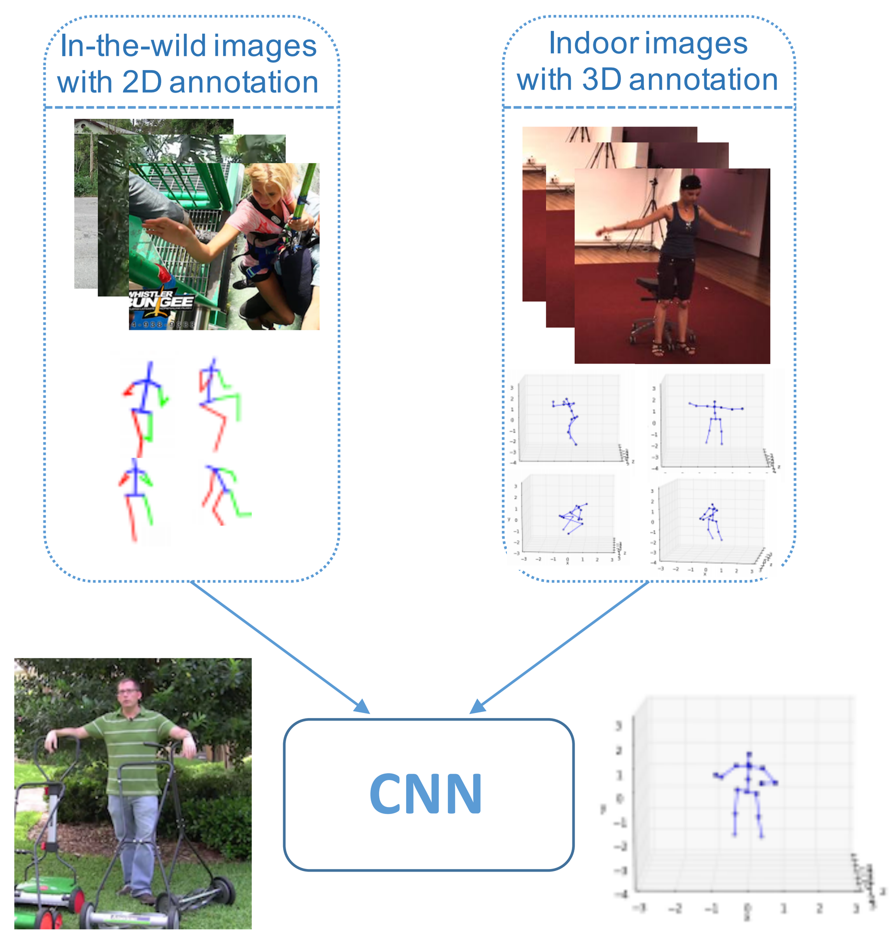
Towards 3D Human Pose Estimation in the Wild: a Weakly-supervised Approach

Xingyi Zhou[1,2], Qixing Huang[2], Xiao Sun[3], Xiangyang Xue[1], Yichen Wei[3]

[1] Fudan University, [2] The University of Texas at Austin, [3] Microsoft Research

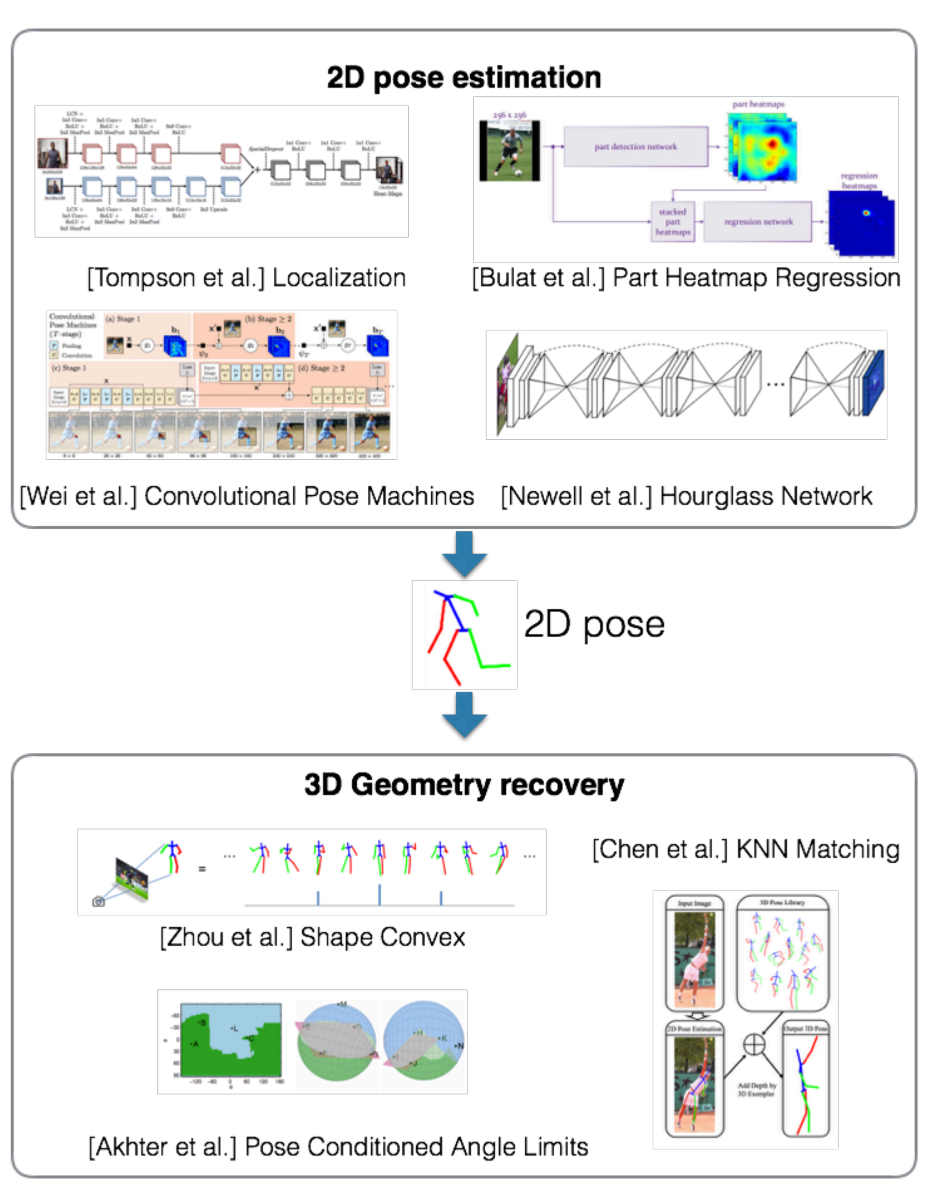
{zhouxy13,xyxue}@fudan.edu.cn, huangqx@cs.utexas.edu, {xias, yichenw}@microsoft.com

Motivation



- Goal: estimate 3D human pose for in-the-wild image.
- In-the-wild images with only 2D annotations.
- 3D annotated images only in indoor environment.

Previous Approaches



The original in-the-wild 2D image, which contains rich cues for 3D pose recovery, is discarded in the second step.

Framework

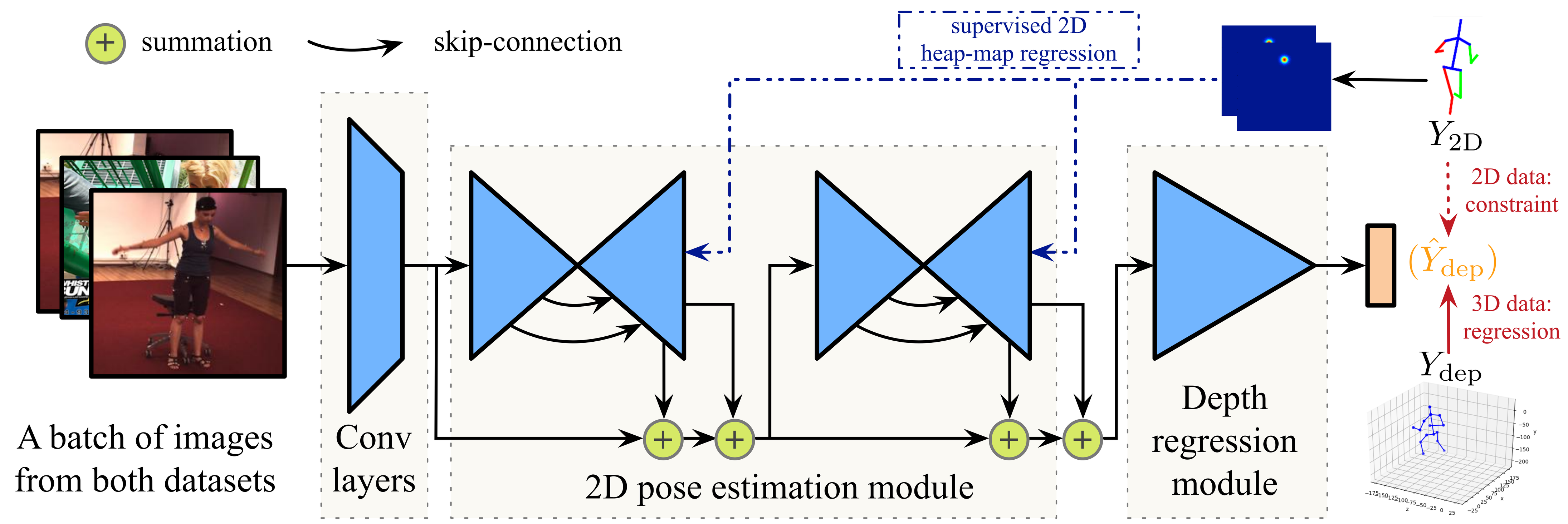


Figure 1: Illustration of our framework: In testing, images go through the stacked hourglass network and turn into 2D heat-maps. The 2D heat-maps and with lower-layer images features are summed as the input of the following depth regression module. In training, images from both 2D and 3D datasets are mixed in a single batch. For the 3D data, the standard regression with Euclidean Loss is applied. For the 2D data, we propose a weakly-supervised loss based on its 2D annotation and prior knowledge of human skeleton.

Method

Task formulation

Assumption: weak-perspective camera

$$Y_{3D} = [Y_{2D}, Y_{dep}]$$

In-the-lab Image with 3D annotation

$$\mathcal{S}_{3D} = \{\mathcal{I}_{3D}, \mathcal{Y}_{2D}, \mathcal{Y}_{dep}\}$$

In-the-Wild Image with 2D annotation

$$\mathcal{S}_{2D} = \{\mathcal{I}_{3D}, \mathcal{Y}_{2D}\}$$

2D Pose Estimation

Stacked hourglass network [Newell et al.]

$$L_{2D}(\hat{Y}_{HM}, Y_{2D}) = \Sigma_h^H \Sigma_w^W (\hat{Y}_{HM}^{(h,w)} - G(Y_{2D})^{(h,w)})^2$$

Depth Regression

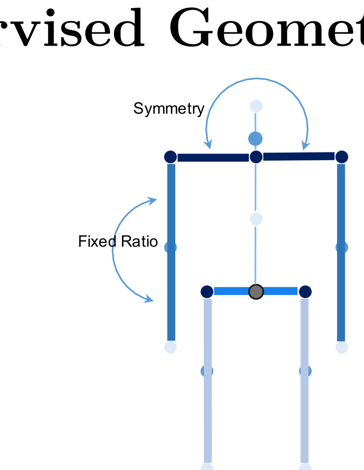
$$L_{dep}(\hat{Y}_{dep}|I, Y_{2D}) = \begin{cases} \lambda_{reg}||Y_{dep} - \hat{Y}_{dep}||^2, if\ I \in \mathcal{I}_{3D} \\ \lambda_{geo} L_{geo}(\hat{Y}_{dep}|Y_{2D}), if\ I \in \mathcal{I}_{2D} \end{cases}$$

- Sum all intermediate image features and 2D prediction as input if depth prediction.
- Ground truth 2D coordinates are used to constraint unsupervised depth prediction.

Overall Training target

$$L(\hat{Y}_{HM}, \hat{Y}_{dep}|I) = L_{2D}(\hat{Y}_{HM}, Y_{2D}) + L_{dep}(\hat{Y}_{dep}|I, Y_{2D})$$

Weakly-supervised Geometry Constraint



$$L_{geo}(\hat{Y}_{dep}|Y_{2D}) = \Sigma_i \frac{1}{|R_i|} \Sigma_{e \in R_i} (\frac{l_e}{\bar{l}_e} - \bar{r}_i)^2,$$

where

$$\bar{r}_i = \frac{1}{|R_i|} \Sigma_{e \in R_i} \frac{l_e}{\bar{l}_e}.$$

- Fact: Ratios between bone lengths remain relative fixed in a human skeleton.
- l_e / $\bar{l}_e$: predicted / canonical length of bone e.
- The sequence $\{\frac{l_e}{\bar{l}_e}\}_{e \in R_i}$ should have a variance of 0.
- L_{geo} is continuous and differentiable with respect to Y_{dep}.
- Training: L_{geo} is activated after training the depth regression module on 3D-only data.

Experiments

Supervised 3D human pose estimation on Human 3.6M dataset

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Chen & Ramanan	89.87	97.57	89.98	107.87	107.31	139.17	93.56	136.09
Zhou et al.	87.36	109.31	87.05	103.16	116.18	143.32	106.88	99.78
Metha et al.	59.69	69.74	60.55	68.77	76.36	85.42	59.05	75.04
Pavlakos et al.	58.55	64.56	63.66	**62.43**	66.93	70.74	57.72	62.51
3D/wo geo	73.25	79.17	72.35	83.90	80.25	81.86	69.77	72.74
3D/w geo	72.29	77.15	72.60	81.08	80.81	77.38	68.30	72.85
3D+2D/wo geo	55.17	61.16	**58.12**	71.75	62.54	67.29	54.81	56.38
3D+2D/w geo	**54.82**	**60.70**	58.22	71.41	**62.03**	**65.53**	**53.83**	**55.58**
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkPair	Average
Chen & Ramanan	133.14	240.12	106.65	106.21	87.03	114.05	90.55	114.18
Zhou et al.	124.52	199.23	107.42	118.09	114.23	79.39	97.70	79.9
Metha et al.	96.19	122.92	70.82	68.45	54.41	82.03	59.79	74.14
Pavlakos et al.	76.84	**103.48**	65.73	**61.56**	67.55	**56.38**	59.47	66.92
3D/wo geo	98.41	141.60	80.01	86.31	61.89	76.32	71.47	82.44
3D/w geo	93.52	131.75	79.61	85.10	67.49	76.95	71.99	80.98
3D+2D/wo geo	**74.79**	113.99	64.34	68.78	52.22	63.97	57.31	65.69
3D+2D/w geo	75.20	111.59	**64.15**	66.05	**51.43**	63.22	**55.33**	**64.90**

3D/wo geo	3D/w geo	3D+2D/wo geo	3D+2D/w geo
90.01%	90.57%	90.93%	91.62%

Transferred 3D Human Pose estimation in the wild

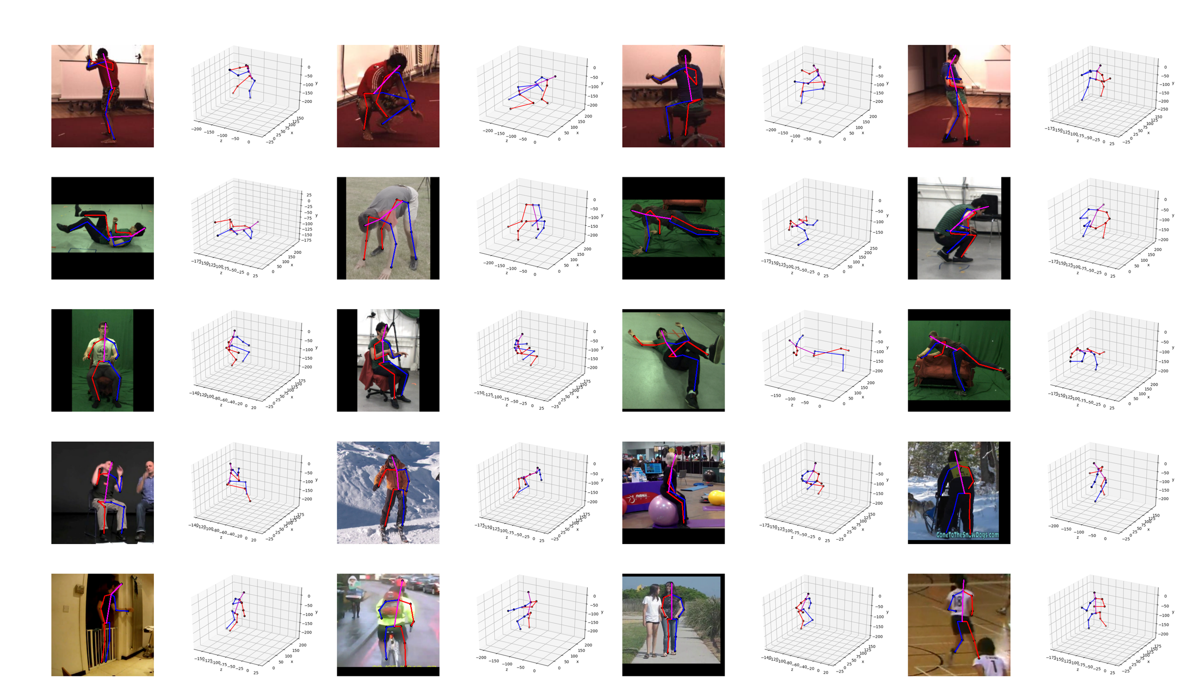
	Studio GS	Studio no GS	Outdoor	ALL PCK	AUC
Metha et al.(H36M+MPII)	70.8	62.3	58.8	64.7	31.7
3D/wo geo	34.4	40.8	13.6	31.5	18.0
3D/w geo	45.6	45.1	14.4	37.7	20.9
3D+2D/wo geo	68.8	61.2	67.5	65.8	32.1
3D+2D/w geo	71.1	64.7	72.7	**69.2**	**32.5**
Metha et al.(MPI-INF-3DHP)	84.1	68.9	59.6	**72.5**	**36.9**

Geometry validity

	3D+2D/wo geo	3D+2D/w geo
Upper arm	42.4mm	**37.8**mm
Lower arm	60.4mm	**50.7**mm
Upper leg	43.5mm	**43.4**mm
Lower leg	59.4mm	**47.8**mm
Upper arm	6.27px	**4.80**px
Lower arm	10.11px	**6.64**px
Upper leg	6.89px	**4.93**px
Lower leg	8.03px	**6.22**px

- State-of-the art performance on supervised 3D task. The benefits are mostly from improved depth regression via shared deep feature representation.
- Transferred performance is close to using the corresponding training data.
- Geometry constraint improves the geometry validity like symmetry.

Qualitative results



Code & Model

https://github.com/xingyizhou/pose-hg-3d