Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images

Andrei Zanfir[2] Elisabeta Marinoiu[2] Mihai Zanfir[2] Alin-Ionut Popa[2] Cristian Sminchisescu[1,3]

[1]Lund University [2]IMAR [3]Google Research

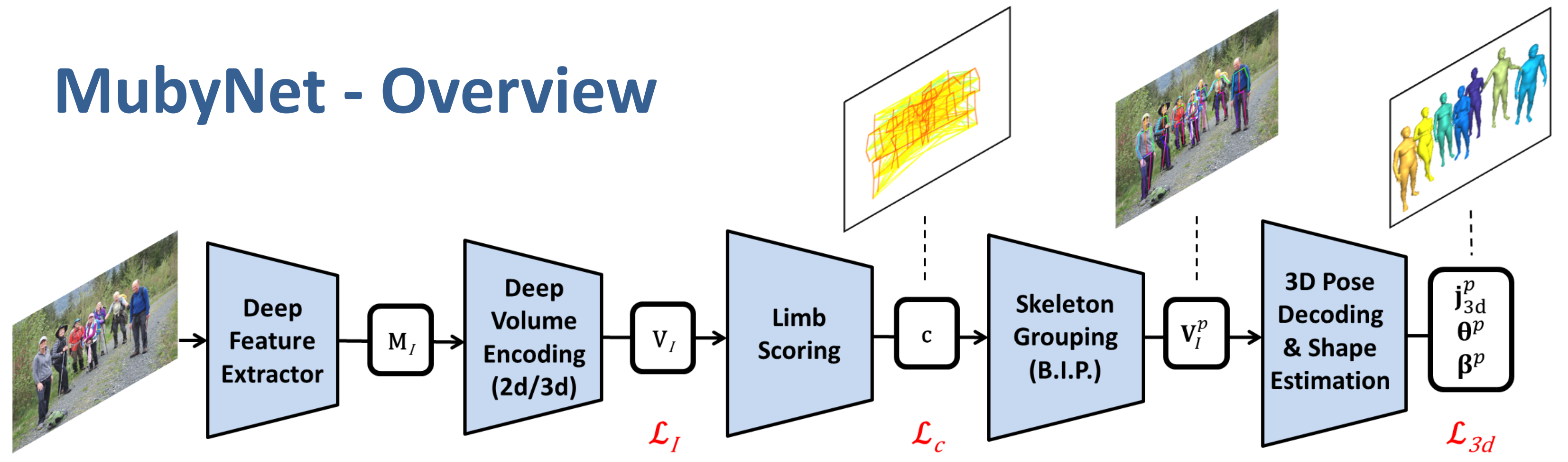
Introduction

We present MubyNet: a multitask, bottom up system for the integrated localization, 3d pose and shape estimation of multiple people in monocular images.

Contributions

- ✓ Novel encoding for 3d pose estimation of multiple people.
- ✓ Learn pairwise scoring functions from 2d and 3d information.
- ✓ Group body structures into 3d human skeleton hypotheses under kinematic tree constraints.
- ✓ State-of-the-art results on single and multiple people 3d datasets.

MubyNet - Overview

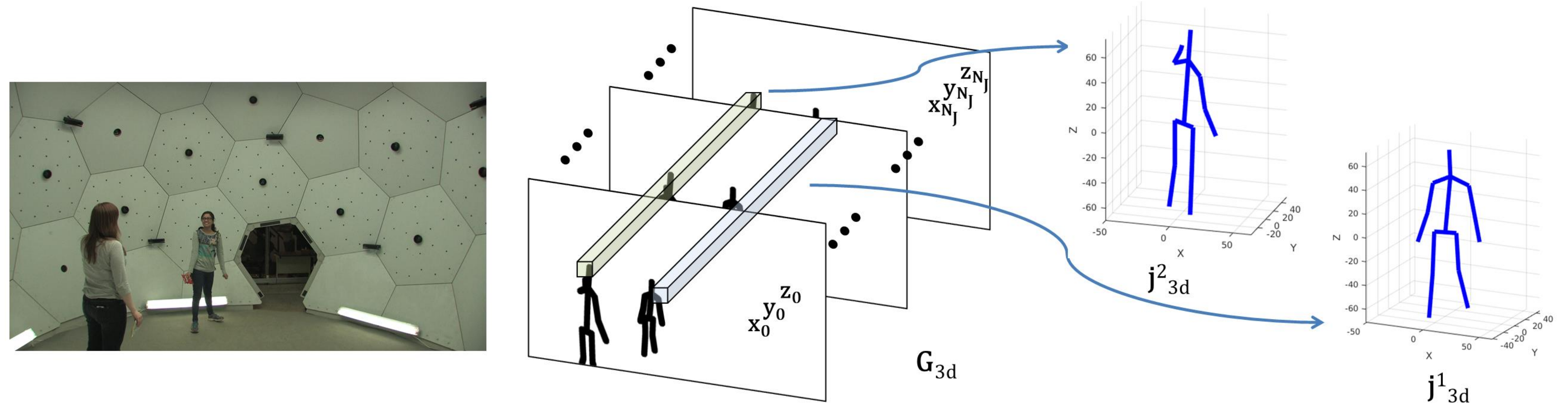

Deep Feature Extractor
M_I
Deep Volume Encoding (2d/3d)
V_I
$\mathcal{L}_I$
Limb Scoring
c
$\mathcal{L}_c$
Skeleton Grouping (B.I.P.)
V_I^p
3D Pose Decoding & Shape Estimation
j_{3d}^p θ^p β^p
$\mathcal{L}_{3d}$

Overview of MubyNet.

- Given an image $\boldsymbol{I}$, the processing stages are as follows:
  - ➢ ***Deep Feature Extractor*** computes image features $\boldsymbol{M}_I$.
  - ➢ ***Deep Volume Encoding*** regresses 2d and 3d pose information volumes, $\boldsymbol{V}_I$.
  - ➢ ***Limb Scoring*** collects all possible kinematic connections between 2d detected joints and predicts corresponding scores $\boldsymbol{c}$.
  - ➢ ***Skeleton Grouping*** assembles limbs into skeletons, $\boldsymbol{V}_I^{\boldsymbol{p}}$ by solving a binary integer linear program.
  - ➢ ***3D Pose Decoding & Shape Estimation*** produces the 3d pose and shape $\boldsymbol{j}_{3d}^{\boldsymbol{p}}$, (θ_p, β_p).
- Multitask losses constrain the output of the network.

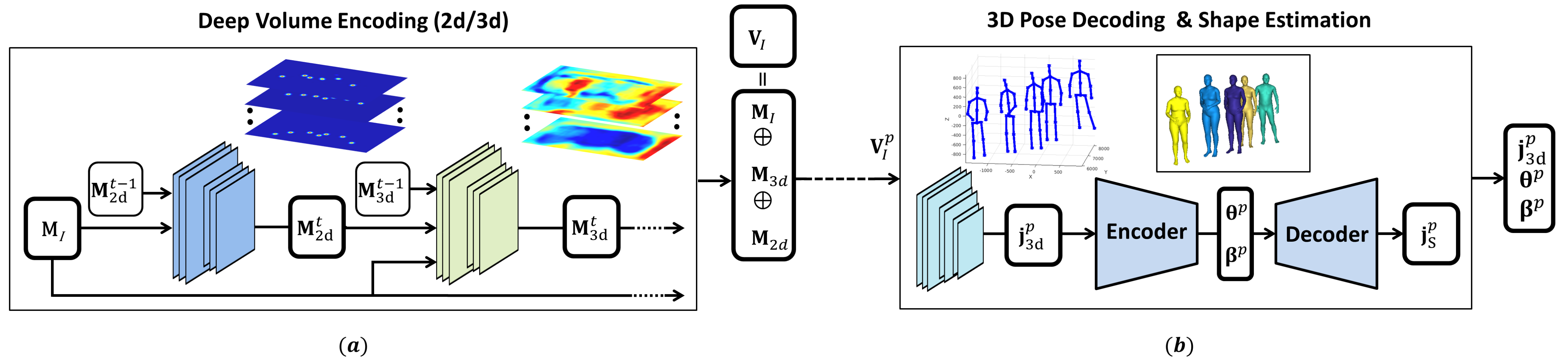
Deep Volume Encoding (2d/3d)

- Associate a slice in the volume to each one of the $\boldsymbol{N}_{J \times 3}$ joint components.
- Encode each 3d skeleton $\boldsymbol{j}_{3d}^{\boldsymbol{p}}$, of a person $\boldsymbol{p}$, by writing its components in the corresponding slices, but only for spatial locations within the image projection of the skeleton.



The volume encoding of multiple 3d ground truth skeletons in a scene.

3D Pose Decoding and Shape Estimation



(a) Detailed view of a single stage $\boldsymbol{t}$ of our multi-stage ***Deep Volume Encoding (2d/3d)*** module.

- The image features $\boldsymbol{M}_I$, together with predictions from the previous stage, $\boldsymbol{M}_{3d}^{\boldsymbol{t-1}}$ and $\boldsymbol{M}_{2d}^{\boldsymbol{t-1}}$, are used to refine the current representations $\boldsymbol{M}_{3d}^{\boldsymbol{t}}$ and $\boldsymbol{M}_{2d}^{\boldsymbol{t}}$.
- The multi-stage module outputs $\boldsymbol{V}_I$, which represents the concatenation of $\boldsymbol{M}_I$, $\boldsymbol{M}_{\boldsymbol{3d}} = \sum_{\boldsymbol{t}} \boldsymbol{M}_{3d}^{\boldsymbol{t}}$ and $\boldsymbol{M}_{\boldsymbol{2d}} = \sum_{\boldsymbol{t}} \boldsymbol{M}_{2d}^{\boldsymbol{t}}$.

(b) Detailed view of the ***3D Pose Decoding & Shape Estimation*** module.

- Given the estimated volume encoding $\boldsymbol{V}_I$, and the person partitions $\boldsymbol{V}_I^{\boldsymbol{p}}$, we decode the 3d pose $\boldsymbol{j}_{3d}^{\boldsymbol{p}}$.
- We recover the model pose and shape parameters (θ_p, β_p) using an auto-encoder.

Experimental Results

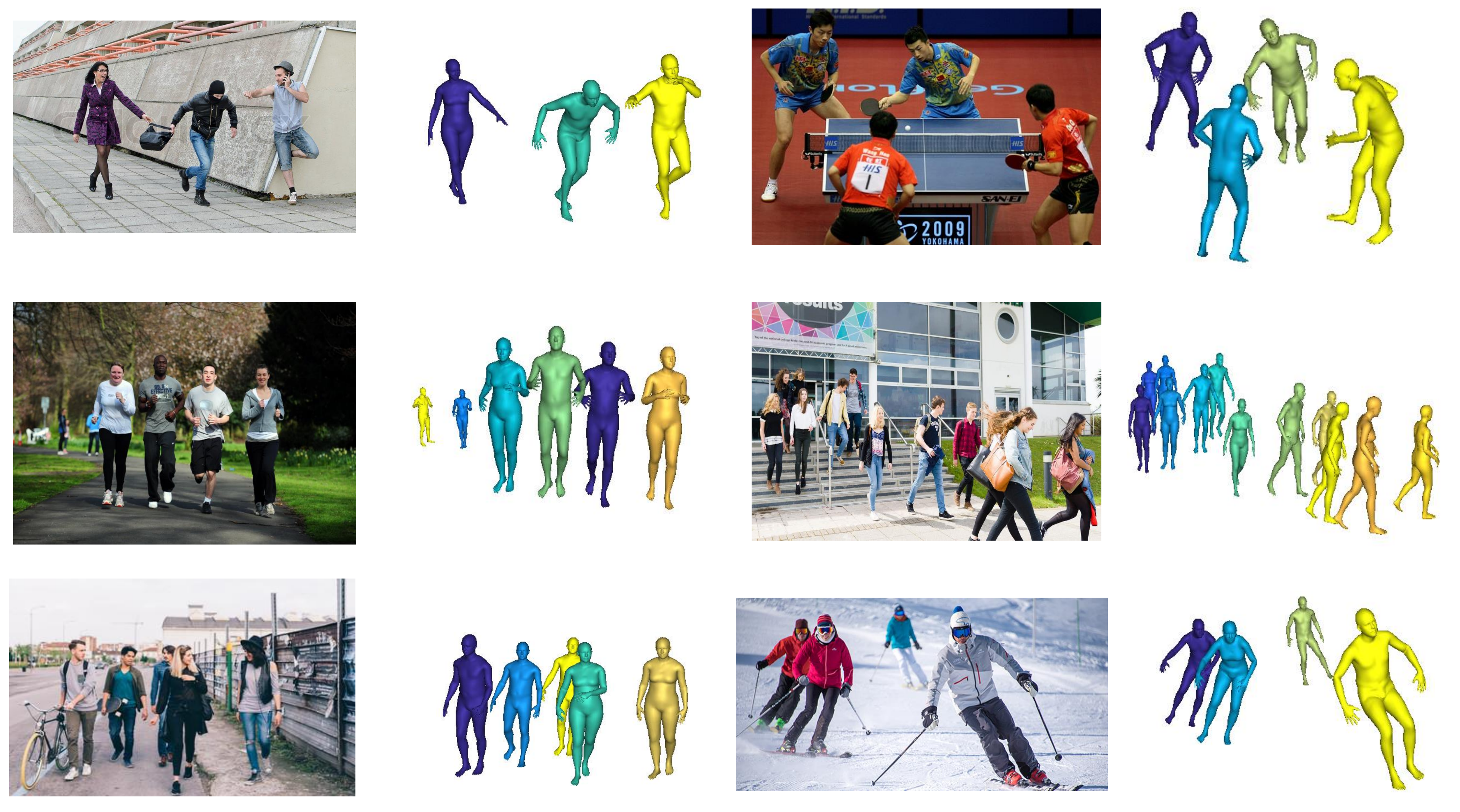
Method	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	Mean
DMHS	60	56	68	64	78	67	68	106	119	77	85	64	57	78	62	73
Zanfir et al.	54	54	63	59	72	61	68	101	109	74	81	62	55	75	60	69
MubyNet	**49**	**47**	**51**	**52**	**60**	**56**	**56**	**82**	**94**	**64**	**69**	**61**	**48**	**66**	**49**	**60**

Mean per joint 3d position error (in mm) on the Human3.6M dataset.

Method	MPJPE (mm)
DMHS	63.35
MubyNet	59.31
MubyNet Attention	**58.40**

Method	Haggling	Mafia	Ultimatim	Pizza	Mean
DMHS	217.9	187.3	193.6	221.3	203.4
Zanfir et al.	140.0	165.9	150.7	156.0	153.4
MubyNet	141.4	152.3	145.0	162.5	150.3
MubyNet Fine-Tuned	**72.4**	**78.8**	**66.8**	**94.3**	**72.1**

(Left) Human 80K. *Our method obtains state-of-the art results. Adding an attention mechanism for decoding 3d information, further improves the performance.* ***(Right) CMU Panoptic dataset.*** *Our method performs better than previous works even when using only 3d supervision from Human80K Fine-tuning on the CMU Panoptic dataset drastically reduces the error.*



Examples of pose and shape reconstructions produced by MubyNet.