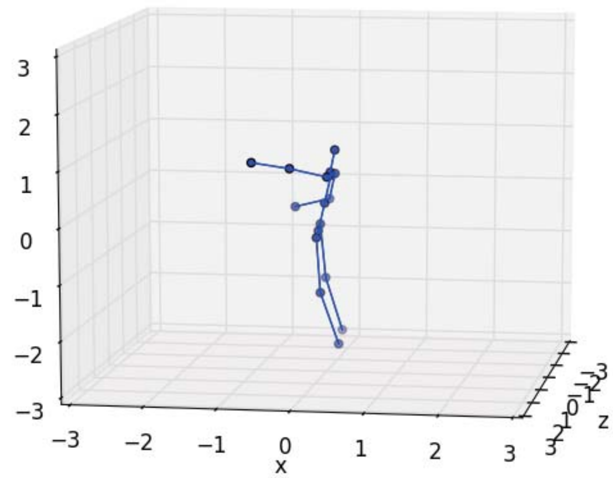
Towards 3D Human Pose Estimation in the Wild: a Weakly-supervised Approach

Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, Yichen Wei
UT Austin & MSRA & Fudan

Human Pose Estimation

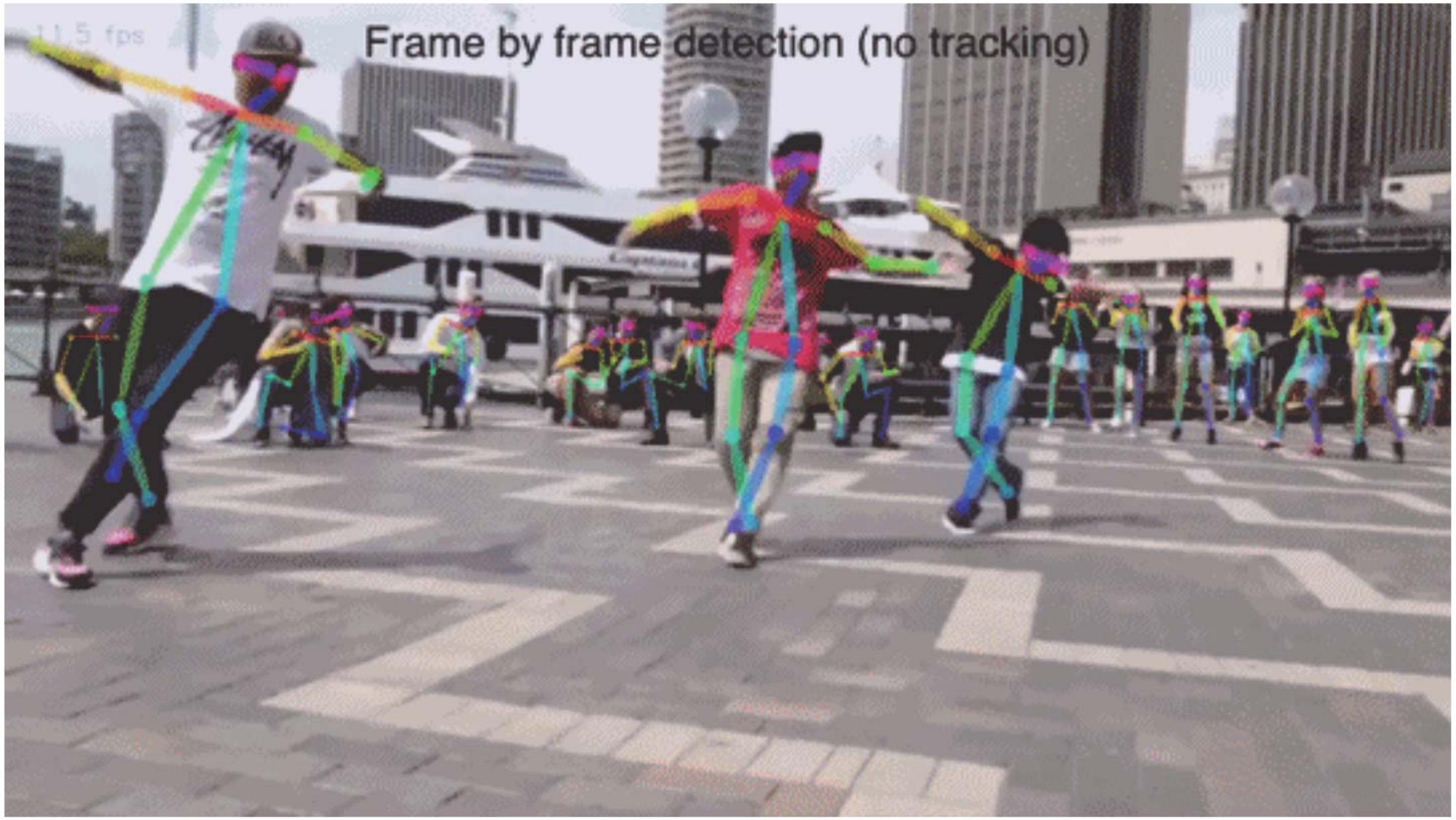
Pose representation



Joint locations

$$y = \{p_1, \cdots, p_N\}$$

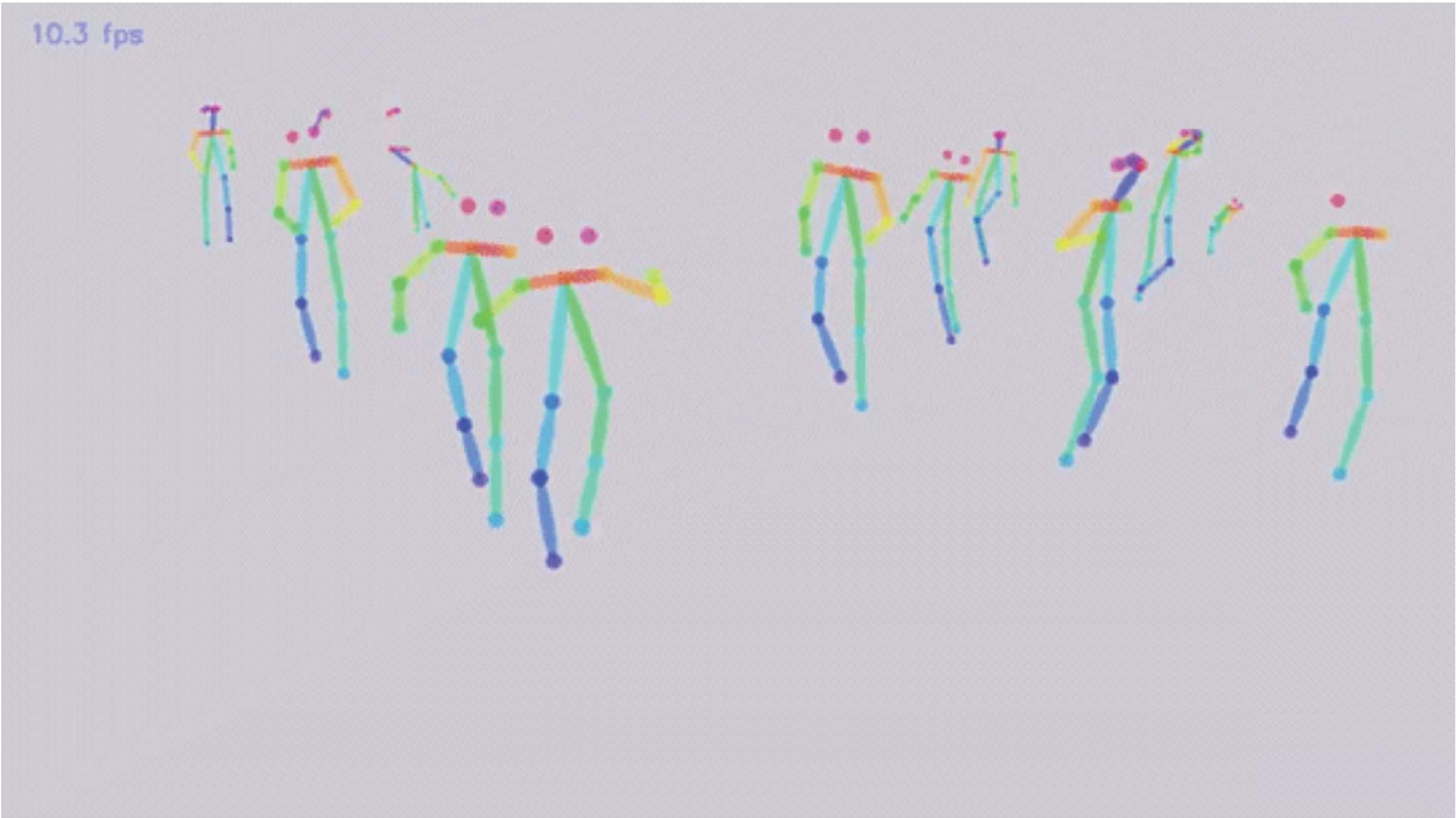
Current Research on 2D Human Pose



- 2D human pose estimation is a well studied problem

Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, CVPR 2017

Is 2D human pose all we need?



- Ambiguous 3D structure

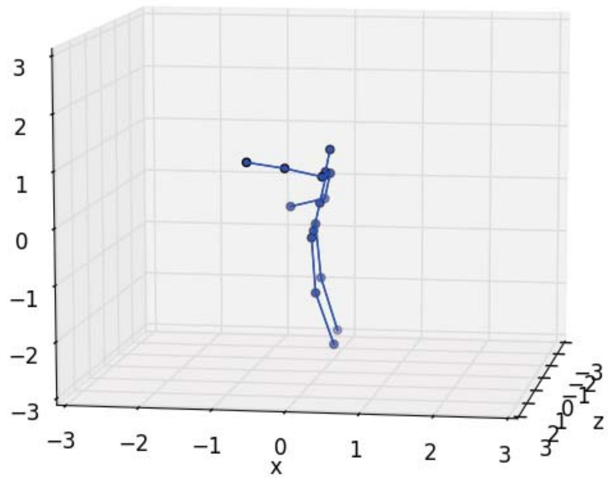
Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, CVPR 2017

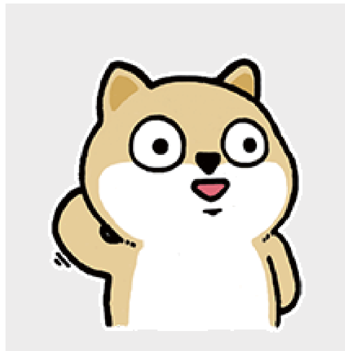
Why we have such a success on 2D?

- 2D human pose data is easy to annotate and largely available

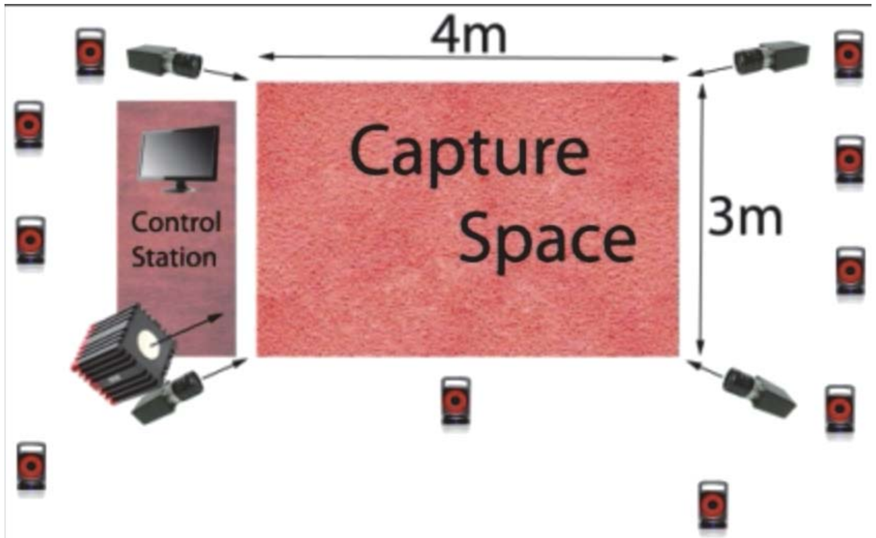
Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, Schiele Bernt, 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, CVPR 2014

3D data not easy to annotate





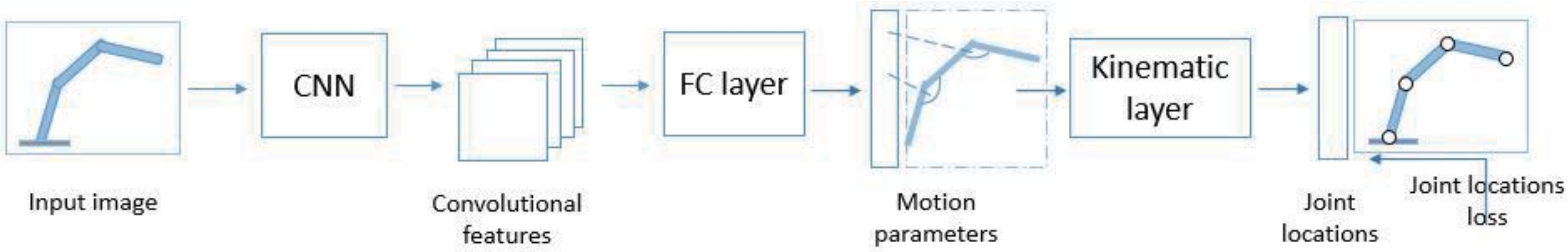
Current 3D human pose data.

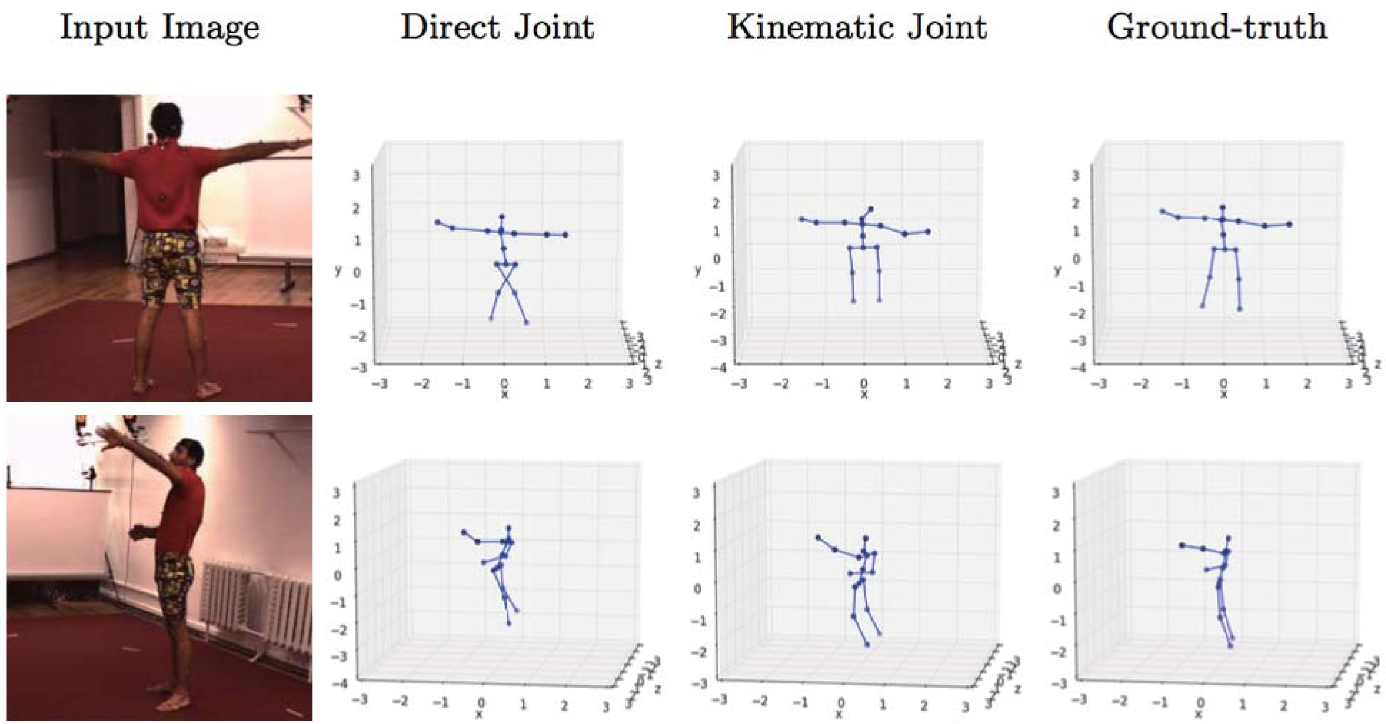


- Captured in control-environment with accurate sensors.

Catalin Ionescu, Dragos Papava, Vlad Olaru and Cristian Sminchisescu, Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, PAMI 2014

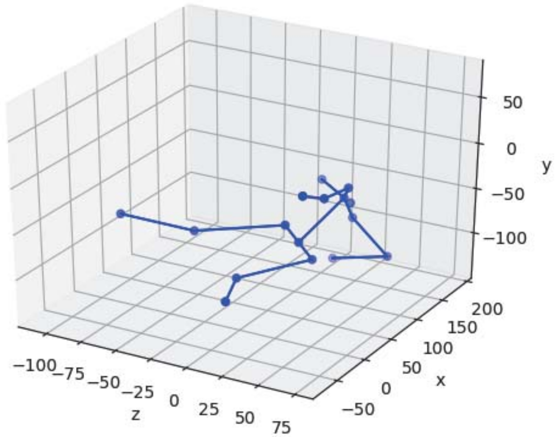
Supervised Pose Regression on Human3.6M





Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, Yichen Wei. Deep Kinematic Pose Regression, In ECCV Workshop on Geometry Meets Deep Learning, 2016

Kinematic Pose Regression-Problems



- Training data is biased to indoor environment

Fail on in-the-wild images!

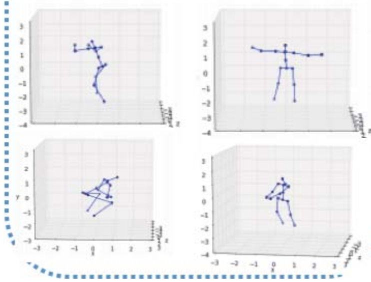
Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, Yichen Wei. Deep Kinematic Pose Regression, In ECCV Workshop on Geometry Meets Deep Learning, 2016

Problem setting

Given:

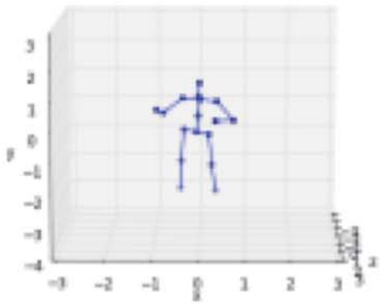
In-the-wild images
with 2D annotation

Indoor images
with 3D annotation



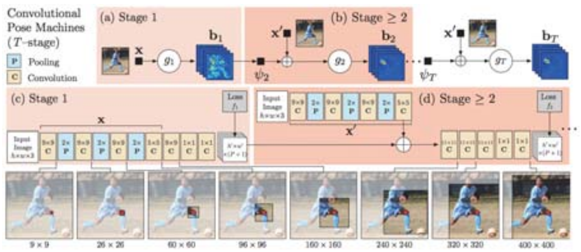
Goal:

In-the-wild image

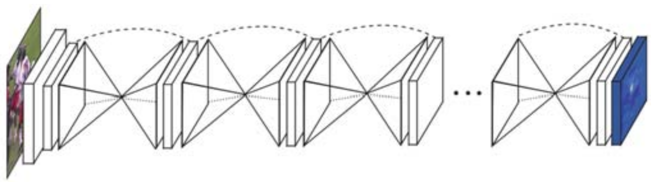


3D pose

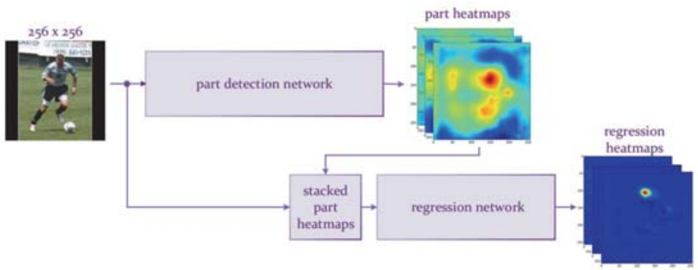
Previous approaches: 2 Stages



Wei et al. Convolutional Pose Machines

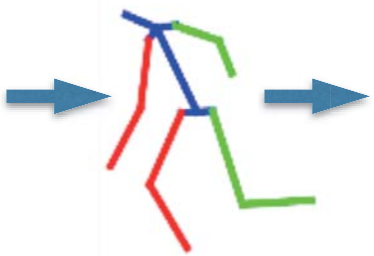


Newell et al. Hourglass Network

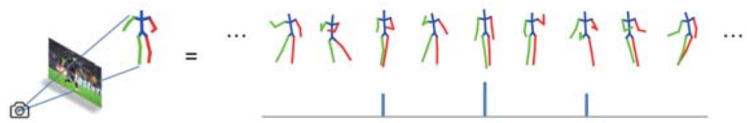


Bulat et al. Part Heatmap Regression

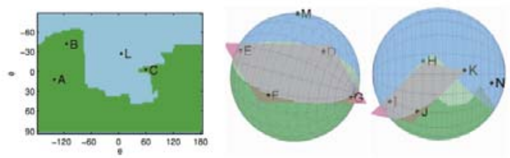
2D pose estimation



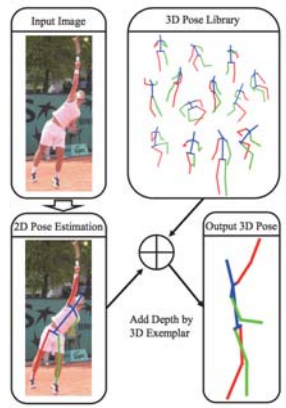
2D pose



Zhou et al. Shape Convex



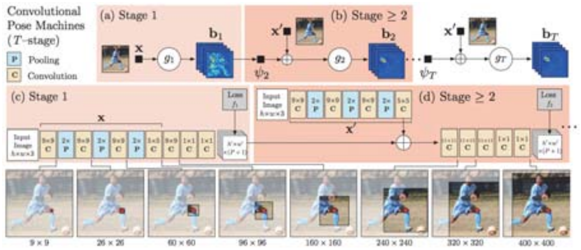
Akhter et al. Pose Conditioned Angle Limits



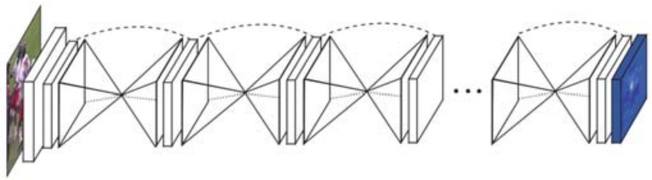
Chen et al. KNN Matching

3D geometry recovery

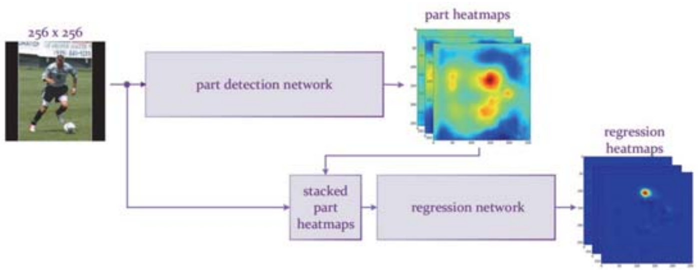
Previous approaches: 2 Stages



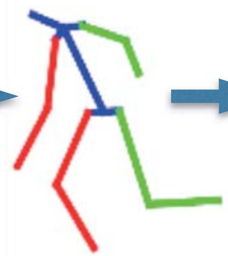
Wei et al. Convolutional Pose Machines



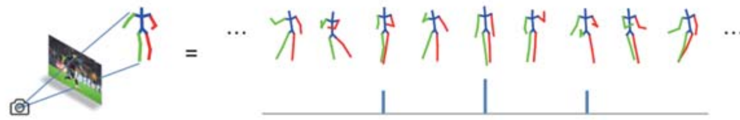
Newell et al. Hourglass Network



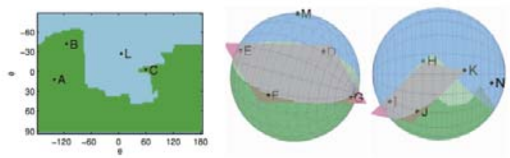
Bulat et al. Part Heatmap Regression



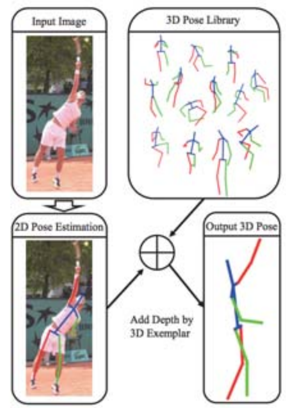
2D pose



Zhou et al. Shape Convex



Akhter et al. Pose Conditioned Angle Limits

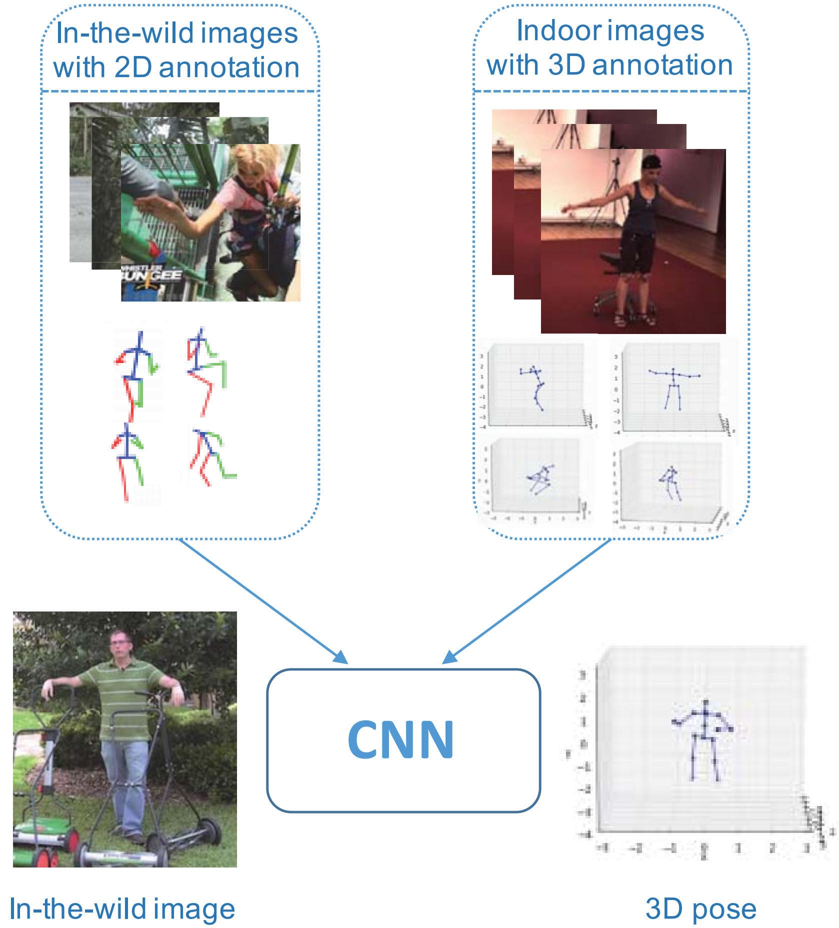


Chen et al. KNN Matching

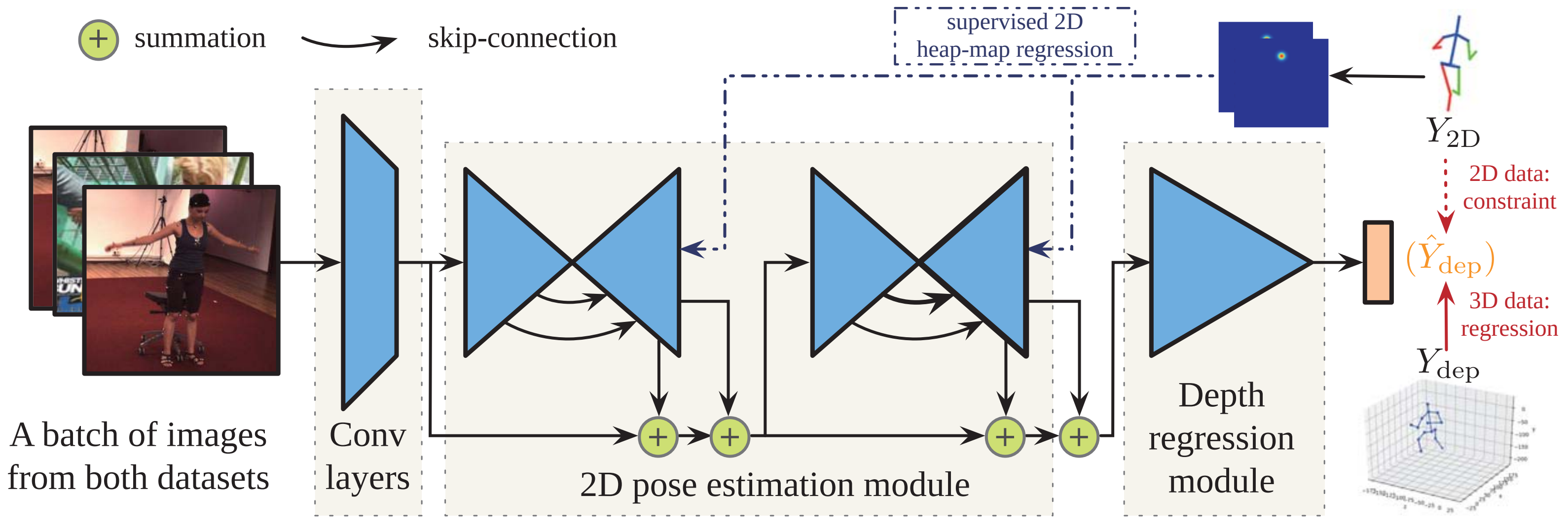
The original in-the-wild 2D image, which contains rich cues for 3D pose recovery, is discarded in the second step. very

Our solution:Weakly-supervised Transfer for 3D Human pose estimation in the wild

- Train a unified neural network using both 2D and 3D annotation.
- 2D and 3D pose are inherently entangled
- 2D-to-3D transfer: provide rich image features
- 3D-to-2D transfer: provide 3D annotation



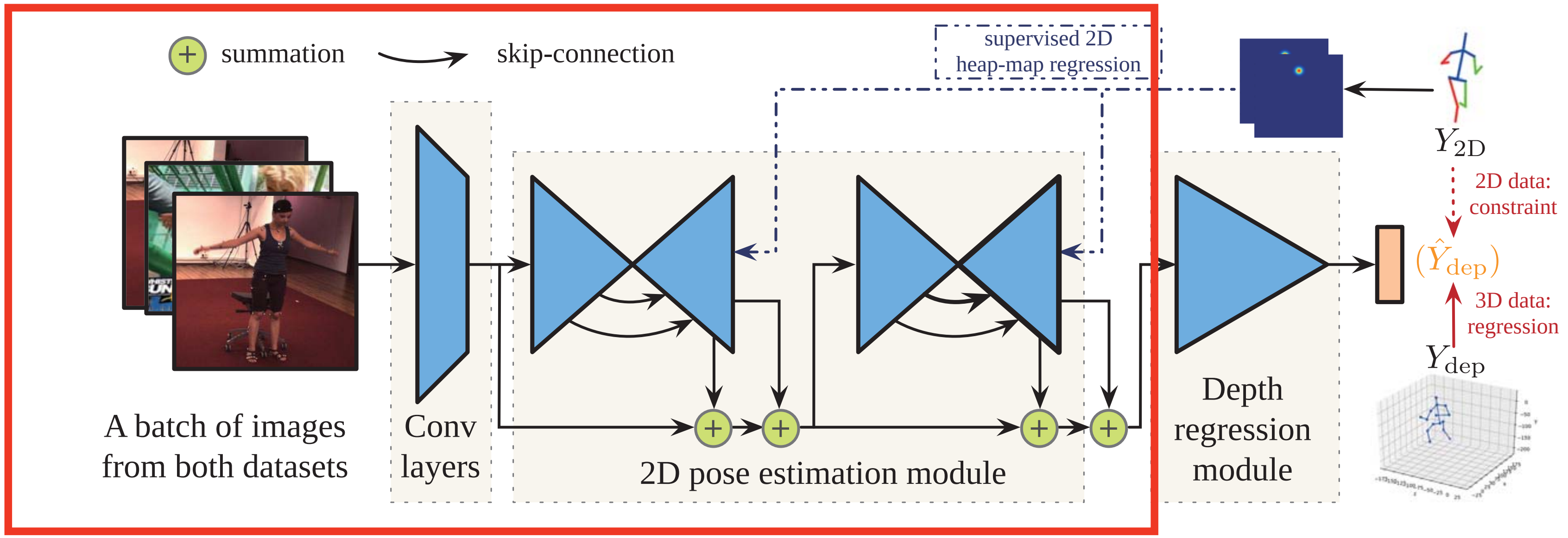
Weakly-supervised Transfer



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

- Images from both dataset are fed into the same mini-batch
- First estimate 2D pose and then regress depth from 2D results and lower layer image features
- Geometry constraint is applied for weakly-labeled 2D data

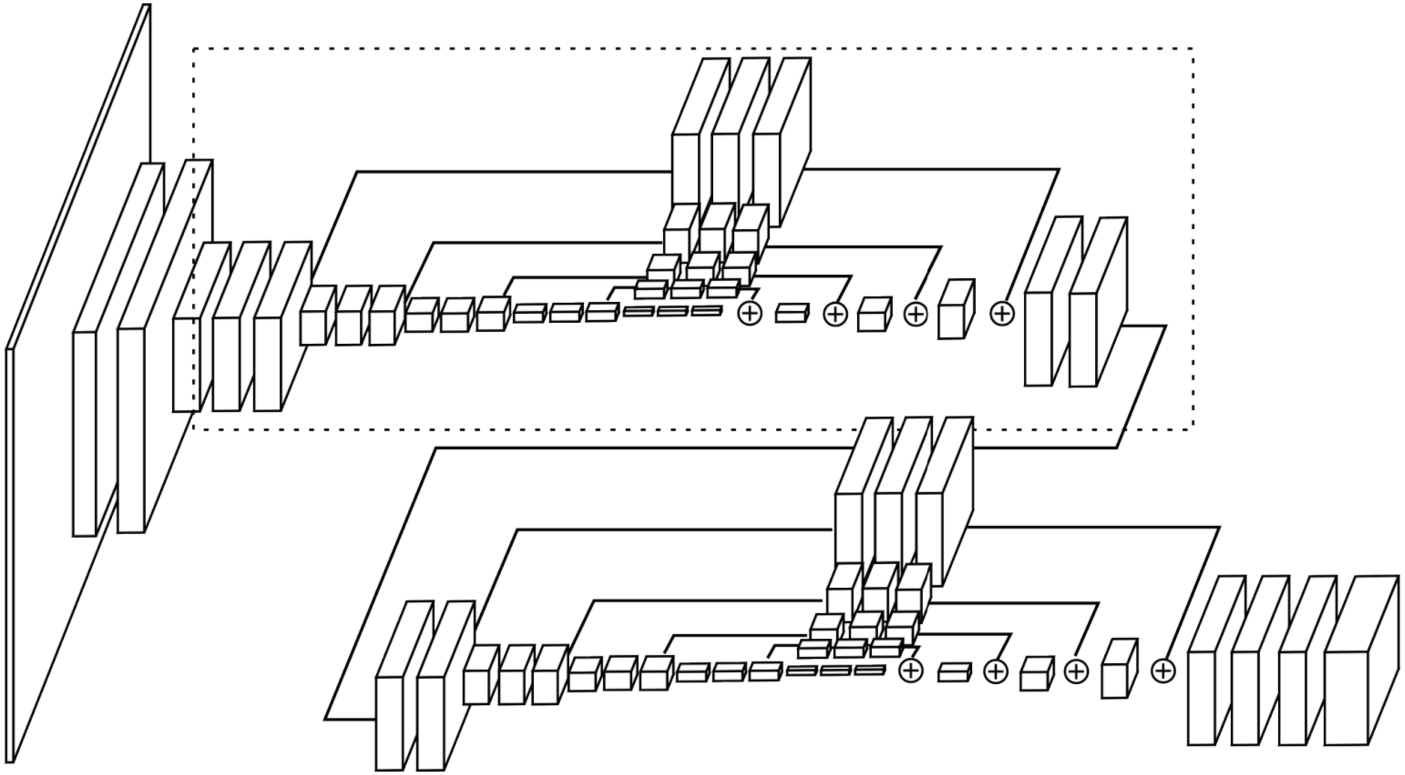
Weakly-supervised Transfer

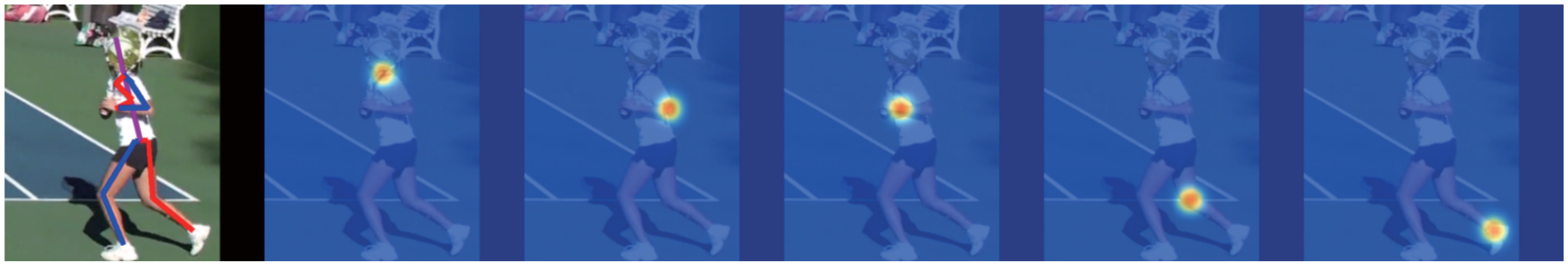


$$\mathcal{S}_{2D} = \{\mathcal{I}_{2D}, \mathcal{Y}_{2D}\} \qquad \mathcal{S}_{3D} = \{\mathcal{I}_{3D}, \mathcal{Y}_{2D}, \mathcal{Y}_{dep}\}$$

- Images from both dataset are fed into the same mini-batch
- First estimate 2D pose and then regress depth from 2D results and lower layer image features
- Geometry constraint is applied for weakly-labeled 2D data

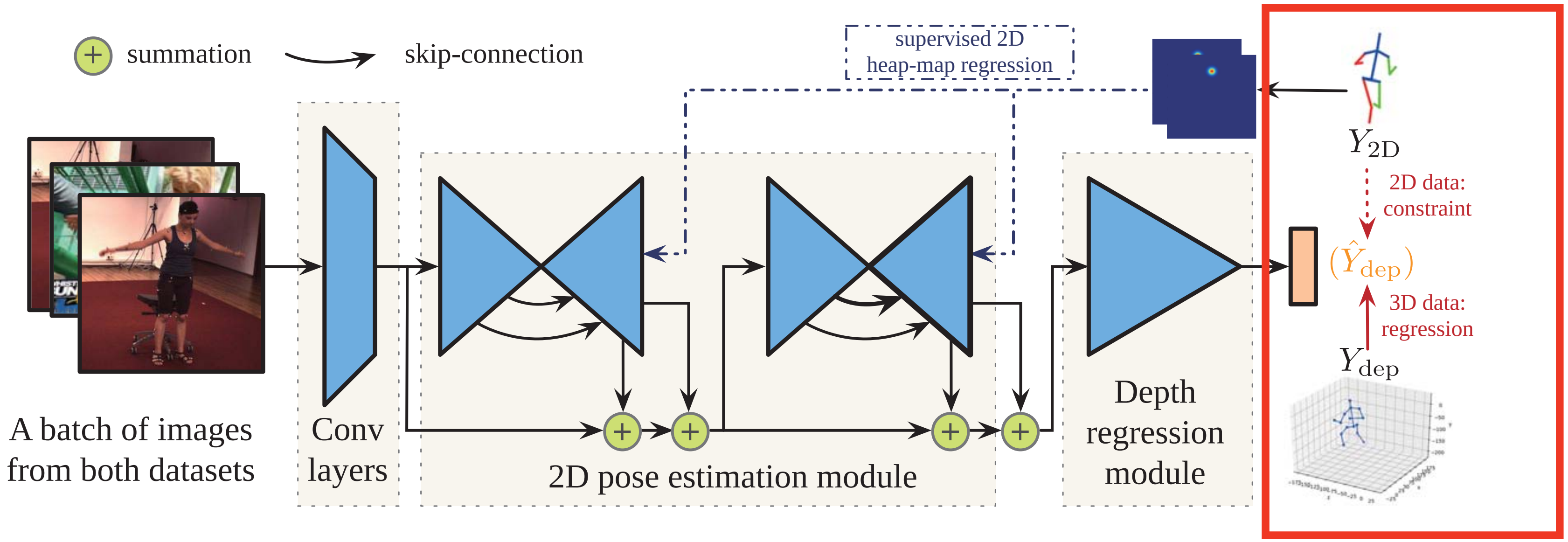
2D Human Pose estimation: HourglassNetwork





Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation, ECCV 2016

Weakly-supervised Transfer

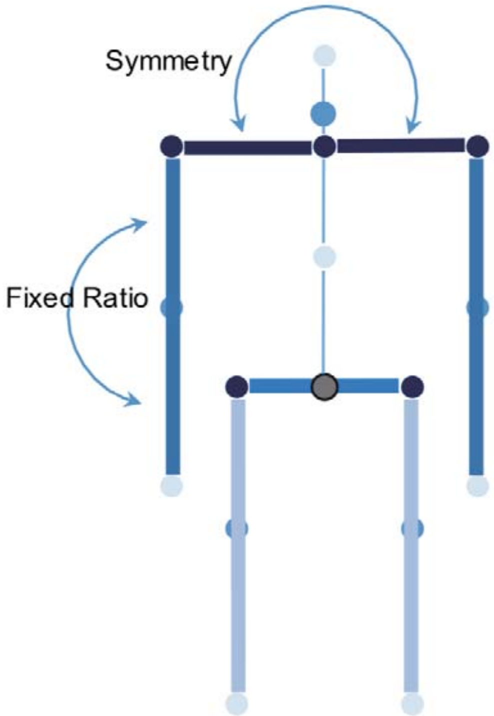


$$\mathcal{S}_{2D} = \{\mathcal{I}_{2D}, \mathcal{Y}_{2D}\} \qquad \mathcal{S}_{3D} = \{\mathcal{I}_{3D}, \mathcal{Y}_{2D}, \mathcal{Y}_{dep}\}$$

- Images from both dataset are fed into the same mini-batch
- First estimate 2D pose and then regress depth from 2D results and lower layer image features
- Geometry constraint is applied for weakly-labeled 2D data

Geometry Constraint

Key idea: Ratios between bone lengths remain relative fixed



R_i : a set of involved bones in a skeleton group

l_e : length of bone e

$\bar{l}_e$: length of bone e in canonical skeleton

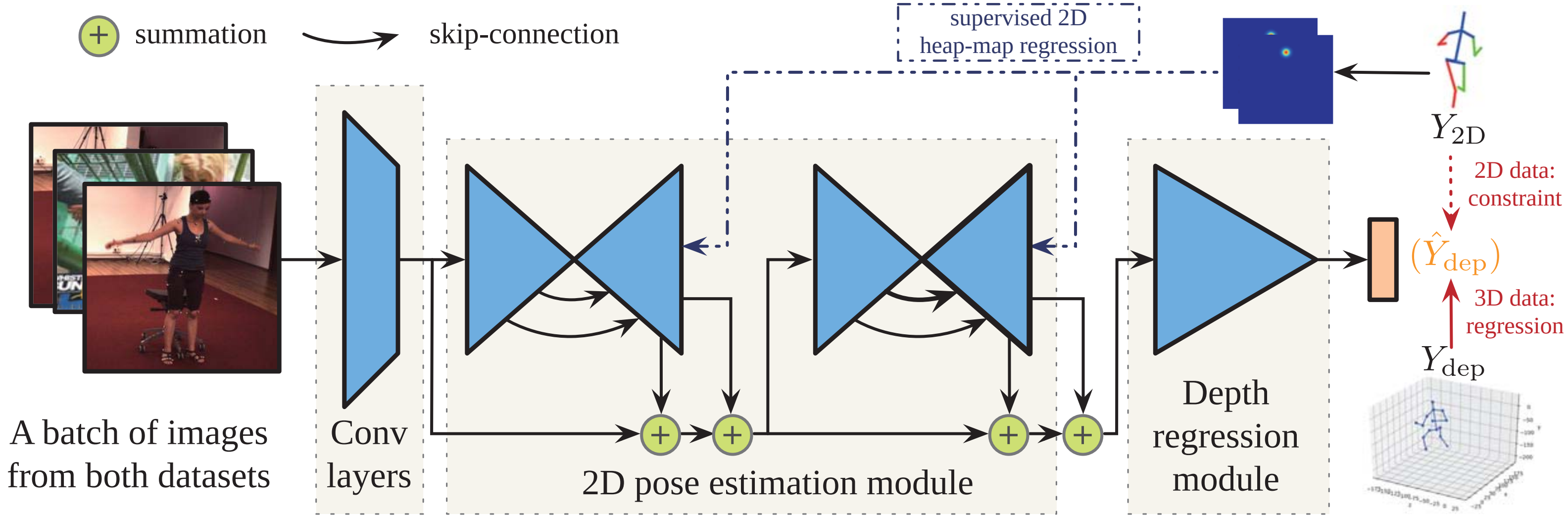
$\{\frac{l_e}{\bar{l}_e}\}_{e\in R_i}$ should remain fixed,
i.e. has zero variance

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

where

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

Weakly-supervised Transfer



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

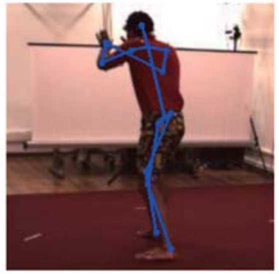
$$L_{dep}(\hat{Y}_{dep}|I, Y_{2D}) = \begin{cases} \lambda_{reg}||Y_{dep} - \hat{Y}_{dep}||^2, if\ I \in \mathcal{I}_{3D} \\ \lambda_{geo}L_{geo}(\hat{Y}_{dep}|Y_{2D}), if\ I \in \mathcal{I}_{2D} \end{cases}$$

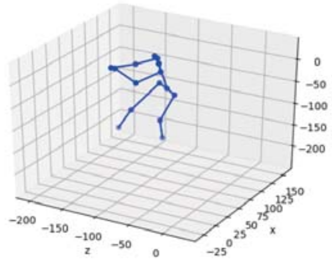
$$L(\hat{Y}_{HM}, \hat{Y}_{dep}|I) = L_{2D}(\hat{Y}_{HM}, Y_{2D}) + L_{dep}(\hat{Y}_{dep}|I, Y_{2D})$$

Evaluation-Datasets

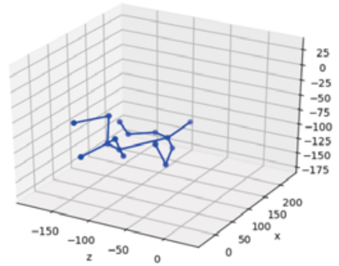
- MPII
  - 2D annotation, in-the-wild images
  - Used for weakly-supervised training

- Human 3.6M
  - MoCap 3D annotation, indoor
  - Used for supervised training

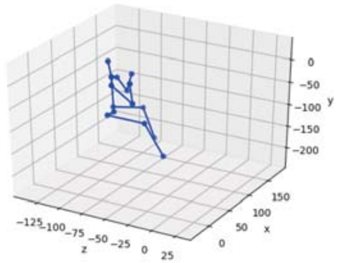




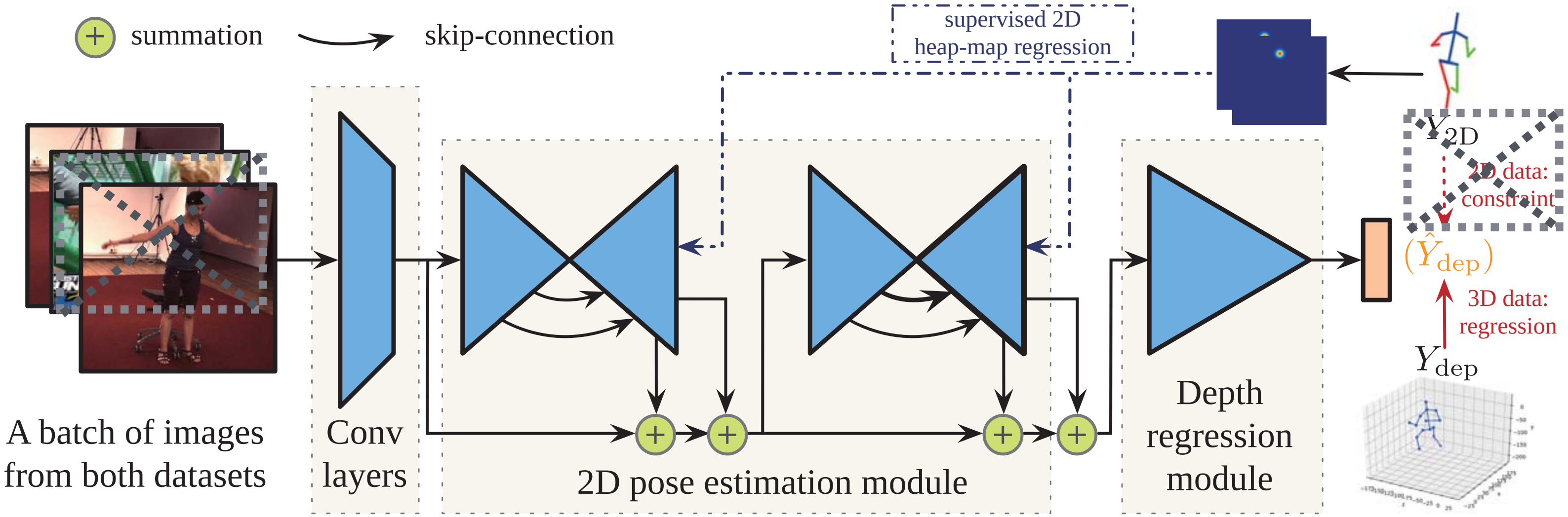
- MPI-INF-3DHP
  - MoCap 3D annotation, indoor & outdoor
  - Used for evaluation



- MPII-Validation
  - Used for evaluation



Evaluation-Baseline setup



	Transfer	Geometry
3D/wo geo	✗	✗
3D/w geo	✗	✓
3D+2D/wo geo	✓	✗
3D+2D/w geo	✓	✓

Table 2. Definition of our baselines. *Transfer* for taking both datasets for training, *Geometry* for the geometry constraint loss.

Supervised 3D pose estimation on Human3.6M dataset

	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkPair	Average
Chen & Ramanan [6]	133.14	240.12	106.65	106.21	87.03	114.05	90.55	114.18
Tome et al. [26]	110.19	172.91	84.95	85.78	86.26	71.36	73.14	88.39
Zhou et al. [35]	124.52	199.23	107.42	118.09	114.23	79.39	97.70	79.9
Metha et al. [16]	96.19	122.92	70.82	68.45	54.41	82.03	59.79	74.14
Pavlakos et al. [20]	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**

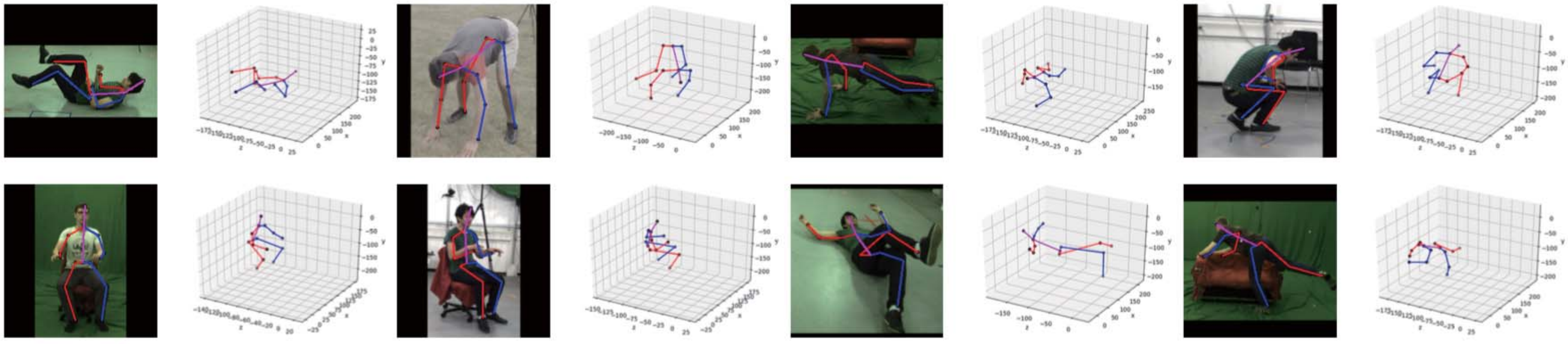
- **3D/wo geo** (82.44mm) shows the effectiveness of our architecture.
- **3D/w geo** shows the geo-constraint is consistent with supervision.
- Training with 3D&2D data (**3D+2D/wo geo**) provides great performance gain.
- Weakly supervised constraint **3D+2D/w geo** brings further improvements.
- Only 2-steps methods Chen & Ramanan(114.18mm) and Zhou et al,(79.9mm) can be applied in-the-wild.

Results Analysis

	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkPair	Average	2D PCK
Chen & Ramanan [6]	133.14	240.12	106.65	106.21	87.03	114.05	90.55	114.18	
Tome et al. [26]	110.19	172.91	84.95	85.78	86.26	71.36	73.14	88.39	
Zhou et al. [35]	124.52	199.23	107.42	118.09	114.23	79.39	97.70	79.9	
Metha et al. [16]	96.19	122.92	70.82	68.45	54.41	82.03	59.79	74.14	
Pavlakos et al. [20]	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	90.01%
3D/w geo	93.52	131.75	79.61	85.10	67.49	76.95	71.99	80.98	90.57%
3D+2D/wo geo	**74.79**	113.99	64.34	68.78	52.22	63.97	57.31	65.69	90.93%
3D+2D/w geo	75.20	111.59	**64.15**	66.05	**51.43**	63.22	**55.33**	**64.90**	91.62%

- Is the improvement from more accurate 2D position or better depth estimation?
  - All baselines have very high 2D pose estimation.
  - This indicates that depth estimation are greatly benefit from more 2D data.
  - 2-stage approaches can not have such benefit.

In-the-wild 3D pose estimation on MPII-INF-3DHP Dataset

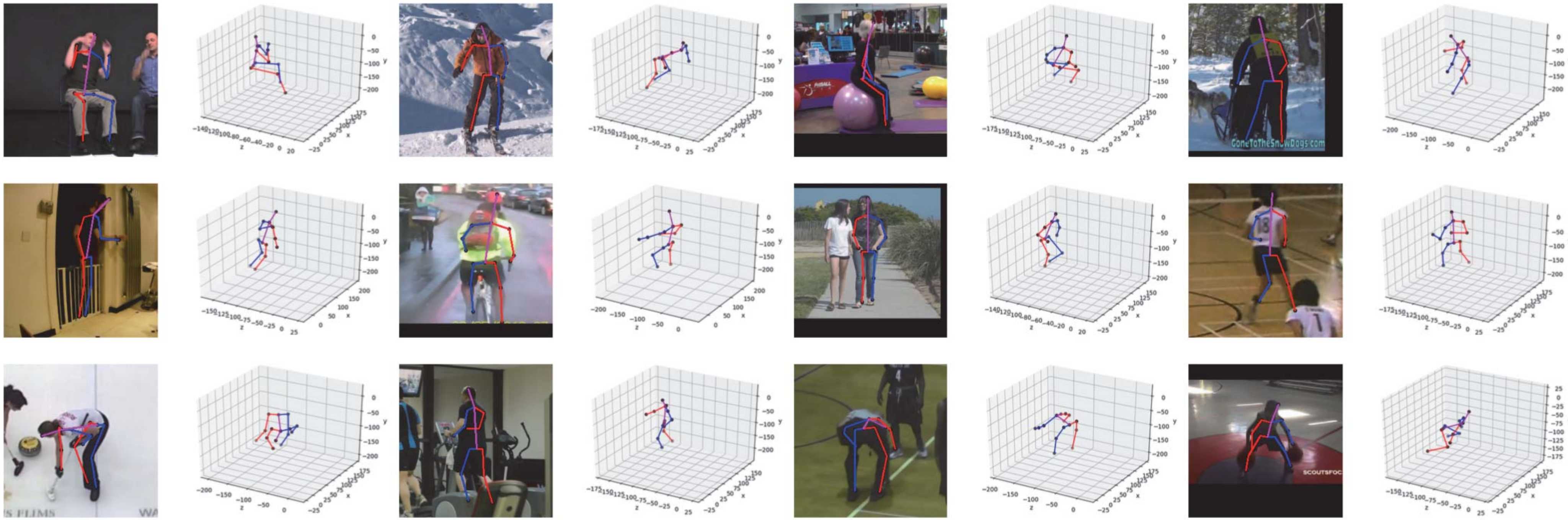


	Stand/Walk	Exercise	Chair	Reach	Ground	Sport	Misc	Total PCK	AUC
Metha et al.(H36M+MPII) [16]	76.4	62.9	58.1	57.4	27.8	66.9	65.6	61.0	28.3
3D/wo geo	28.6	41.2	41.4	34.3	19.7	36.4	36.4	31.5	18.0
3D/w geo	37.0	44.5	45.4	38.8	22.9	50.1	30.8	37.7	20.9
3D+2D/wo geo	82.3	66.4	60.3	69.2	37.1	65.7	67.8	65.8	32.1
3D+2D/w geo	85.4	71.0	60.7	71.4	37.8	70.9	74.4	**69.2**	**32.5**
Metha et al.(MPI-INF-3DHP) [16]	85.0	70.1	72.7	65.2	47.0	79.0	70.3	**70.8**	**35.9**

Table 2. Results of MPI-INF-3DHP Dataset. The results are shown in PCK and AUC.

- 3D data-only methods fail on in-the-wild images.
- **3D+2D/wo geo** wins its counterpart of Metha et al.
- Geo-constraint provides further improvements, whose results are close to training on the corresponding training set.

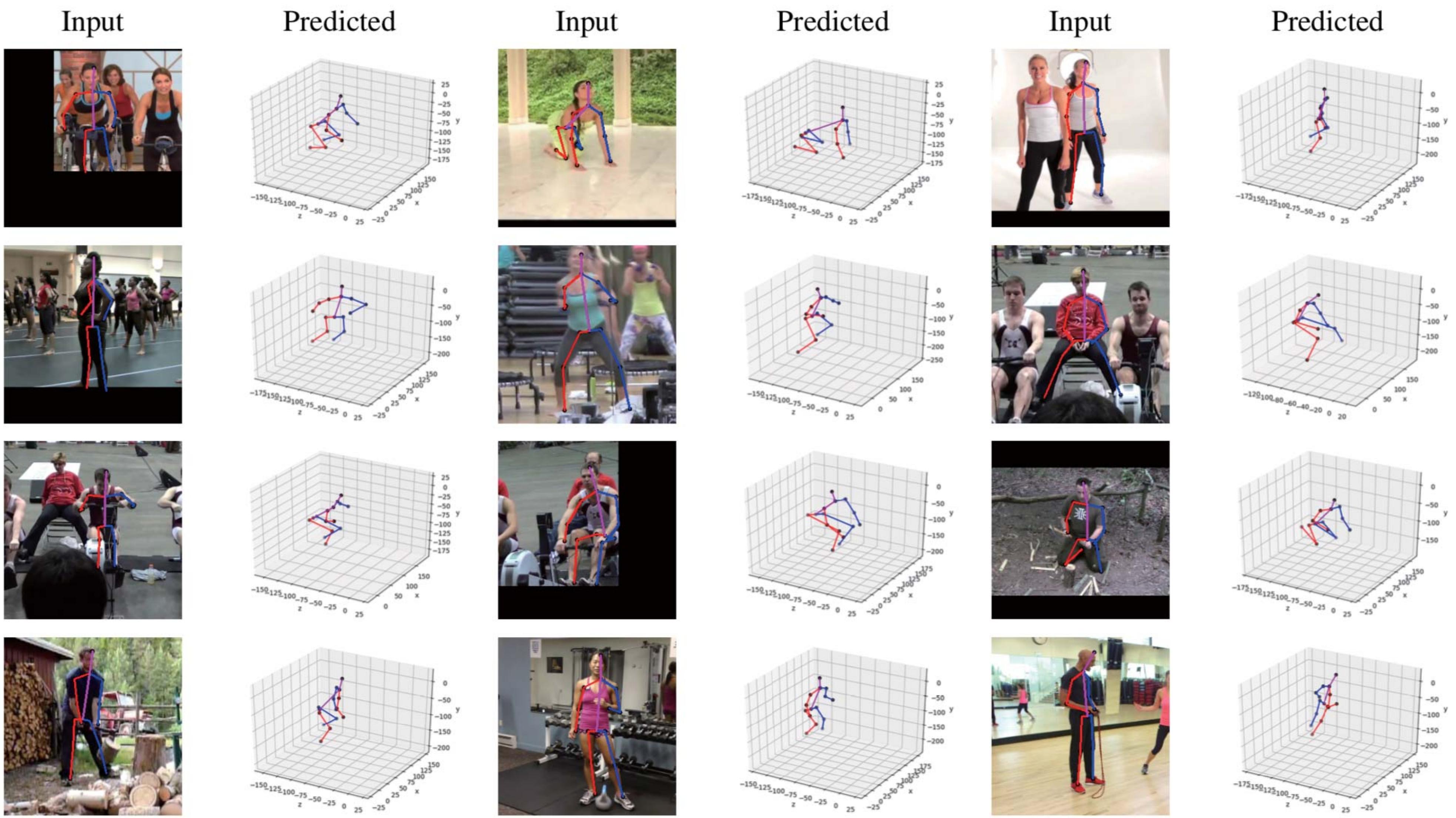
In-the-wild 3D pose estimation on MPII-Validation-3D Set



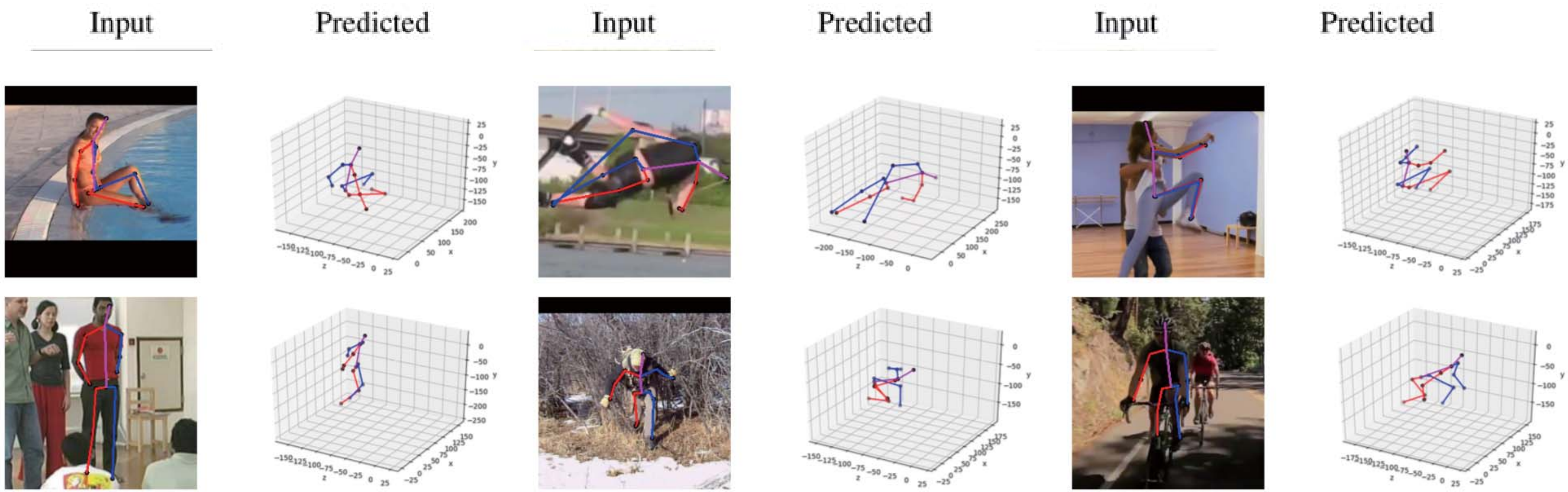
- **3D+2D/w geo** performs better and correct the symmetry invalidity.
- Our framework keeps 2D accuracy.

	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

More qualitative results

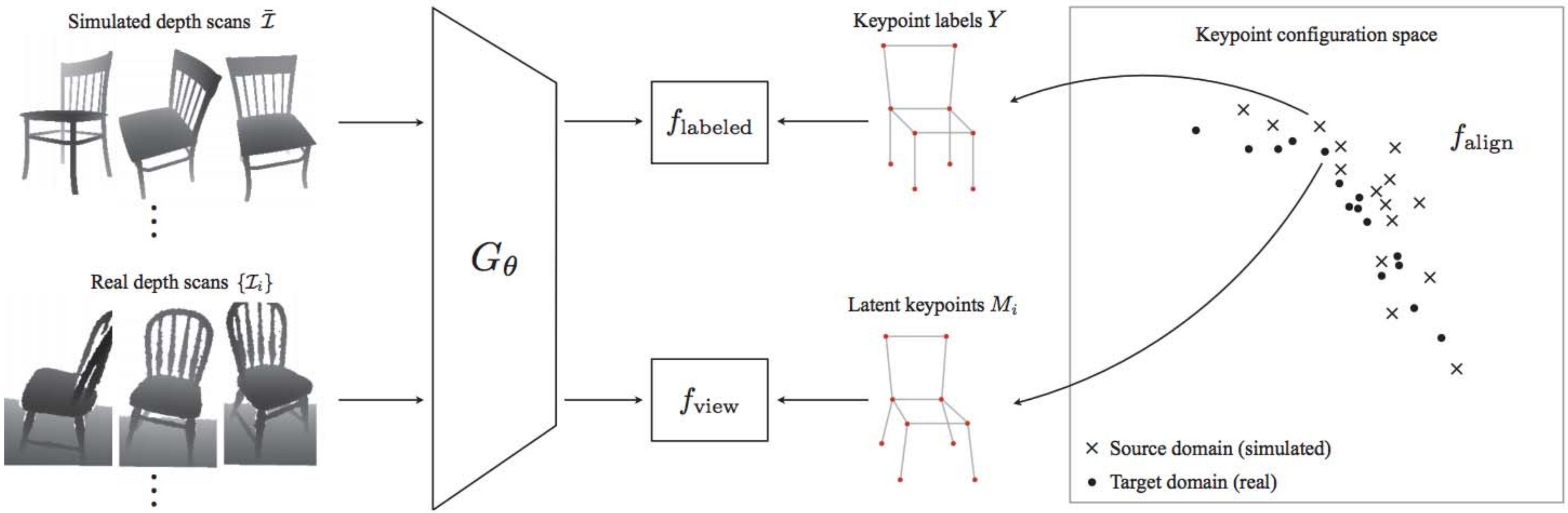


Failure Cases



inaccurate 2D prediction/ ambiguous depth/ false torso length.

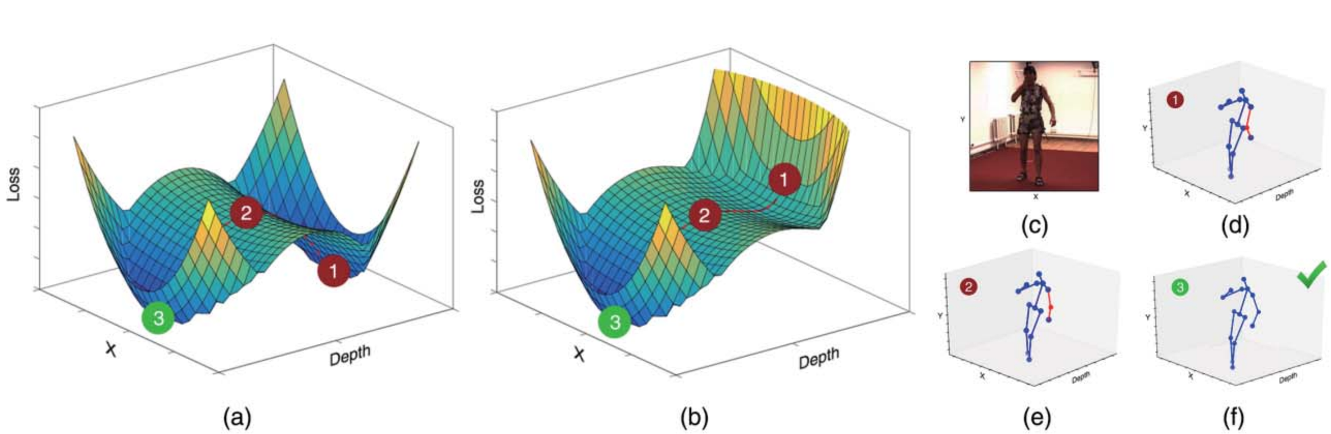
Extension

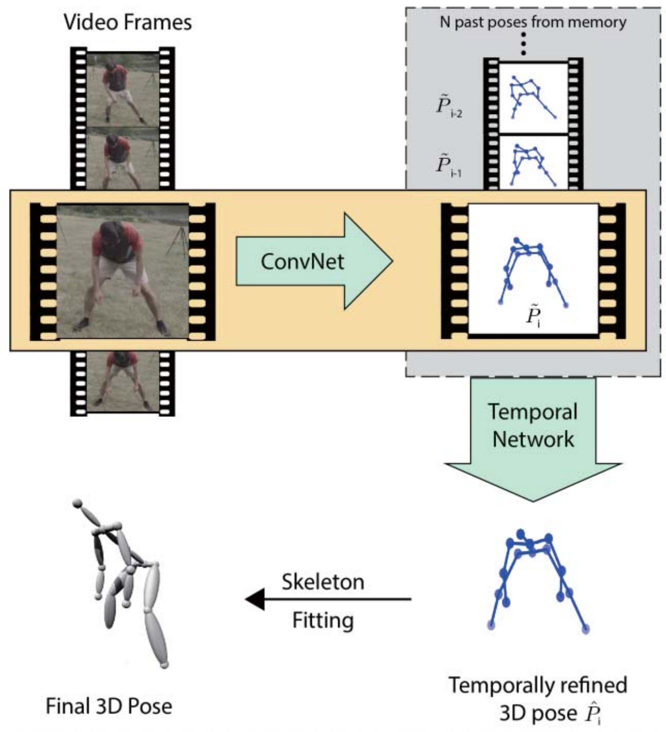


- An improved weak-supervision for rigid objects.
- The predicted pose of the same object from different viewpoint should be consistent with each other.

Xingyi Zhou, Arjun Karpur, Chuang Gan, Linjie Luo, Qixing Huang, Unsupervised Domain Adaptation for 3D Keypoint Prediction from a Single Depth Scan, arXiv 1712.05765, 2017

Extension





- Add temporal refinement.
- Add angle constraint.

Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Arjun Jain, Structure-Aware and Temporally Coherent 3D Human Pose Estimation, arXiv:1711.09250

Demo

Q & A

Code & Model Available!

Torch

PyTorch