On Triangulation as a Form of Self-Supervision for 3D Human Pose Estimation

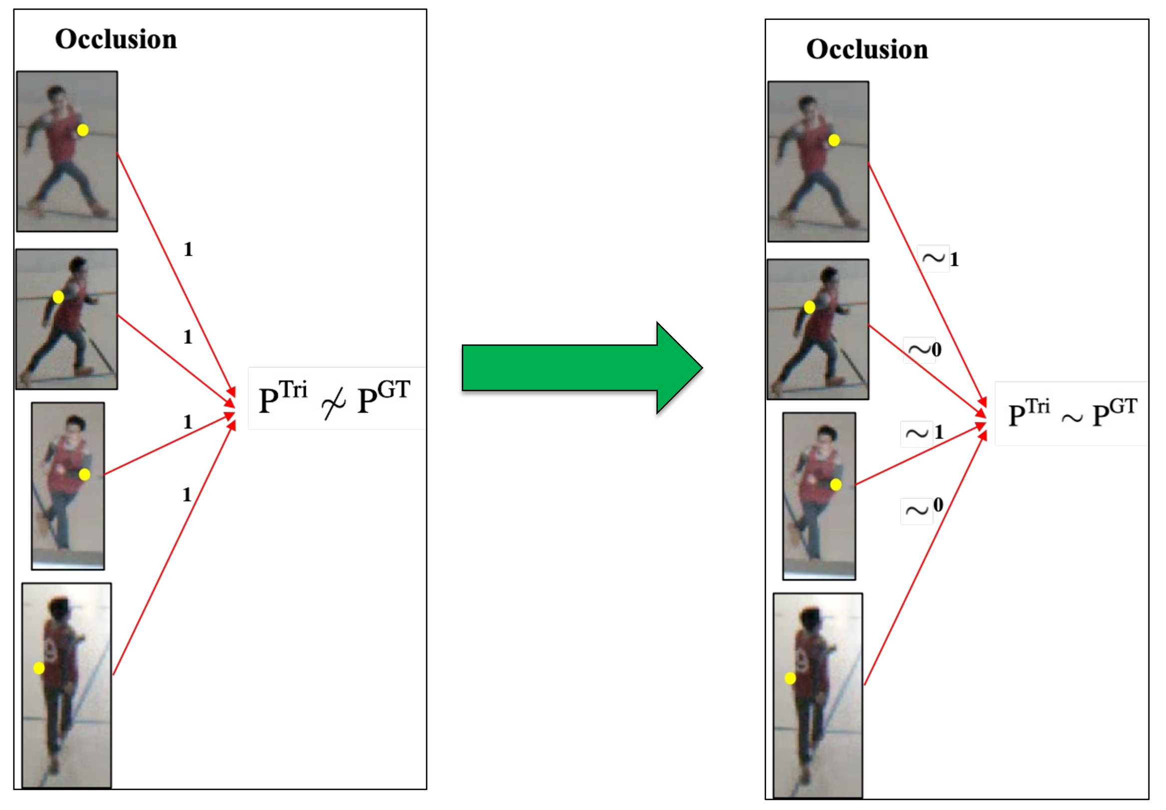
Soumava Kumar Roy* Leonardo Citraro* Sina Honari Pascal Fua
Computer Vision Lab, EPFL

3DV 2022

Motivation:

- Manual 3D pose Annotation is a labor intensive and time consuming task.
- Multi-view 2D pose estimator models are used to obtain pseudo 3D labels.
- These models fail to generate reliable pseudo labels due to:
  - **Occlusion,**
  - **Change in Illumination,**
  - **Low Image Resolution.**

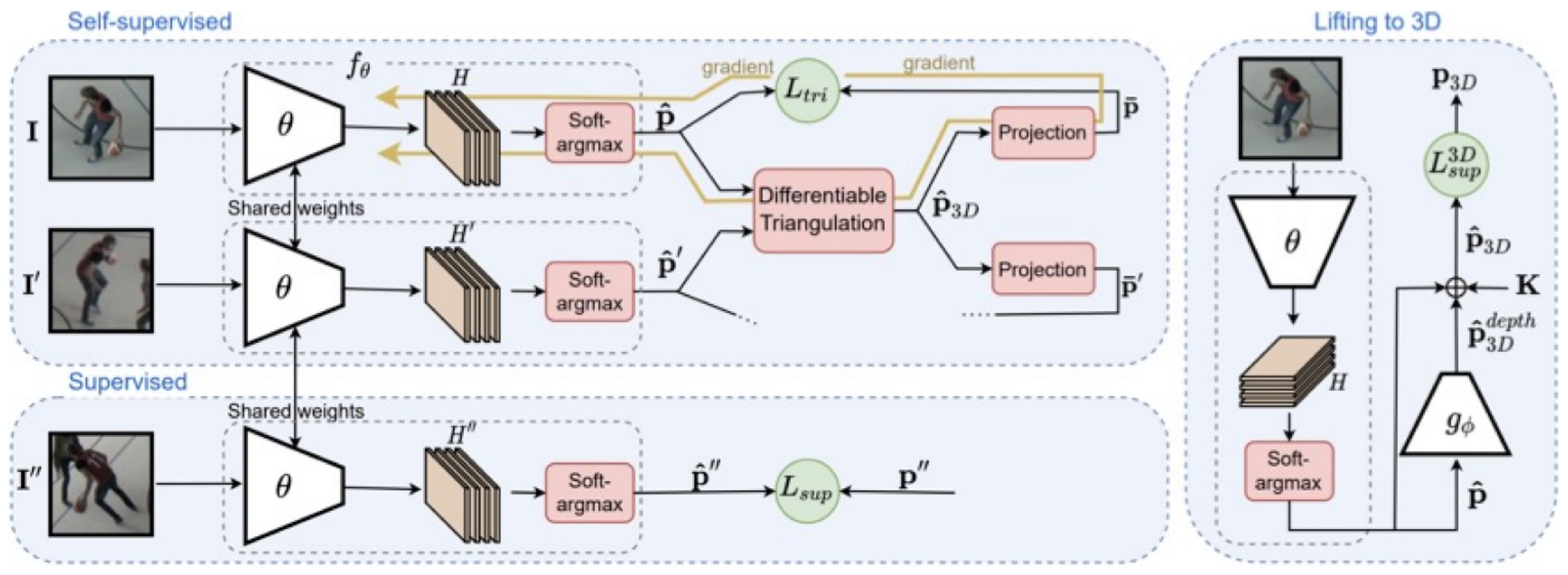
$\Big\}$ NOISY DETECTED KEYPOINTS



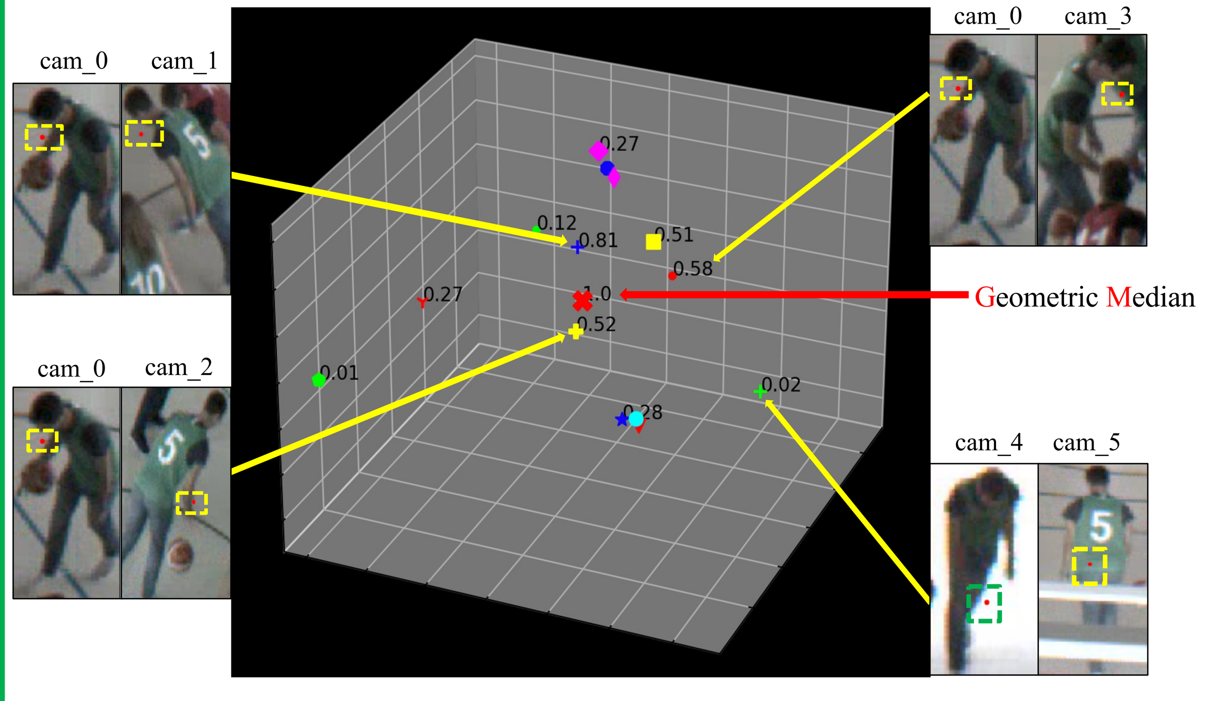
Objective:

1. Minimize the influence of noisy detections for Pose Estimation : **We propose a robust weighting method**.
2. Train Pose Estimator Models with minimum amount of 3D annotations : **Semi-Supervised Learning**.

Training Setup:



Robust Weighting Algorithm:



	cam_1	cam_2	cam_3	cam_4	cam_5
cam_0	0.58	0.02	0.25	0.27	0.51

Median

1. For each joint, compute its 3D pose for all pairwise combination of cameras.
2. Calculate the **Geometric Median (GM)**.
3. Fit a 3D Gaussian with the **GM** as its mean.
4. For a camera, compute the median of weights of pair of cameras containing it to attain its final weight.
5. Repeat step 4 for all $\mathbf{N_c}$ cameras.
6. Repeat steps 1-5 for all $\mathbf{N_j}$ joints to objtain $\mathbf{N_j \times N_c}$ weight matrix.

Training Losses: Semi Supervised Learning

Training f_θ

$$L_{sup}(\boldsymbol{\theta};\mathcal{L}) = \|\hat{\mathbf{p}} - \mathbf{p}\|_2^2$$

$$L_{tri}(\boldsymbol{\theta};\mathcal{U}) = \|\hat{\mathbf{p}} - \bar{\mathbf{p}}\|_2^2$$

- $\hat{\mathbf{p}} \longrightarrow$ **Detected** 2D joints
- $\bar{\mathbf{p}} \longrightarrow$ **Projected** 2D joints
- $\mathbf{p} \longrightarrow$ **Ground Truth** 2D joints

Training g_ϕ

$$L^g(\phi; f_\theta) = \|\hat{\mathbf{p}}_{3D} - \mathbf{p}_{3D}\|_2^2$$

$\mathbf{p}_{3D}$ is

- **triangulated** 3D pose for **unsupervised** samples,
- **ground truth** 3D pose for **supervised** samples.

SportCenter Dataset:

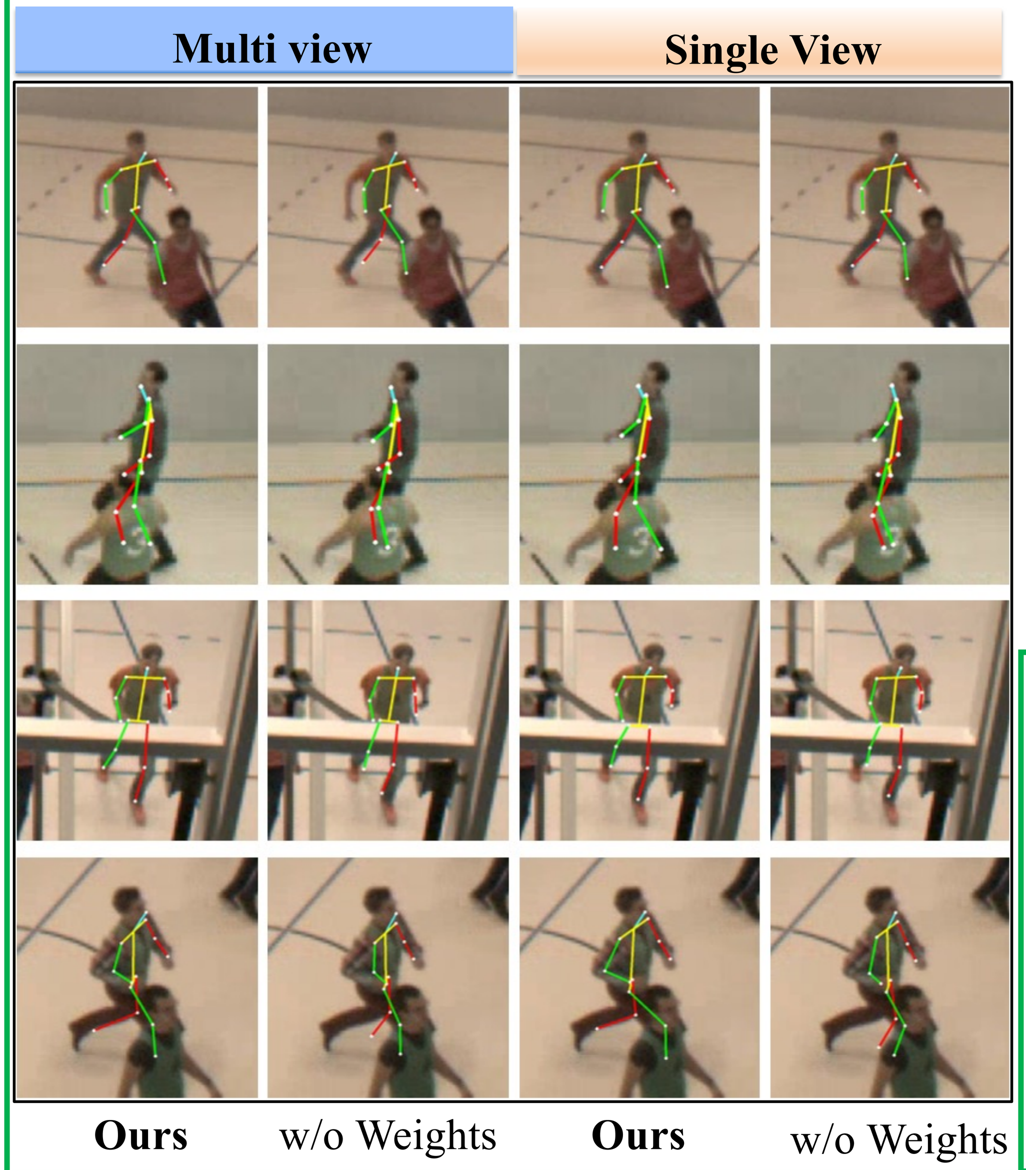
- 13 subjects playing a game of basketball (10 subjects at given instant of time).
- 8 fixed and calibrated cameras, 6 of which are used pose estimation.
- Occlusion by other players or static structures.
- Annotated 3740 2D and 700 3D poses.
- Various Ligthing Conditions.
- 0.3M Images in Total.

Experiments:

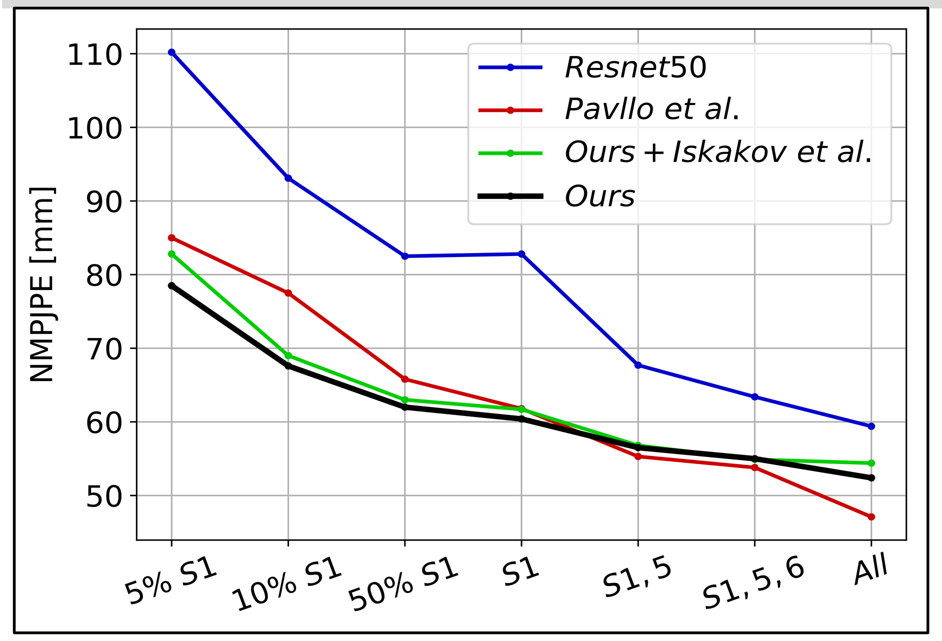
Quantitative Results on **SportCenter** Dataset

Method	Weights	Differentiable	MPJPE (in mms)	
			Multi View	**Single View**
Ours non-differentiable w/o weights	No	No	109.7	142.9
Ours non-differentiable	Yes	No	83.0	111.4
Ours w/o weights	No	Yes	80.5	118.5
Ours+Iskakov [1]	Yes	Yes	88.3	121.1
Ours	Yes	Yes	**66.9**	**104.4**

Qualitative Results on **SportCenter** Dataset



Results on **Human3.6M** Dataset



Future Work:

- Rectifying abnormal 3D poses using discriminators.
- Integrate temporal or spatial constraints using Graphs or Transformers.

Contributions:

- Self Supervised Learning via differentiable triangulation.
- Robust weighting algorithm to generate reliable pseudo 3D labels.
- Multi-Person amateur basketball dataset featuring occlusion and difficult lighting conditions.

References

[1] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable Triangulation of Human Pose. ICCV, 2019