Korea University
Pattern Recognition &
Machine Learning Lab

ACCV 2022 Macau

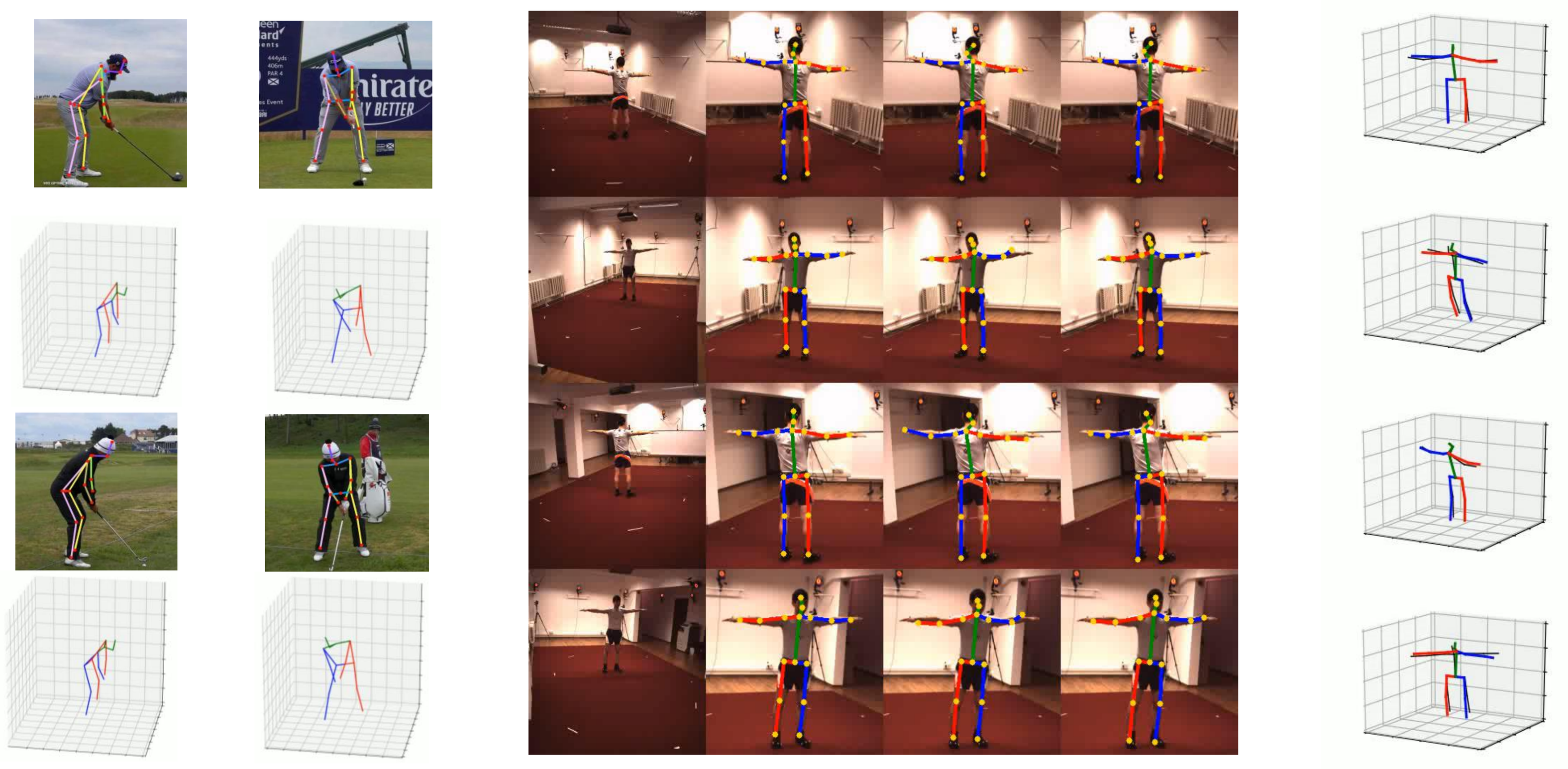
Cross-View Self-Fusion for Self-Supervised 3D Human Pose Estimation in the Wild

Hyun-Woo Kim, Gun-Hee Lee, Myeong-Seok Oh, and Seong-Whan Lee

Department of Artificial Intelligence, Korea University

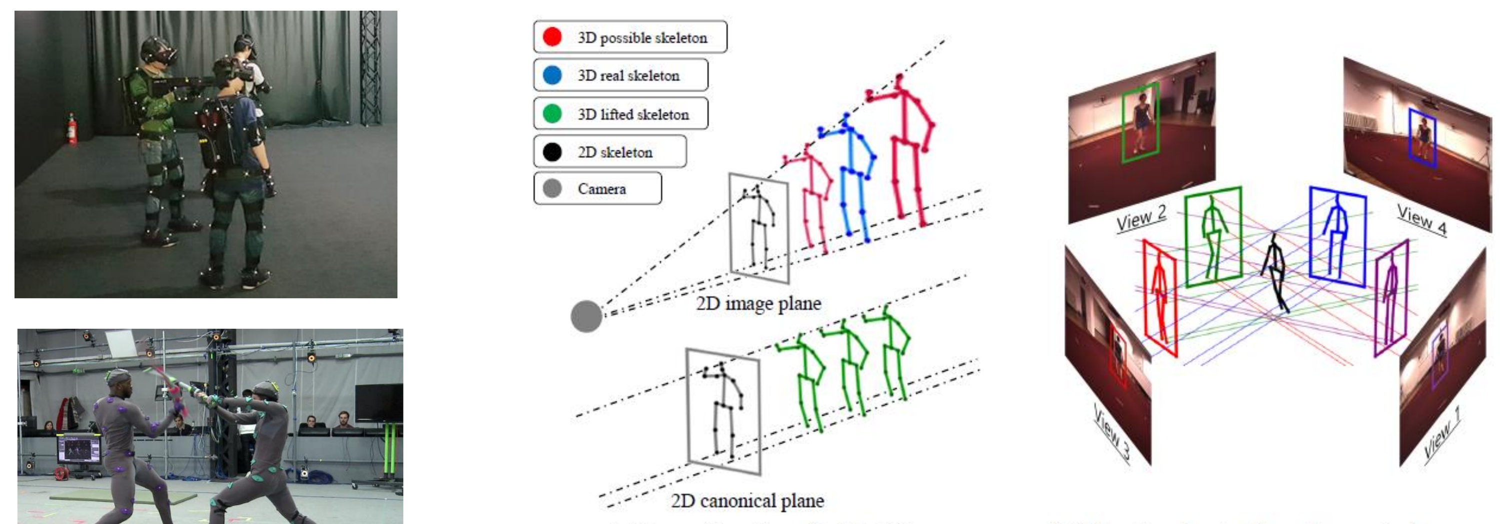
Goal

- ✓ We propose a self-supervised approach that learns a monocular 3D human pose estimation from unlabeled multi-view data without any camera calibrations.
- ✓ Our goal is to train a network without any additional information on the any images captured in a spatially unconstrained in-the-wild environment.



Introduction

- ✓ Collecting 3D annotated data is expensive and mostly limited to fully controlled indoor settings that require motion capture systems.
- ✓ To solve the problem, there are 3D human pose estimation methods using multi-view without 3D annotation, but they require parameters of each camera to use multi-view information.
- ✓ We don't need to estimate the parameters of all cameras by using the canonical form and normalizing the scale and position of 2D poses observed in the different views.

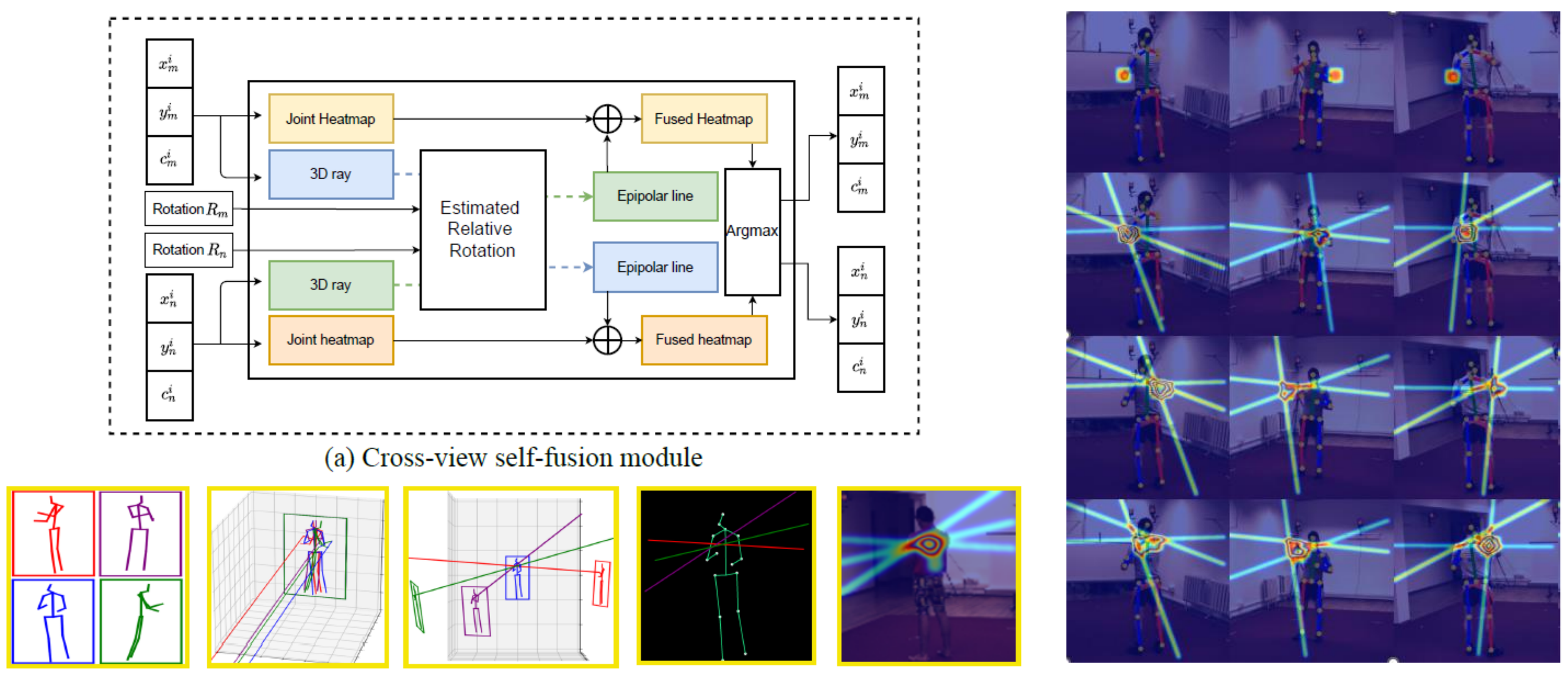


(a) Examples of results lifted from general camera (top) and canonical camera (bottom)

(b) Visual understanding of canonical space

Cross-view Self-fusion Module

- ✓ We propose a Cross-view Self-fusion Module (CSM) that refines an incorrect 2D pose using the input 2D poses and predicted rotations.
- ✓ 2D pose errors not only propagate to the 3D prediction, but also may affect the multi-view consistency requirement during training, which can yield an inaccurate camera rotation estimation.



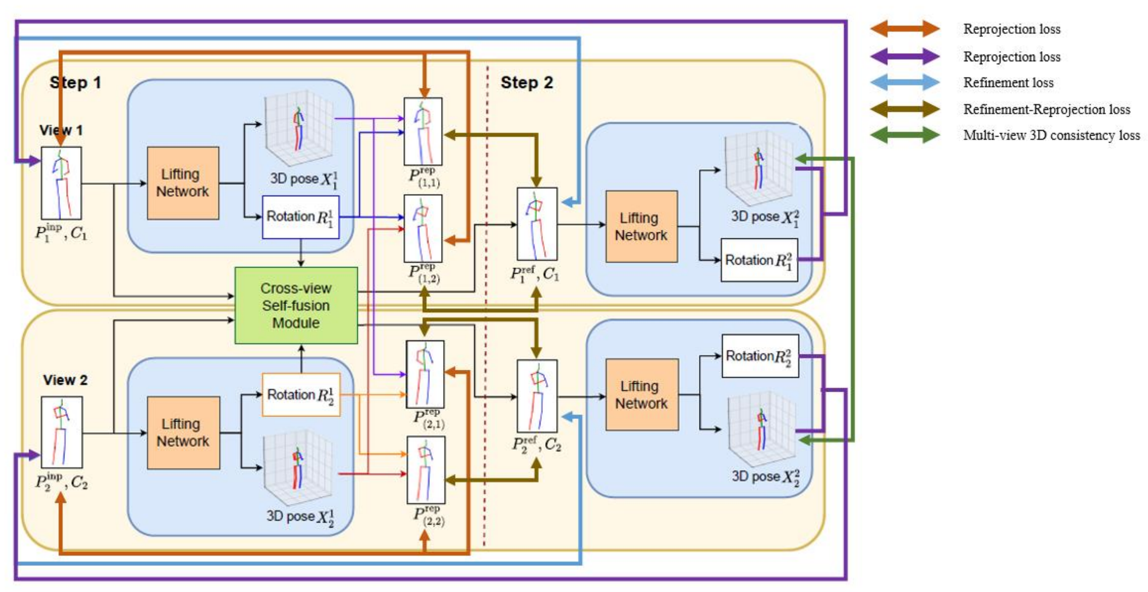
(a) Cross-view self-fusion module

1. 2D poses 2. 3D rays 3. Rotated 3D rays 4. Epipolar lines 5. Fused heatmap

(b) Process of cross-view self-fusion

Two step – two stage training strategy

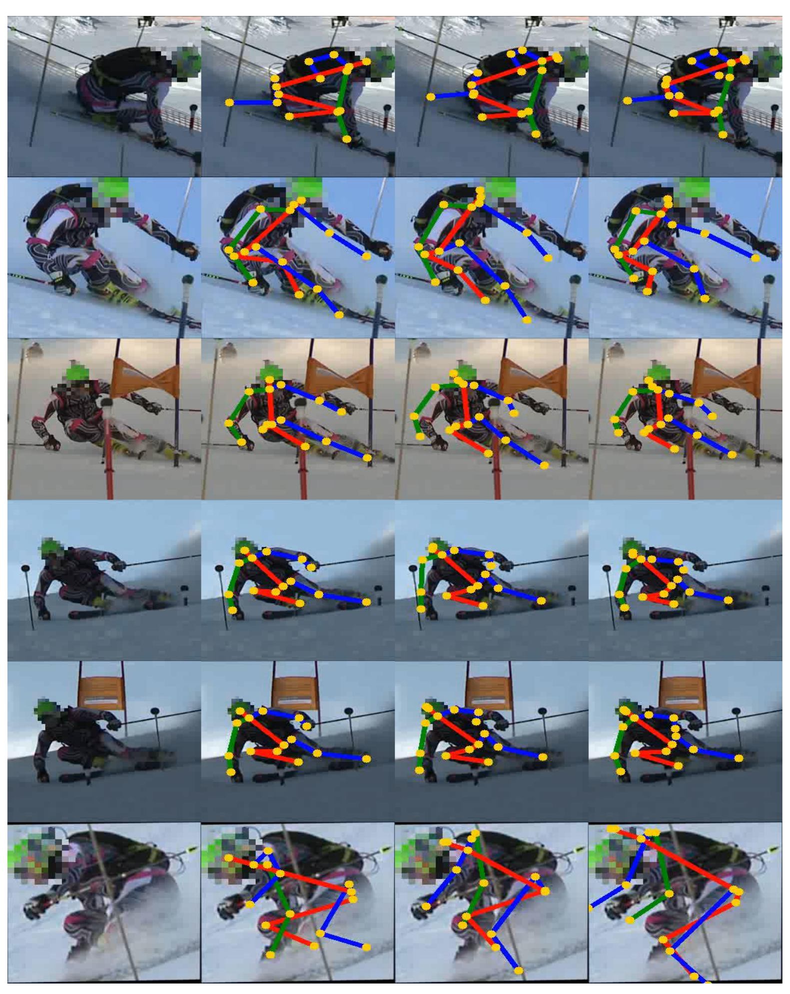
- Step 1
- ✓ A reprojection loss between all combinations of rotations and 3D poses estimated in each view and input 2D poses is defined
- ✓ A refinement loss between refined 2D poses and input 2D poses is defined because the refined 2D poses are affected by initial incorrect rotation estimation
- Step 2
- ✓ Step 2 performs the same process as Step 1 using refined 2D poses
- ✓ We were inspired by the lift-reproject-lift processing [Chen at al. 2019]
- ✓ We add a multi-view 3D consistency loss between the 3D poses estimated in Step 2
- ✓ Weight of each loss is set differently for each stage



Results

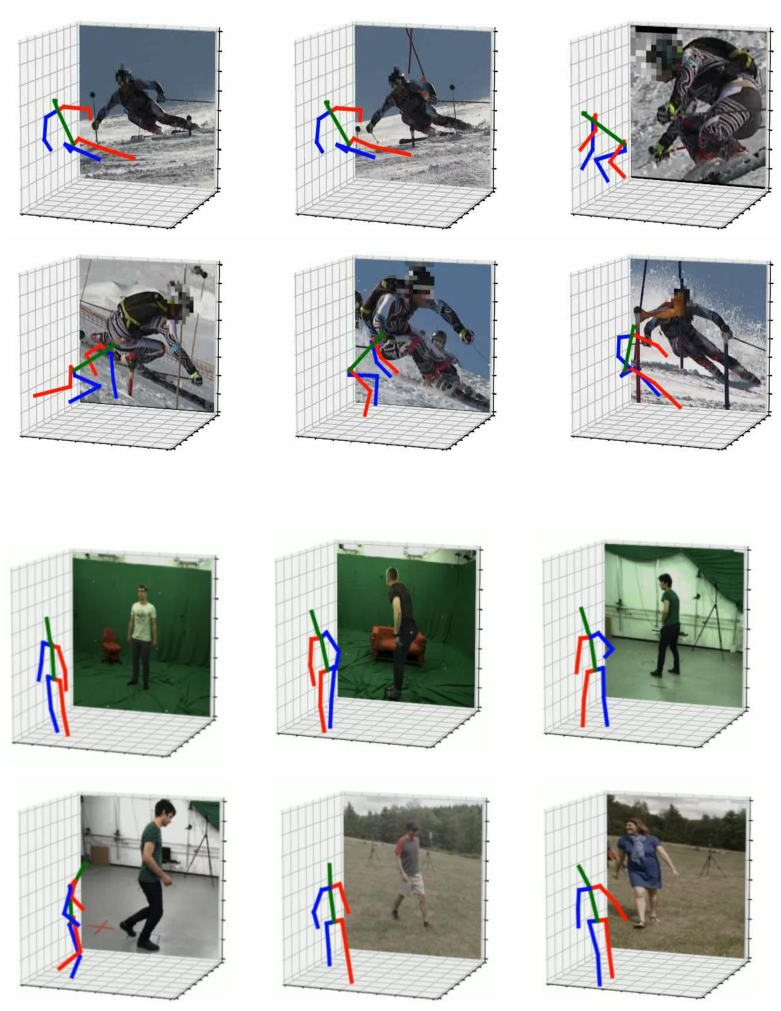
Evaluation of 2D pose refinement accuracy for each dataset. We show JDR for six important joints about each dataset

Method	Dataset	Hip	Knee	Ankle	Shoulder	Elbow	Wrist
Single	H36M	97.1	97.5	97.5	98.5	96.7	98.2
Ours	H36M	**98.2**	**98.5**	**97.8**	**98.9**	**98.5**	**99.6**
Single	3DHP	97.4	97.8	99.8	96.9	97.0	96.9
Ours	3DHP	**98.8**	**97.8**	**99.9**	**98.4**	**98.4**	**98.3**
Single	Ski	97.0	73.7	**81.2**	90.0	70.0	**60.9**
Ours	Ski	**98.7**	**77.0**	75.1	**91.9**	**71.7**	56.4



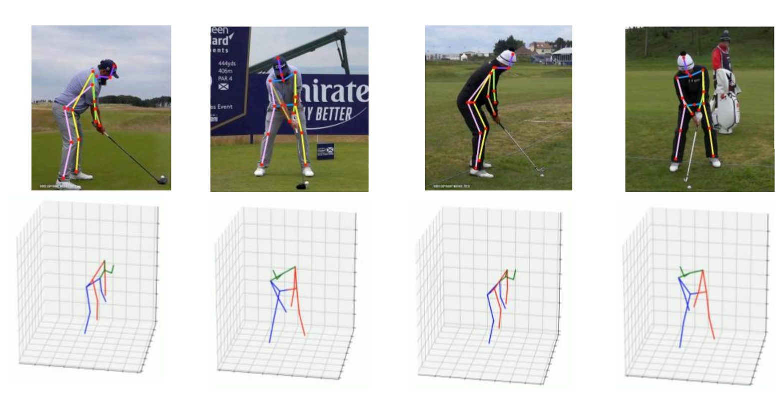
Evaluation of results on the SkiPose. NMPJPE and PMPJPE are given in *mm*, N-PCK is in %. The best results are marked in bold.

Supervision	Method	NMPJPE ↓	PMPJPE ↓	N-PCK ↑
Weak	Rhodin [53]	**85.0**	-	**72.7**
Self	Wandt [12]	128.1	89.6	67.1
	Ours (S1)	118.2	79.3	70.1
	Ours (S1+S2)	115.2	**78.8**	72.4



Evaluation of results on the MPI-INF-3DHP. NMPJPE and PMPJPE are reported in millimeters, and N-PCK is in %. The best results are marked in bold

Supervision	Method	NMPJPE ↓	PMPJPE ↓	N-PCK ↑
Weak	Rhodin [53]	121.8	-	72.7
	Kolotouros [57]	124.8	-	66.8
	Li [59]	-	-	74.1
	Kundu [58]	103.8	-	**82.1**
Self	Kocabas [9]	125.7	-	64.7
	Iqbal [11]	110.1	68.7	76.5
	Wandt [12]	104.0	70.3	77.0
	Ours (S1)	95.2	57.3	79.3
	Ours (S1+S2)	**94.6**	**56.5**	81.9



Evaluation of results on the Human3.6M and comparison of the NMPJPE and PMPJPE (*mm*). The best results are marked in bold. Our model outperforms all self-supervised methods

Supervision	Method	NMPJPE ↓	PMPJPE ↓
Full	Martinez [34]	67.5	52.5
Weak	Rhodin [55]	122.6	98.2
	Rhodin [53]	80.1	65.1
	Wandt [56]	89.9	65.1
	Kolotouros [57]	-	62.0
	Kundu [58]	85.8	-
Self	Kocabas [9]	76.6	67.5
	Jenni [10]	89.6	76.9
	Iqbal [11]	69.1	55.9
	Wandt [12]	74.3	53.0
	Ours (S1)	63.6	46.1
	Ours (S1+S2)	**61.4**	**45.9**

Conclusion

In this paper, we introduced a novel self-supervised learning method for monocular 3D human pose estimation from unlabeled multi-view images without camera calibration. We exploited multi-view consistency to disentangle 2D estimations into canonical predictions (a 3D pose and camera rotation) that were used to refine the errors of the 2D estimations and reproject the 3D pose on the 2D for self-supervised learning.