce synth-to-real domain gap
- Semi-Perspective Decoupled Heatmaps (SPDH)
- SimBa dataset

Scan for project website:

https://aimagelab.ing.unimore.it/go/simba

KYOTO 2022
IEEE/RSJ International Conference on Intelligent Robots and Systems

THANK YOU!

"Semi-Perspective Decoupled Heatmaps for 3D Robot Pose Estimation from Depth Maps"

Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani

alessandro.simoni@unimore.it, s.pini@unimore.it, guido.borghi@unibo.it, roberto.vezzani@unimore.it

University of Modena and Reggio Emilia, Italy

UNIMORE AImage Lab AIRI AI Research & Innovation Center