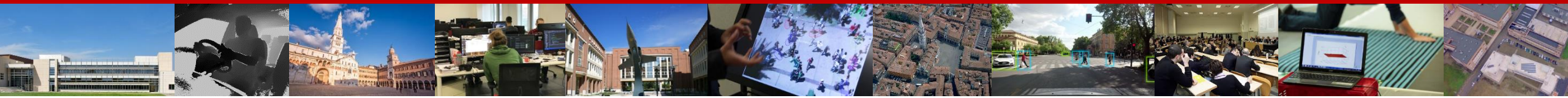
ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

RefiNet: 3D Human Pose Refinement with Depth Maps



Andrea D'Eusanio, Stefano Pini, Guido Borghi, Roberto Vezzani, Rita Cucchiara

{andrea.deusanio, s.pini, guido.borghi, roberto.vezzani, rita.cucchiara}@unimore.it

University of Modena and Reggio Emilia, Italy

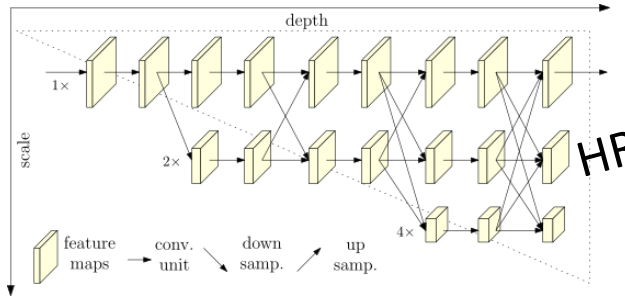
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

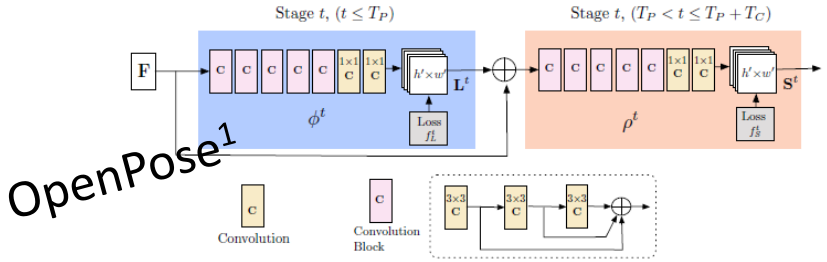
AImage Lab

Centro Interdipartimentale di Ricerca
Softech: ICT per le Imprese

Human Pose Estimation

- The *Human Pose Estimation* (HPE) is the localization of significant joints of the human body
- The combination of effective **deep learning approaches** and **huge datasets of RGB images** have led to impressive performance, in terms of **accuracy**, **computational load**, and **generalization capabilities**.





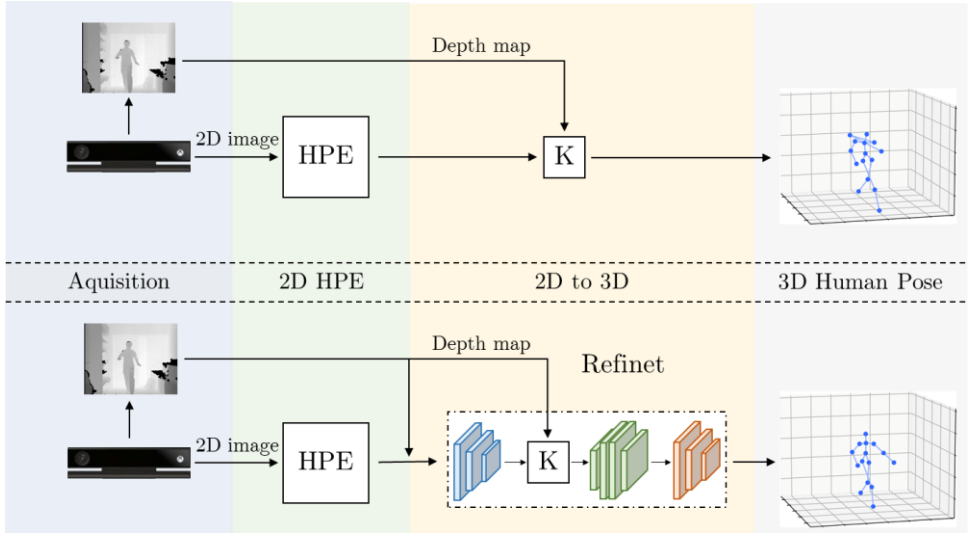
- These methods provide the pose only in **2D image coordinates**, thus without the **depth value** and **lacking any metric information**

1. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields”, in CVPR, 2017
2. *K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation”, in CVPR, 2019*

Our proposal

We aim to combine:

- the successful deep learning architectures for **2D human pose estimation**
- the 3D measurement capabilities of **depth sensors**



Thus, we propose **RefiNet**:

- Given: a **depth map** + an initial **2D human pose estimation**
- Regresses: a **precise 3D human pose** in world coordinates
- **Multi-stage modular** system:
  - Each module is **specialized** in a particular type of **refinement**
  - Modules can be **activated or deactivated** according to the needs
  - **Different data representations** are exploited, ranging from **2D depth patches** to **3D point clouds**

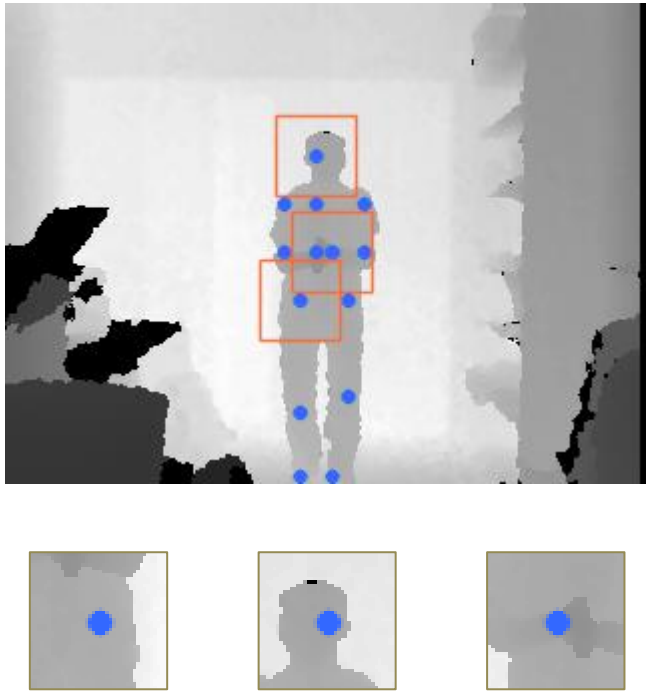
Initial 2D Human Pose Estimation

The starting point is the prediction of a set of human body joints over the depth map, in the form of

- (x,y) coordinates located on the **2D** depth image

This introduces some **problems**:

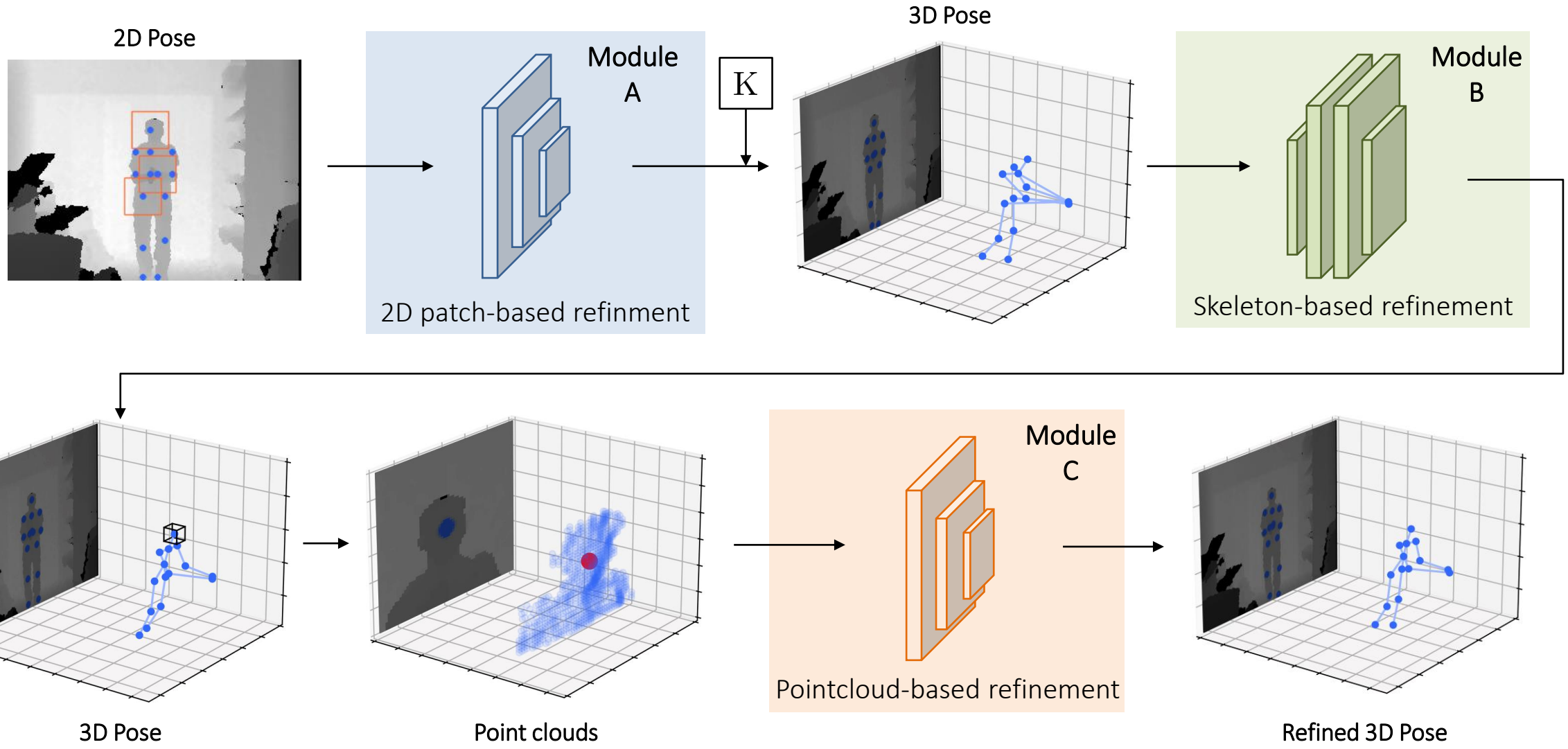
- **Obstructions:** if any occlusion (external or self occlusion) is present, the depth value at that location is wrong
- **z-value approximation:** even with an accurate 2D estimation, the 3D location calculated using the camera intrinsic parameters is always positioned on the outer surface of the body



Joints are not always locally precise

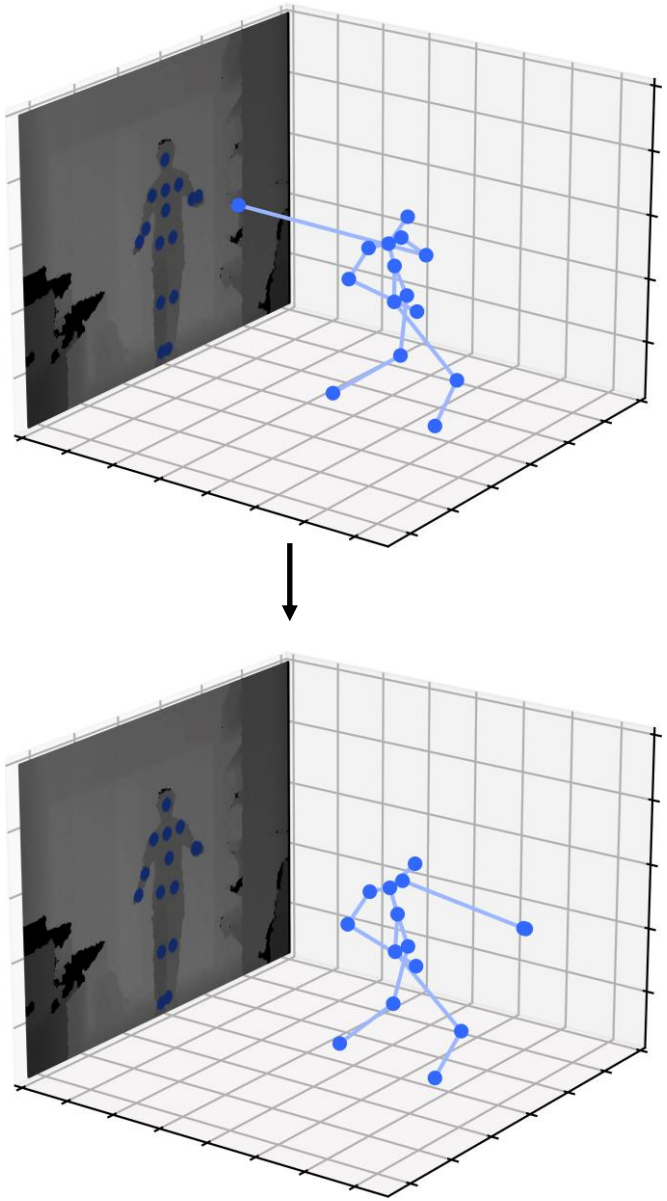
2.5D depth data is not exploited!

Method overview



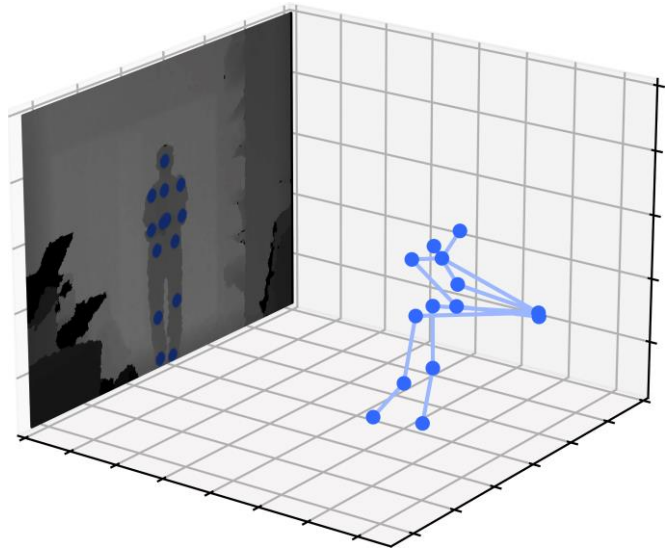
Module A: 2D patch-based refinement

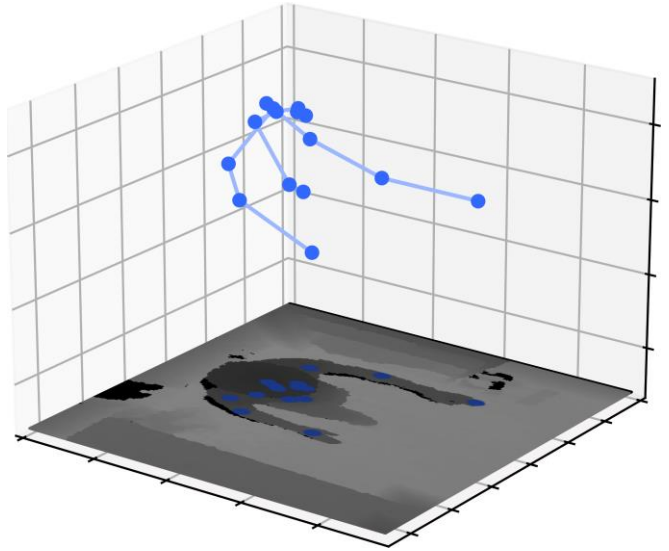
- It **addresses 2D displacement error** before the 2D to 3D conversion.
- It takes **2D patches** as input.
  Patches are a squared area centered at the detected joints location. Their size can be adapted to control how much information is provided for the refinement.
- It **predicts 2D offsets,** (x, y) displacement values from the previous coordinates (i.e. the center of the patch)
- Architecture:
  - 2 ResNet-like blocks
  - A FC layer for displacement regression
  - **0.8M** parameters, inference time of **1.7ms**



Module A: 2D patch-based refinement

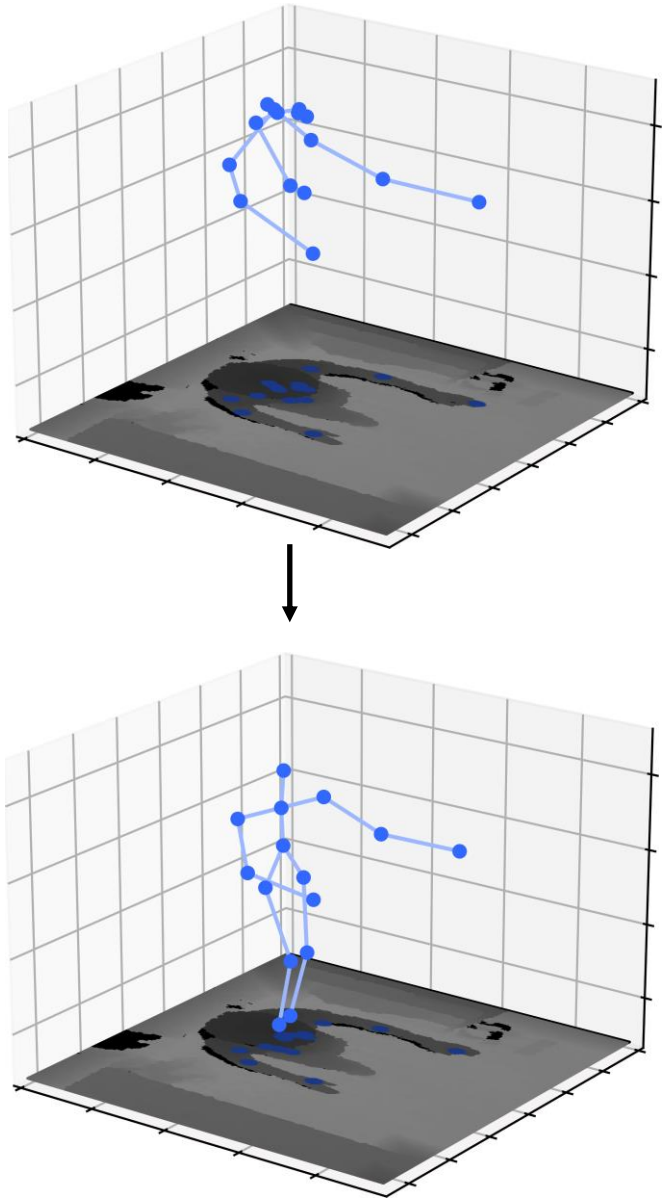
- This module **refines the 2D joints coordinates only**.
- Thus, it can't handle:
  - **Occlusions**
    This problem is not solved since the 2D to 3D conversion is done with the camera intrinsic parameters and the z-value retrieved from the depth frame
  - **Top views**
    having only depth information, is not directly possible to retrieve a correct depth value for the lower body part, obtaining a not plausible 3D skeleton.





Module B: skeleton-based refinement

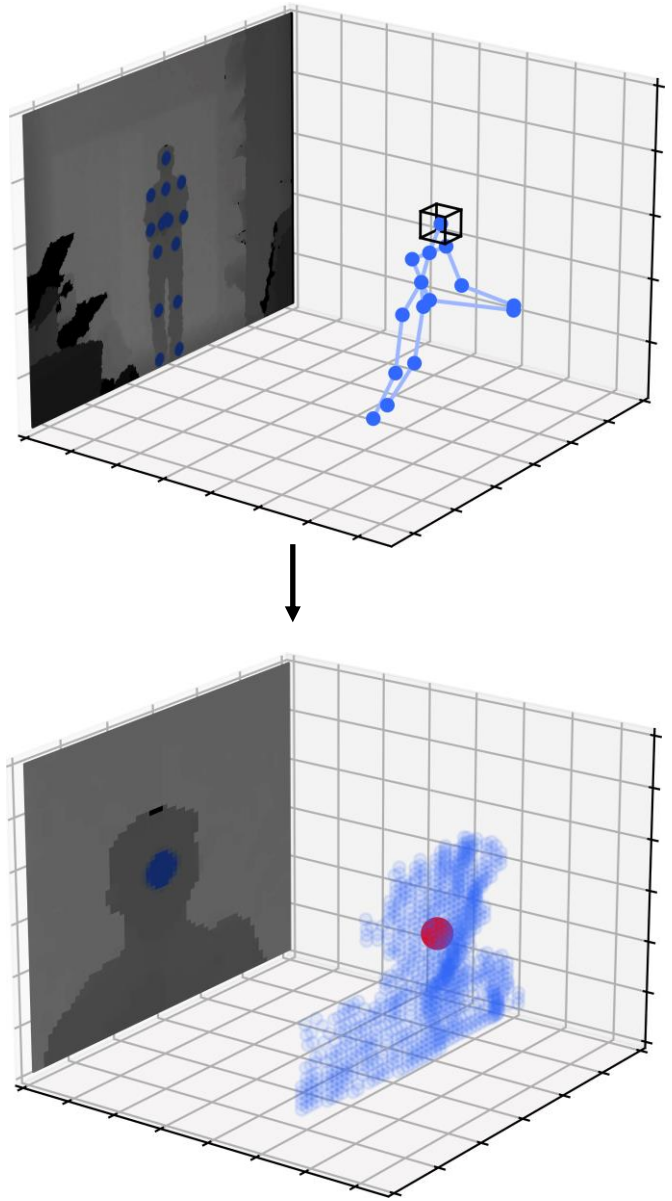
- It **refines** the global **3D skeleton structure**
- It takes **3D coordinates** of the **whole 3D skeleton** as input Coordinates are calculated from 2D locations using the camera matrix and the depth value at the 2D location.
- It **predicts 3D offsets,** (x, y, z) displacement values from the previous coordinates.
  - **z-values** are no more constrained on depth map values
  - The result is a 3D estimation in **world coordinates**
- Architecture:
  - 2 residual FC blocks, similar to [1]
  - **4.3M** parameters, inference time of just **0.8ms**!



1. *Martinez et al., "A simple yet effective baseline for 3d human pose estimation", in CVPR, 2017*

Module C: point cloud-based refinement

- It **refines** the local **3D body joints** in **world coordinates**
- It takes **3D point cloud patches** as input
  Point cloud is obtained from the depth map and a fixed-size **volume** around the 3D joint coordinate is extracted.
- It **predicts 3D offsets**, (x, y, z) displacement values from the previous coordinates.
  - Both the input and the output are 3D locations in **world coordinates**
- Architecture:
  - Based on PointNet[1]
  - **2.9M** parameters, 1.7GB RAM
  - Inference time of **13.8ms**

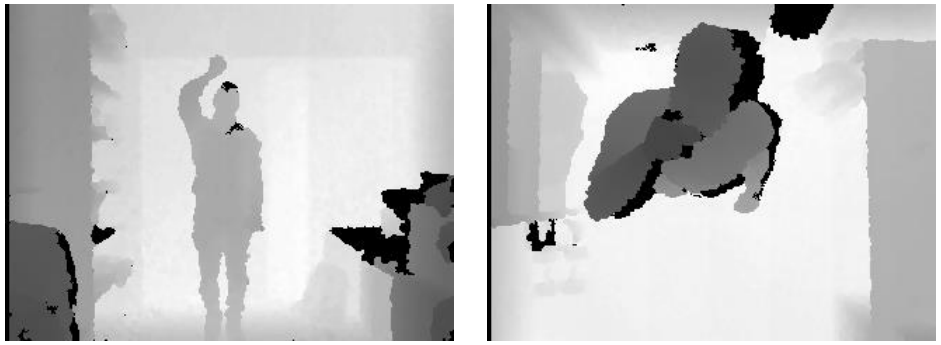


1. *Qi et al., "PointNet: Deep learning on point sets for 3d classification and segmentation", in CVPR, 2017*

Quantitative Results

ITOP dataset[1]:

- 100K depth images of a moving person in a static scene
- Annotations: 15 human body parts are labeled in 3D in camera frame
- Recorded with two Asus Xtion PRO (SL sensor, 320×240px) – Side and top views



Comparison between the **baseline** and **RefiNet**

Side view

Mod. A	Mod. B	Mod. C	OpenPose [4]				HRNet [5]			
			mAP ↑	Improv.	mDE ↓	Improv.	mAP ↑	Improv.	mDE ↓	
			0.646	-	12.634	-	0.670	-	10.711	-
✓			0.687	6.35%	10.442	17.4%	0.699	4.32%	10.060	6.08%
	✓		0.775	20.0%	8.463	33.0%	0.787	17.5%	8.185	23.6%
		✓	0.719	11.3%	11.834	6.33%	0.734	9.55%	10.693	0.17%
✓	✓		0.796	23.2%	8.042	36.3%	0.804	20.0%	7.790	27.3%
✓	✓	✓	**0.818**	**26.6%**	**7.646**	**39.5%**	**0.824**	**23.0%**	**7.447**	**30.5%**

Top view

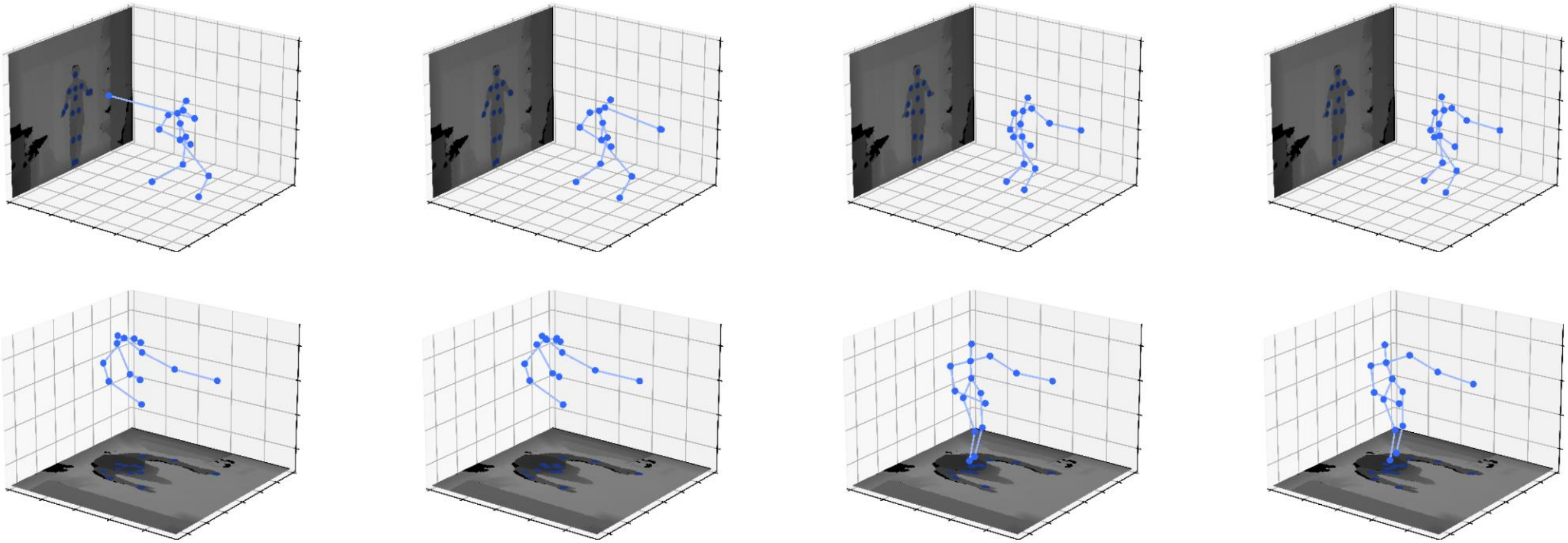
Mod. A	Mod. B	Mod. C	OpenPose [4]				HRNet [5]			
			mAP ↑	Improv.	mDE ↓	Improv.	mAP ↑	Improv.	mDE ↓	
			0.153	-	70.672	-	0.175	-	68.755	-
✓			0.164	7.19%	69.137	2.17%	0.173	-1.14%	68.580	0.25%
	✓		0.665	334.6%	10.464	85.2%	0.713	307.4%	9.836	85.7%
		✓	0.205	34.0%	68.218	3.47%	0.215	22.9%	66.444	3.36%
✓	✓		**0.675**	**341.2%**	**10.349**	**85.4%**	**0.718**	**310.3%**	**9.550**	**86.1%**
✓	✓	✓	0.619	304.6%	10.973	84.5%	0.663	278.9%	10.160	85.2%

Comparison with **competitors**

Body part	ITOP side view				ITOP top view			
	[38]	[19]	Bas.	Ours	[38]	[19]	Bas.	Ours
Upper Body	**84.8**	84.0	71.2	77.9	84.8	**91.4**	32.8	72.1
Lower Body	72.5	67.3	62.3	**85.7**	46.1	54.7	0.1	**71.4**
Full Body	80.5	77.4	67.0	**81.8**	68.2	**75.5**	17.5	71.8

1. *Haque et al., "Towards viewpoint invariant 3d human pose estimation", in ECCV, 2016*
2. *Downloadable at https://doi.org/10.5281/zenodo.3932973*

Qualitative Results

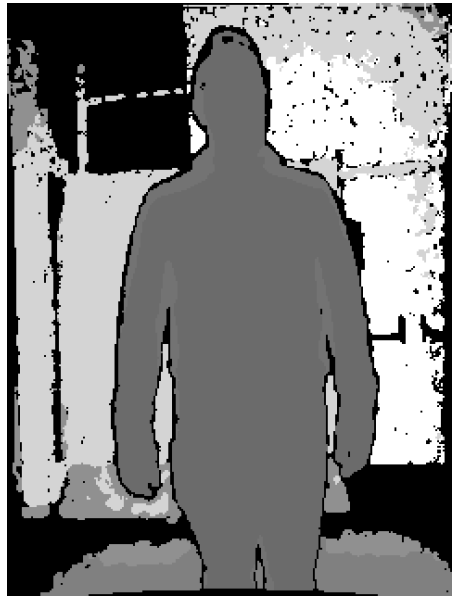


(a) 2D prediction (b) Module A (c) Module B (d) Module C

Estimation of anthropometric measurements

Baracca - A Multimodal Dataset for Anthropometric Measurements in Automotive:

- >9k frames in 4 modalities (RGB, IR, Depth, Thermal)
- 2 synchronized cameras (ToF RGB-D, Thermal)
- 4 streams: 8 viewpoints (3 in-car, 5 outside)
- 30 subjects
- **Multiple anthropometric measurements** (e.g. height, shoulder width, forearm length)
- **Soft-biometric traits** (age, sex, weight)



Method	Baseline		Ours	
	Mean	Std	Mean	Std
LR	**5.586**	1.468	5.656	**1.330**
AdaBoost	4.347	1.018	**3.372**	**0.755**
RF	2.230	0.377	**1.983**	**0.321**
kNN	**0.783**	0.503	1.276	**0.348**

Results on **height estimation**

Using HRNet + RefiNet, the **height estimation** from human body joints is more stable and precise.

1. *Pini et al., "Baracca: a multimodal dataset for anthropometric measurements in automotive", in IJCB, 2020*
2. *Downloadable at https://aimagelab.ing.unimore.it/go/baracca*

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Project page

Thank you for your attention

Code

RefiNet: 3D Human Pose Refinement with Depth Maps

Andrea D'Eusanio, Stefano Pini, Guido Borghi, Roberto Vezzani, Rita Cucchiara

{andrea.deusanio, s.pini, guido.borghi, roberto.vezzani, rita.cucchiara}@unimore.it

University of Modena and Reggio Emilia, Italy

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

Centro Interdipartimentale di Ricerca
Softech: ICT per le Imprese