VirtualPose: Learning Generalizable 3D Human Pose Models from Virtual Data

Jiajun Su[1,2] Chunyu Wang[5] Xiaoxuan Ma[2,3] Wenjun Zeng[6] Yizhou Wang[2,3,4]

[1] Center for Data Science, Peking University
[2] Center on Frontiers of Computing Studies, Peking University
[3] Dept. of Computer Science, Peking University [4] Inst. for Artificial Intelligence, Peking University
[5] Microsoft Research Asia [6] EIT Institute for Advanced Study

Background

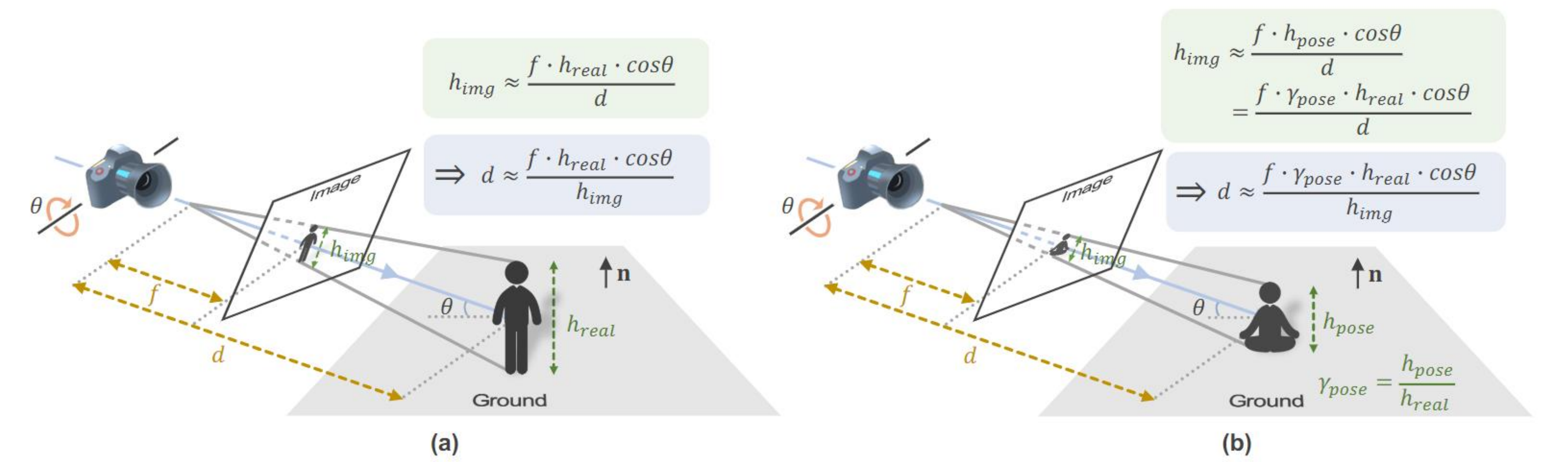
Problem

- Generalization ability of monocular 3D human pose estimation.

Motivation

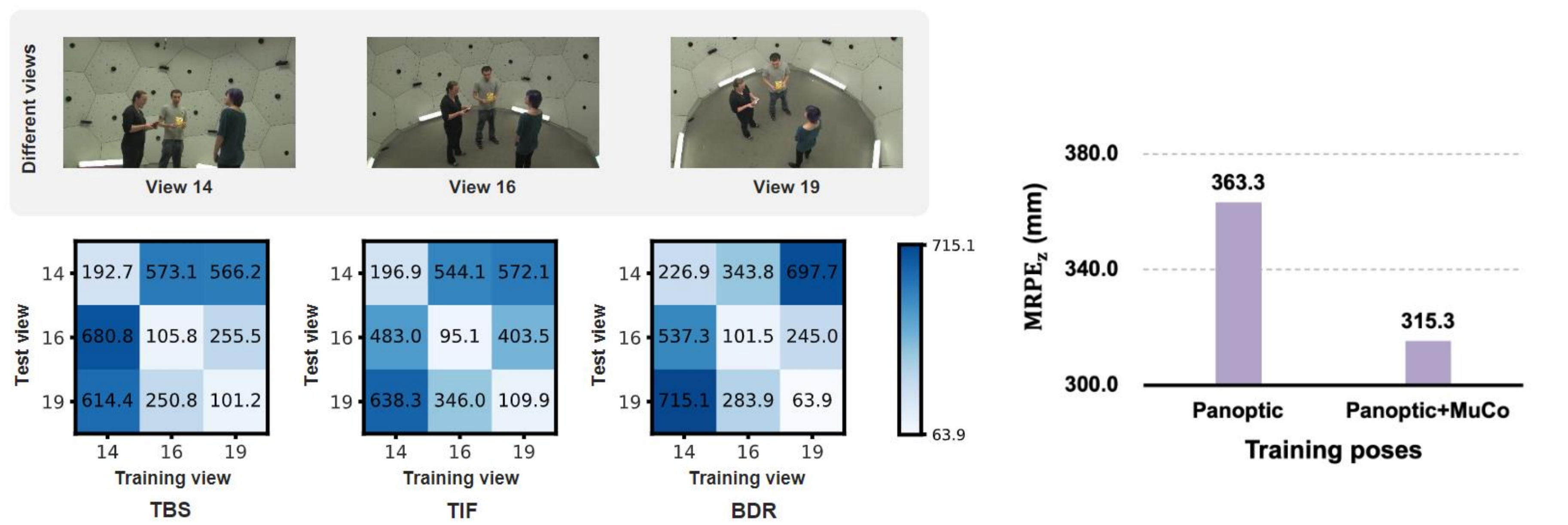
- Although monocular 3D pose estimation have achieved good results on the public datasets, their generalization ability is largely overlooked.
- 3D pose datasets have extremely limited variations in terms of cameras, human poses and appearance.

Projection Geometry



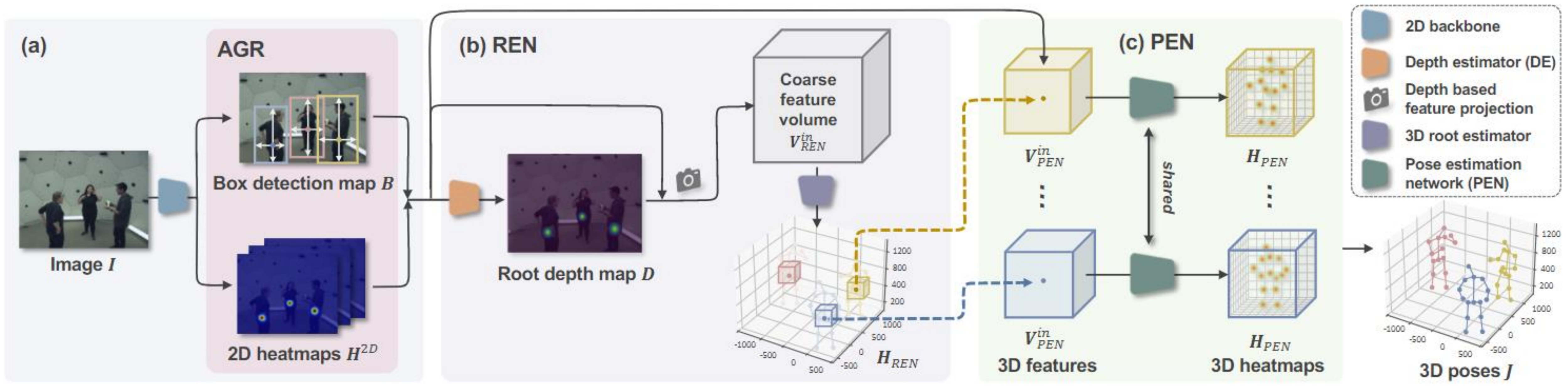
Generalization Study

- Systematically evaluate the robustness of the existing methods to the variations of the key factors.
- Models get notably larger errors when tested on different cameras and human poses.



(a) Camera views (b) Human Pose

Pipeline



- Abstract Geometry Representation (AGR) splits a 3D pose estimator into two successive modules.

1. The first module maps a raw image to AGR, can be trained on the diverse 2D datasets.
2. The second module maps AGR to the corresponding 3D pose, can be trained on synthesized <AGR, Pose> data.

- Loss function
  - First module

$$\mathcal{L}_{heat}^{2D} = ||\boldsymbol{H}^{2D} - \widetilde{\boldsymbol{H}}^{2D}||_2^2$$

$$\mathcal{L}_{bbox} = \sum_{p \in \mathcal{P}} ||\boldsymbol{B}_p - \widetilde{\boldsymbol{B}}_p||_1,$$

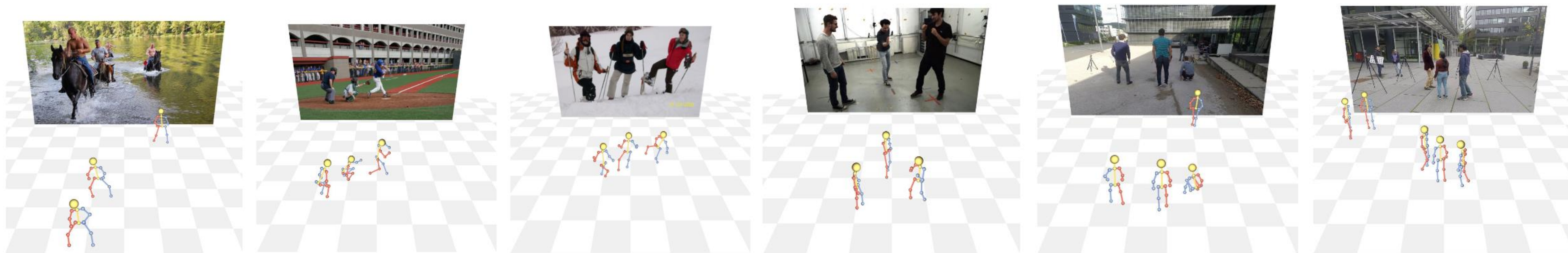
  - Second module

$$\mathcal{L}_{depth} = \sum_{p \in \mathcal{P}} |\boldsymbol{D}_p - \widetilde{\boldsymbol{D}}_p|$$

$$\mathcal{L}_{REN} = ||\boldsymbol{H}_{REN} - \widetilde{\boldsymbol{H}}_{REN}||_2^2$$

$$\mathcal{L}_{PEN} = \frac{1}{N} \sum_{k=1}^{N} ||\boldsymbol{J}_k - \widetilde{\boldsymbol{J}}_k||_1$$

Experiments



Method	Haggling	Mafia	Ultimatum	Pizza	Mean
RootNet	89.6	91.3	79.6	90.1	87.6
Zanfir et al.	72.4	78.8	66.8	94.3	78.1
SMAP	<u>63.1</u>	**60.3**	<u>56.6</u>	<u>67.1</u>	<u>61.8</u>
Ours	**54.1**	<u>61.6</u>	**54.6**	**65.4**	**58.9**

Method	REN	PEN	MRPE	MRPE$_z$
DE	2D	×	113.9	104.1
REN	2D+3D	×	115.7	93.6
Ours	2D+3D	√	**97.0**	**86.0**

CMU Panoptic

Method	Matched People		All People
	PCK$_{abs}$	PCK$_{root}$	PCK$_{abs}$
RootNet	31.8	31.0	31.5
HDNet	35.2	39.4	-
SMAP	38.7	45.5	35.4
HMOR	-	-	43.8
Ours	**47**	**52.3**	**44.0**

Generalization ability

MuPoTS-3D

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022
October 23-27, 2022, Tel Aviv

GitHub