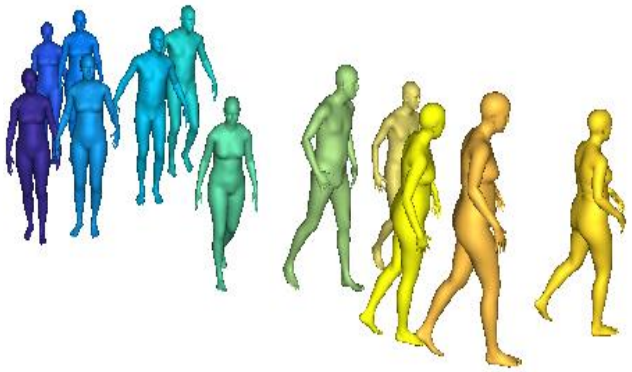
Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images

Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu

Objective

Single input image

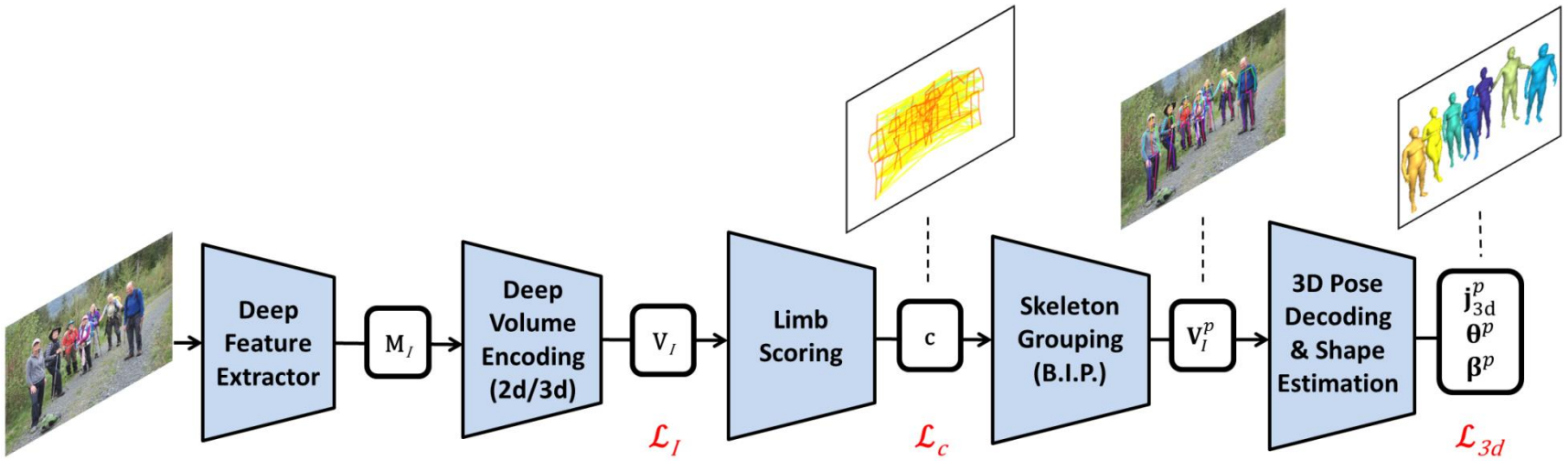
Automatic 3d pose and shape reconstruction



Automatic, feed-forward model, to predict the 3d body shape and pose of multiple people, given a single input image

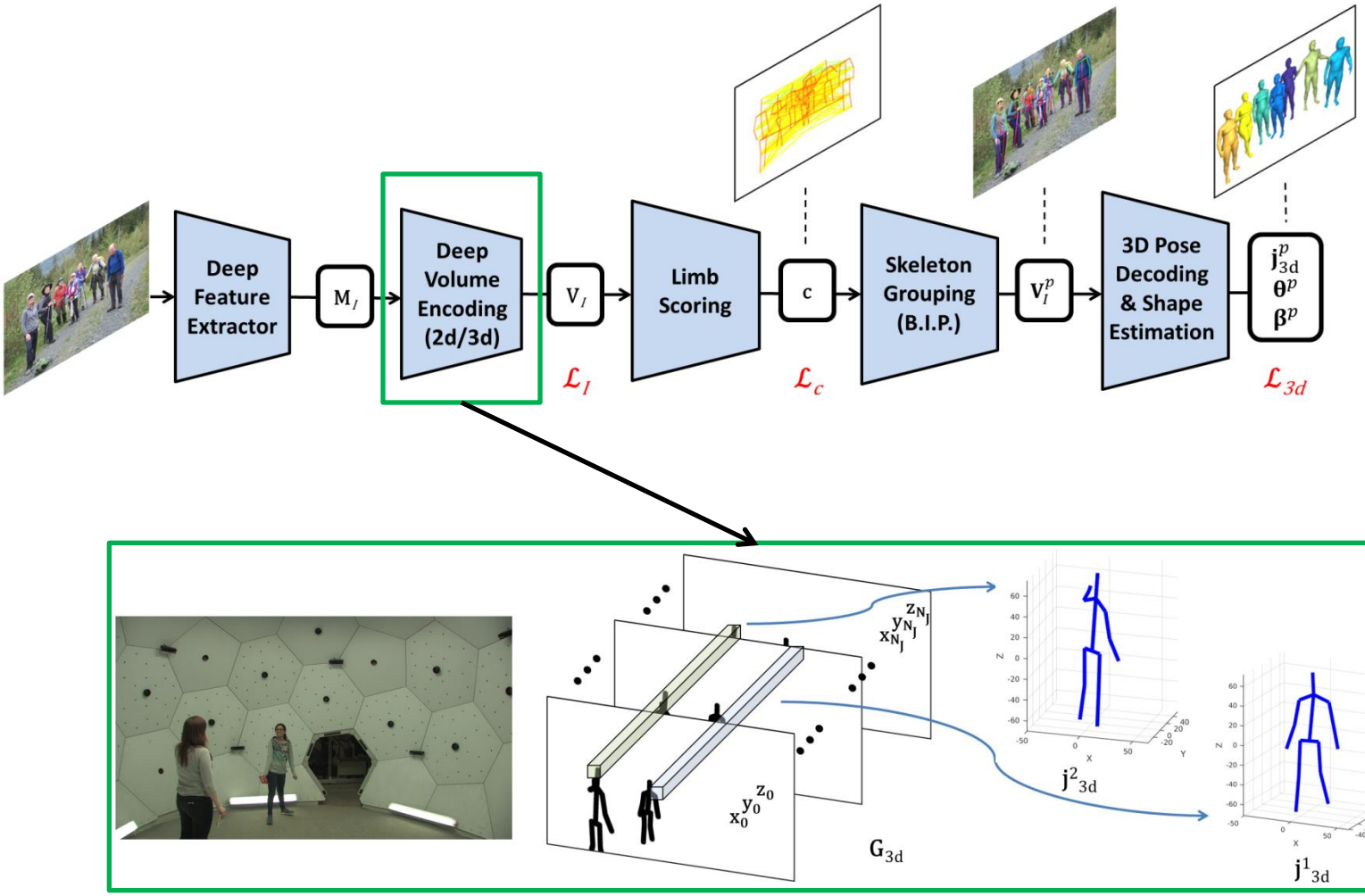
Challenges: multiple people, occlusions, depth ambiguities, difficult to formulate a single cost function and an integrated learning process

MubyNet
(Multi Body Net)

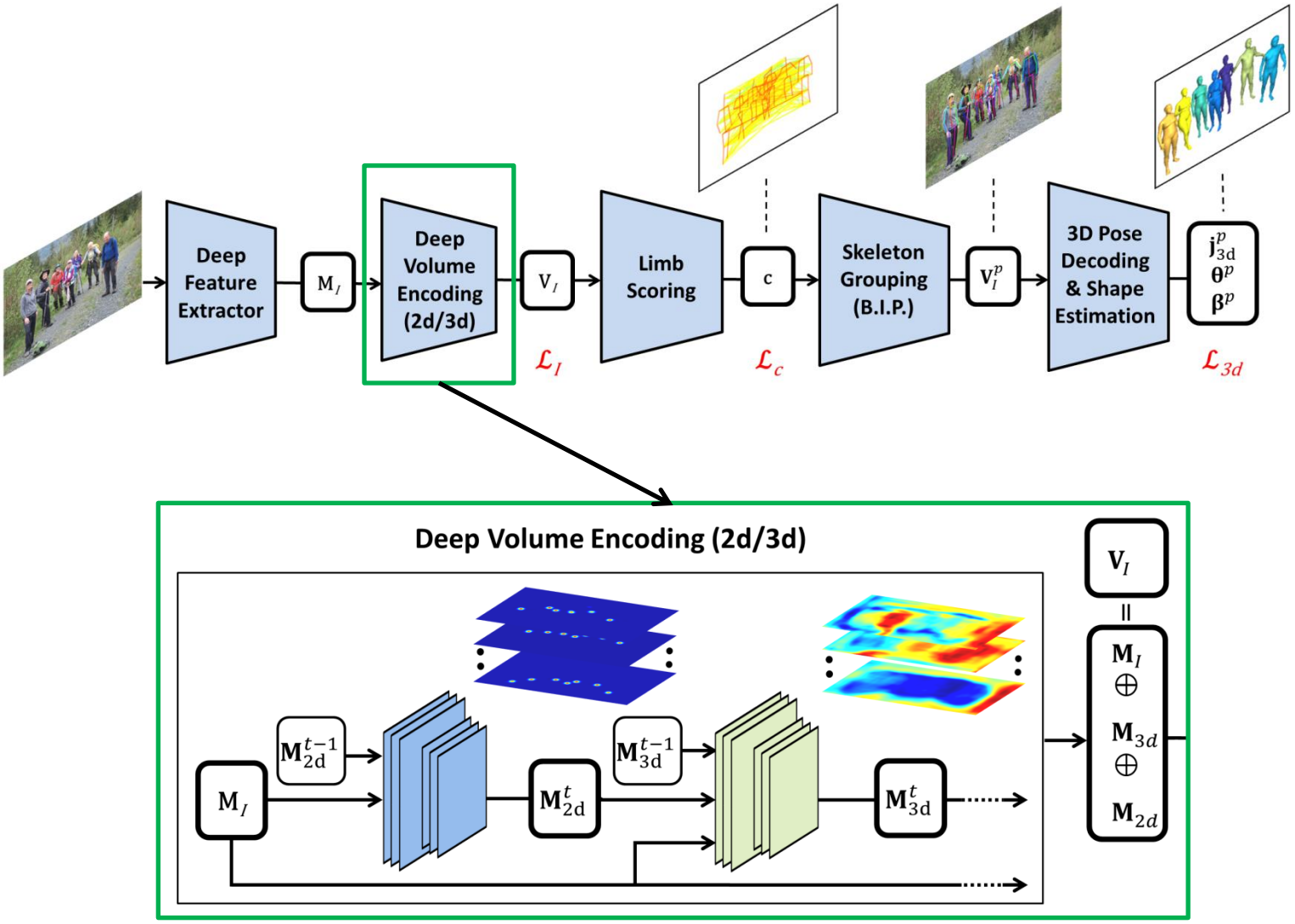


- Formulate a single, feedforward model with discrete and continuous components
- Multiple tasks: body joint detection, person grouping, pose and shape estimation
- Integrated representation based on 3d reasoning at all stages

Deep Volume Encoding

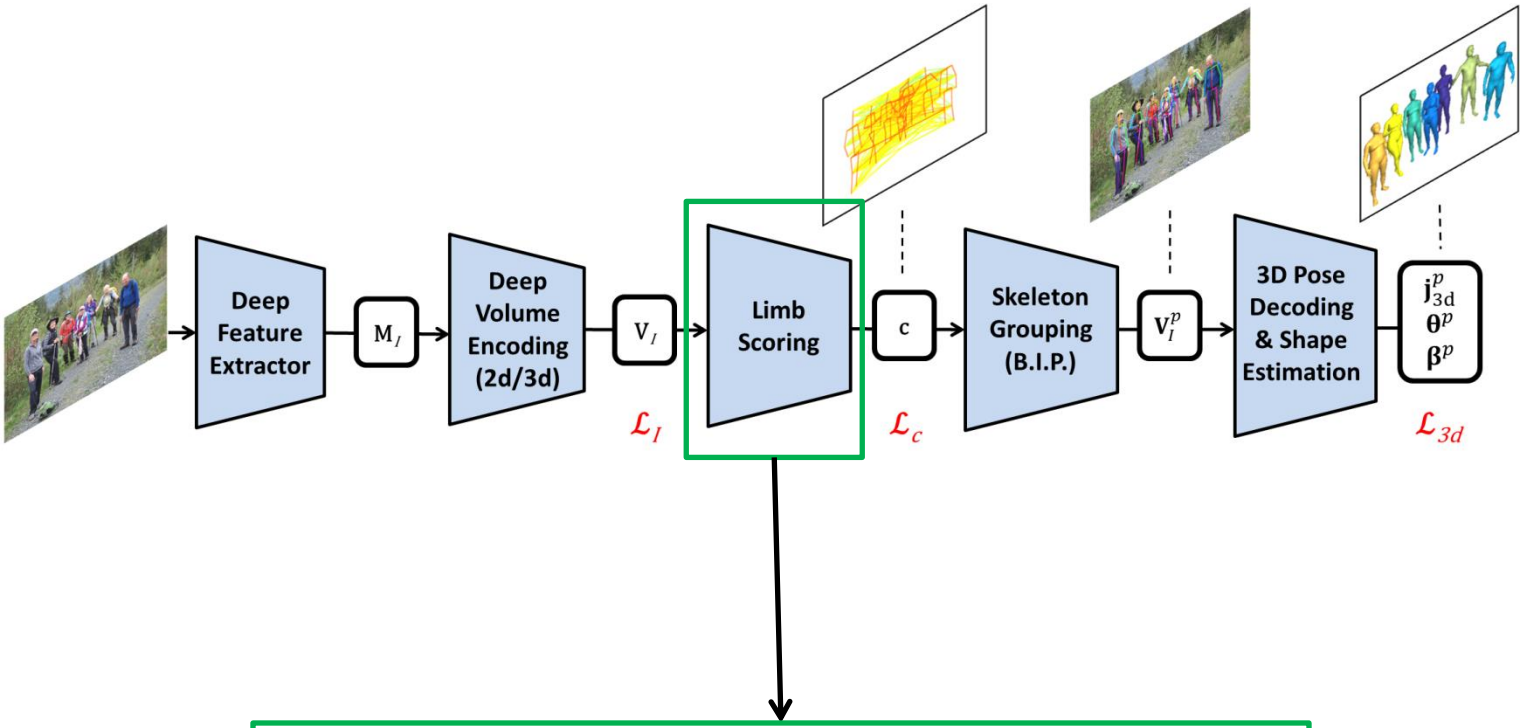


Deep Volume Encoding



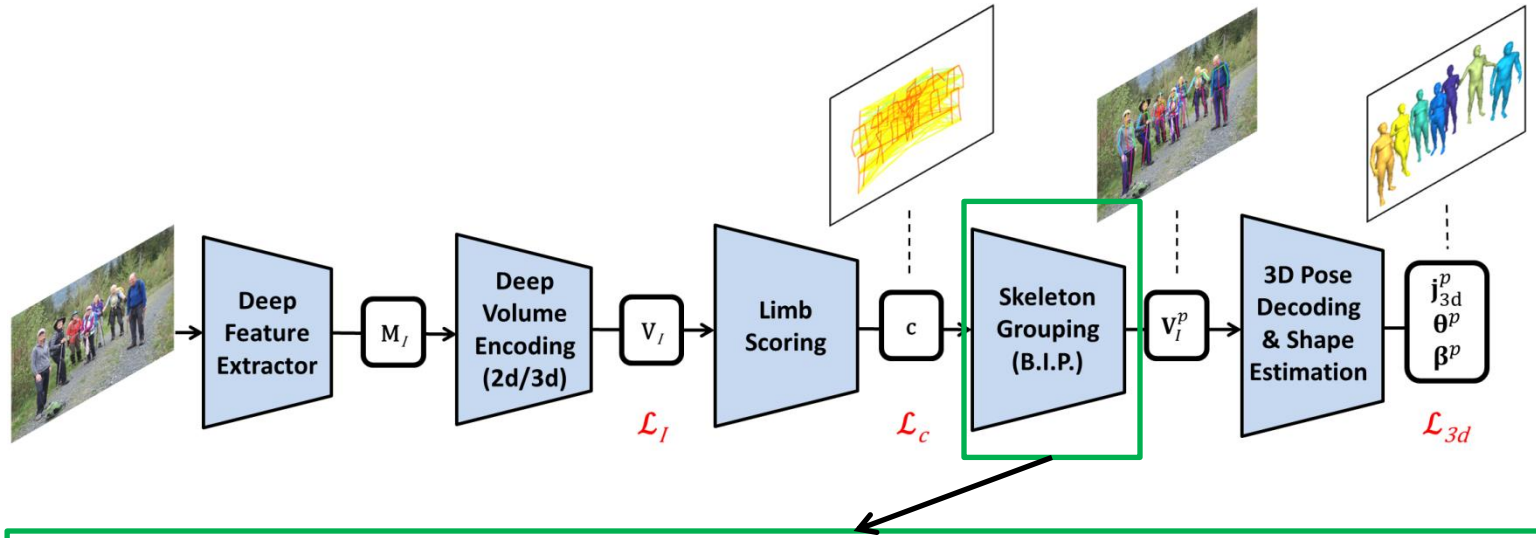
Multi-stage architecture

Limb Scoring

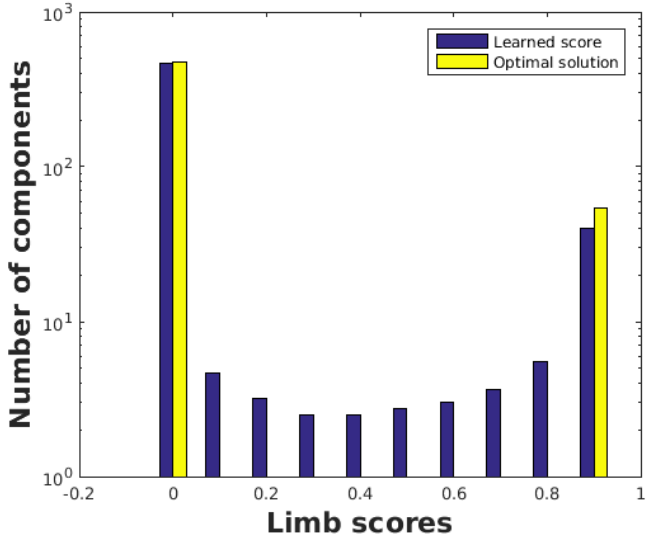


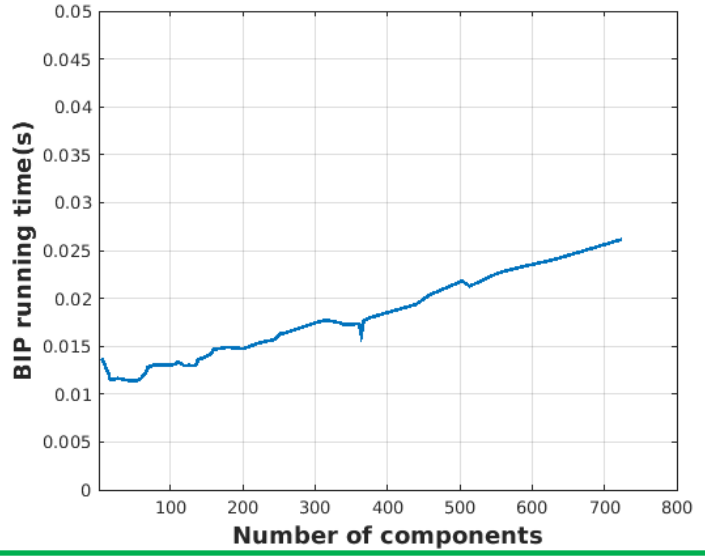
Limb Scoring collects all possible kinematic connections between 2D detected joints and predicts corresponding scores $\boldsymbol{c}$.

Skeleton Grouping via B.I.P

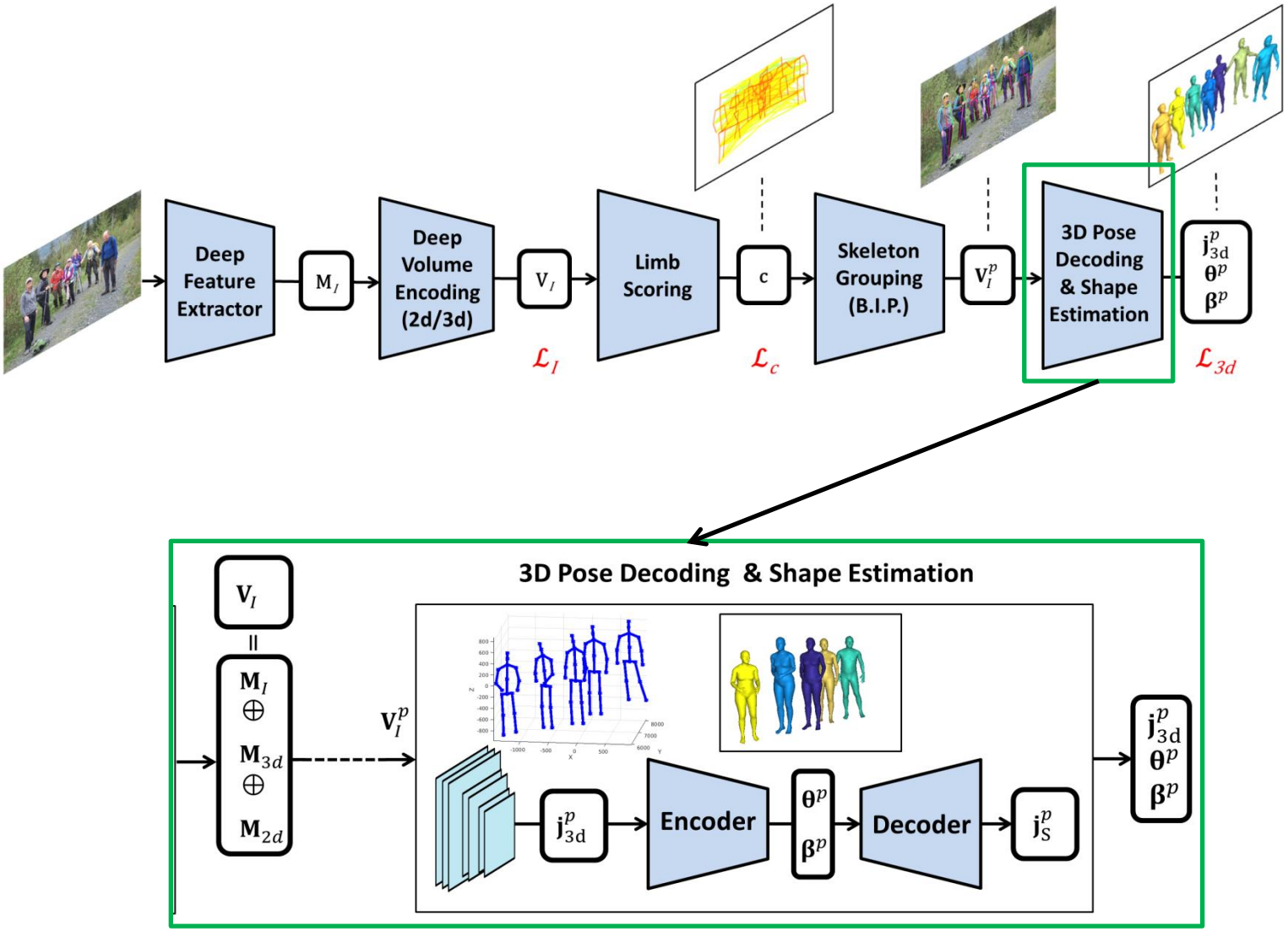


$$\mathbf{x}^*(\mathbf{c}) = \operatorname{argmax} \mathbf{c}^\top \mathbf{x}, \text{ subject to } \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \{0,1\}^{N_L \times 1}$$





3D Pose Decoding & Shape Estimation



M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," SIGGRAPH

Results

Visit our poster for videos!
Room 210 & 230 AB #120

Method	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	Mean
[1]	60	56	68	64	78	67	68	106	119	77	85	64	57	78	62	73
[2]	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**

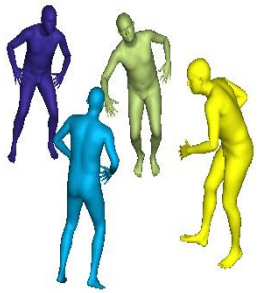
*- Mean per joint 3d position error (in mm) on the **Human3.6M** dataset -*

Method	Haggling	Mafia	Ultimatim	Pizza	Mean
[1]	217.9	187.3	193.6	221.3	203.4
[2]	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**

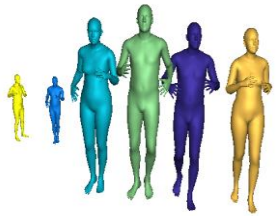
*- MPJ3DPE on the **CMU Panoptic** dataset -*

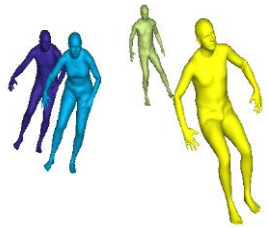
Method	MPJPE (mm)
[1]	63.35
MubyNet	59.31
MubyNet Attention	**58.40**

*- MPJ3DPE on the **Human80k** dataset -*







[1] A. I. Popa, M. Zanfir, and C. Sminchisescu, "Deep multitask architecture for integrated 2d and 3d human sensing," in CVPR, 2017
[2] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes – The Importance of Multiple Scene Constraints," in CVPR, 2018.