Max Planck Institute for Intelligent Systems – Perceiving Systems Department

Learning to Train with Synthetic Humans

David Hoffmann, Dimitrios Tzionas, Michael J. Black, Siyu Tang

MAX-PLANCK-GESELLSCHAFT

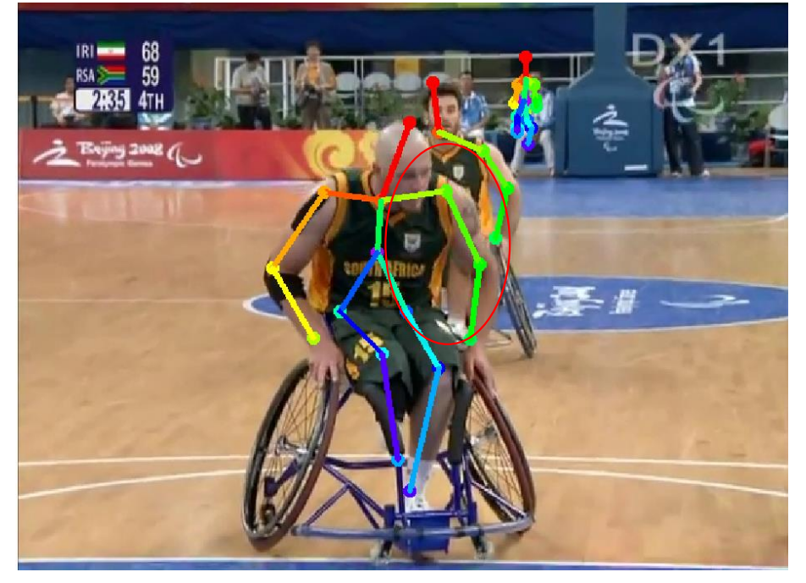
Introduction

Problem

- Not enough hard examples

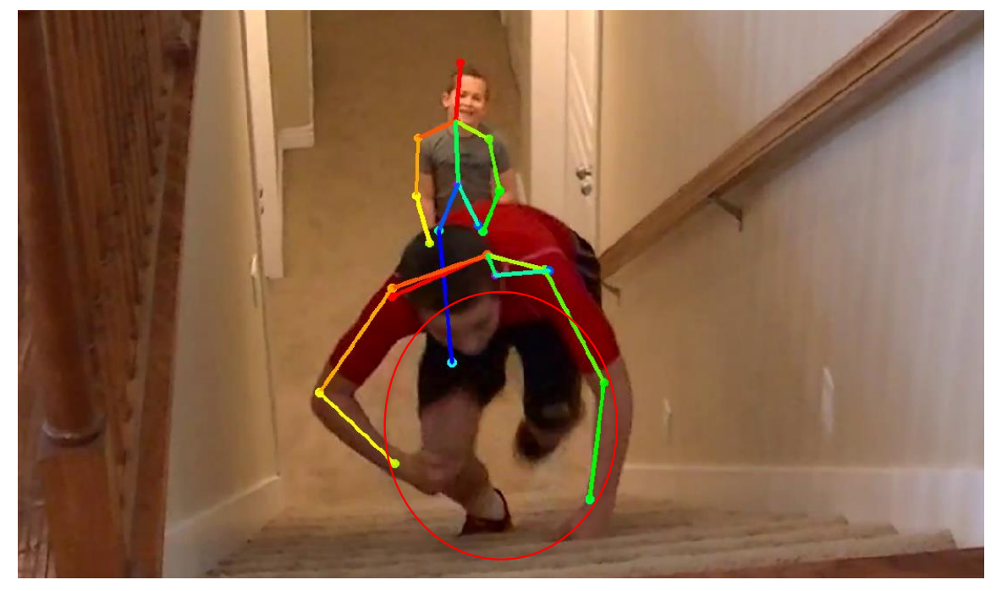
  Frequent failure cases:

Cao et al. [1]



Occlusion

Cao et al. [1]

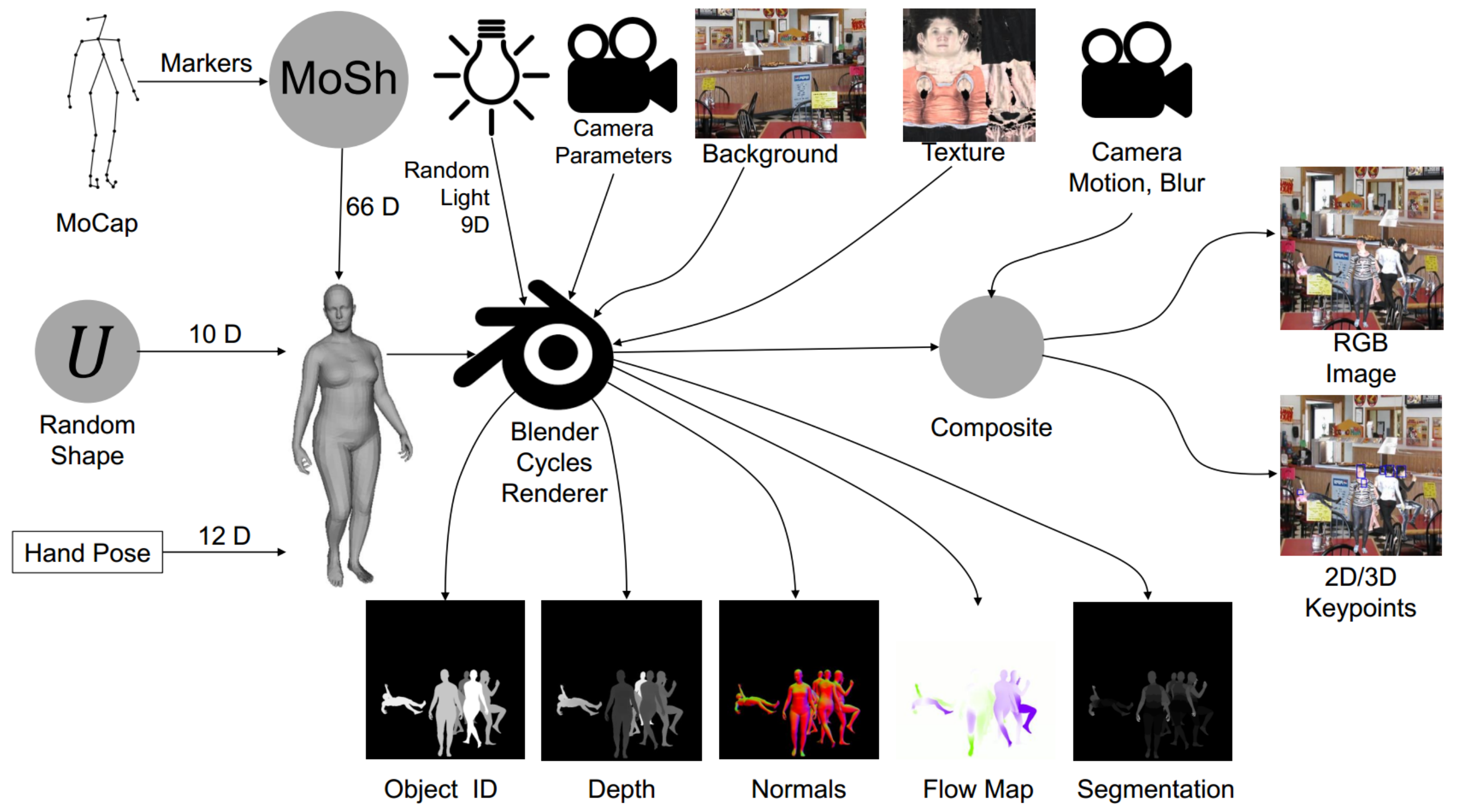


Camera Position

Goal

- Robust multi-person pose estimation

Synthetic Data Generation

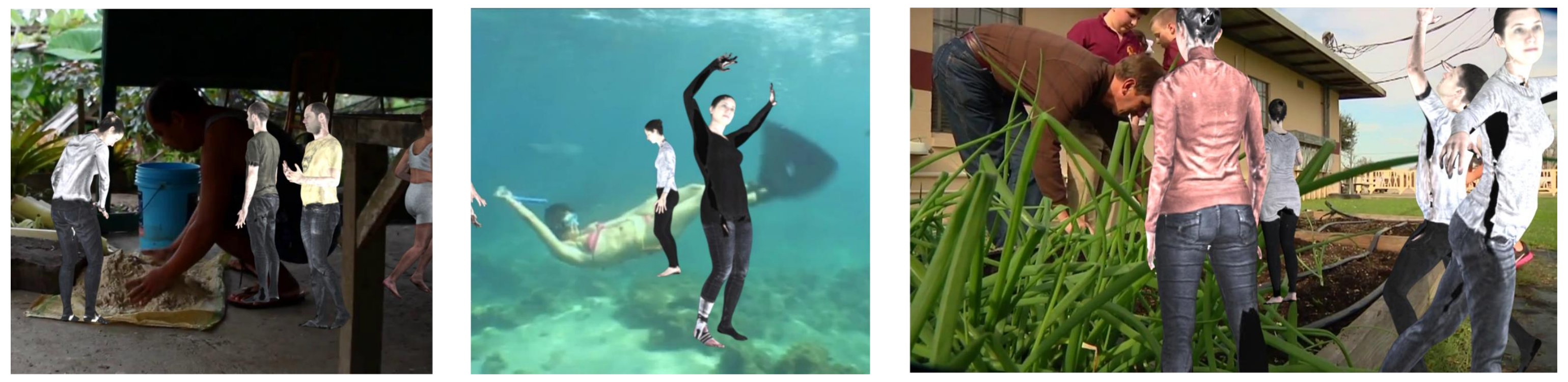


References

1. Cao et al.: Realtime multi-person 2d pose ^estimation using part affinity fields. In: CVPR (2017)
2. Dundar et al.: Domain stylization:A strong, simple baseline for synthetic to real image domain adaptation. arXiv preprint arXiv:1807.09384 (2018)
3. He, K. et al.: ICCV (2017)
4. Peng et al.: Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In: CVPR (2018)

Mixed and Domain Adapted Dataset

Augment real data with synthetic humans



Using a variant of photorealistic style transfer algorithm [2]

- Smaller domain gap
- Occasional artifacts when person detector [3] fails

Results Datasets

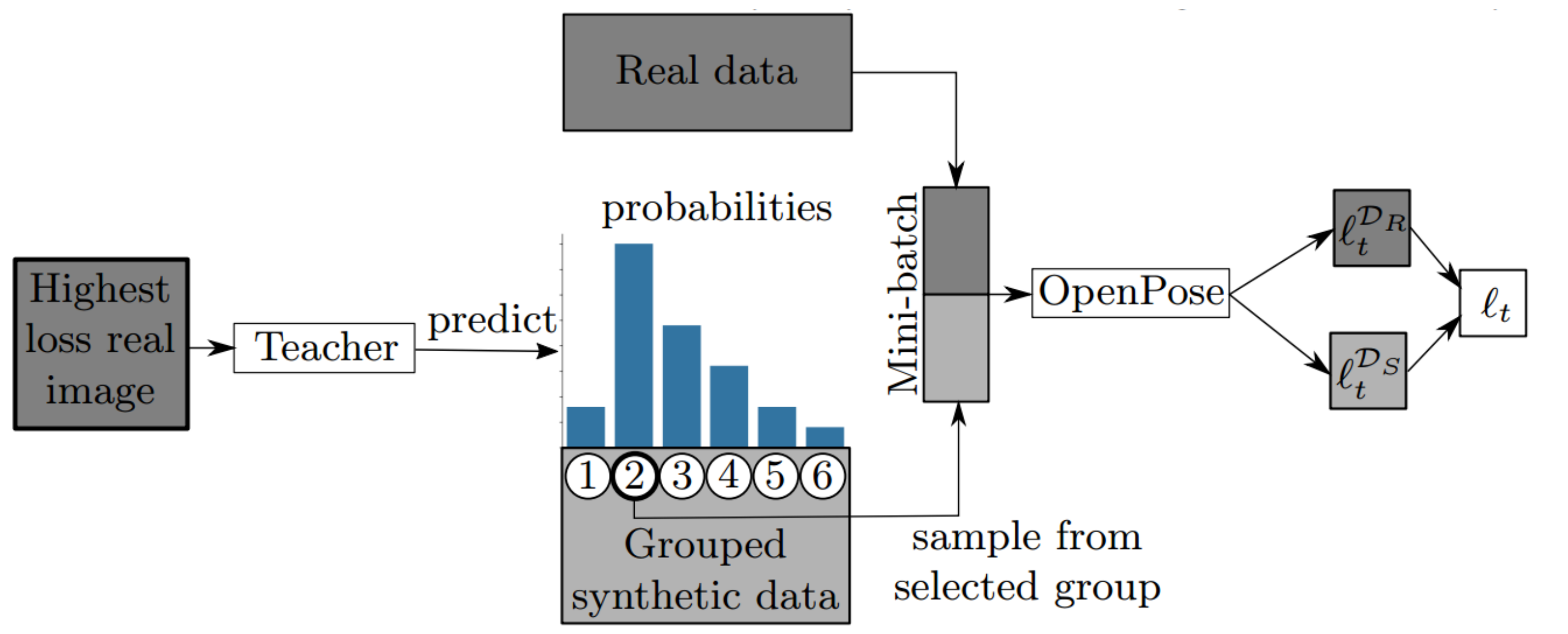
Model	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
$\mathcal{M}_{\mathcal{D}_R}$	**91.3**	89.1	79.2	70.4	**75.9**	71.5	66.7	77.7
$\mathcal{M}_{\mathcal{D}_S}$	37.9	23.5	12.7	7.3	5.6	3.4	3.2	13.4
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_S}$	91.0	89.4	80.4	71.2	75.3	**73.3**	**68.1**	**78.4**
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M}$	**91.3**	**89.5**	**80.7**	**71.7**	75.4	72.5	67.7	**78.4**
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_{Style}}$	91.8	89.8	80.4	70.9	75.5	71.6	67.9	78.3

Masking out Synthetic Humans

Model	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M}$	91.3	89.5	80.7	71.7	75.4	**72.5**	67.7	78.4
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M+\text{masks}}$	**92.3**	**90.9**	80.5	**72.2**	76.0	71.7	68.3	78.9
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_{Style}}$	91.8	89.8	80.4	70.9	75.5	71.6	67.9	78.3
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_{Style}+\text{masks}}$	91.6	90.6	**80.8**	71.8	**77.7**	72.2	**68.8**	**79.1**

Teacher Network

Not all synthetic data samples provide new information



Inspired by [4] we

Reward if

$$\ell_t^{\mathcal{D}_S} \geq \frac{1}{H}\sum_{h=0}^{H} \ell_{t-1-h}^{\mathcal{D}_S},$$

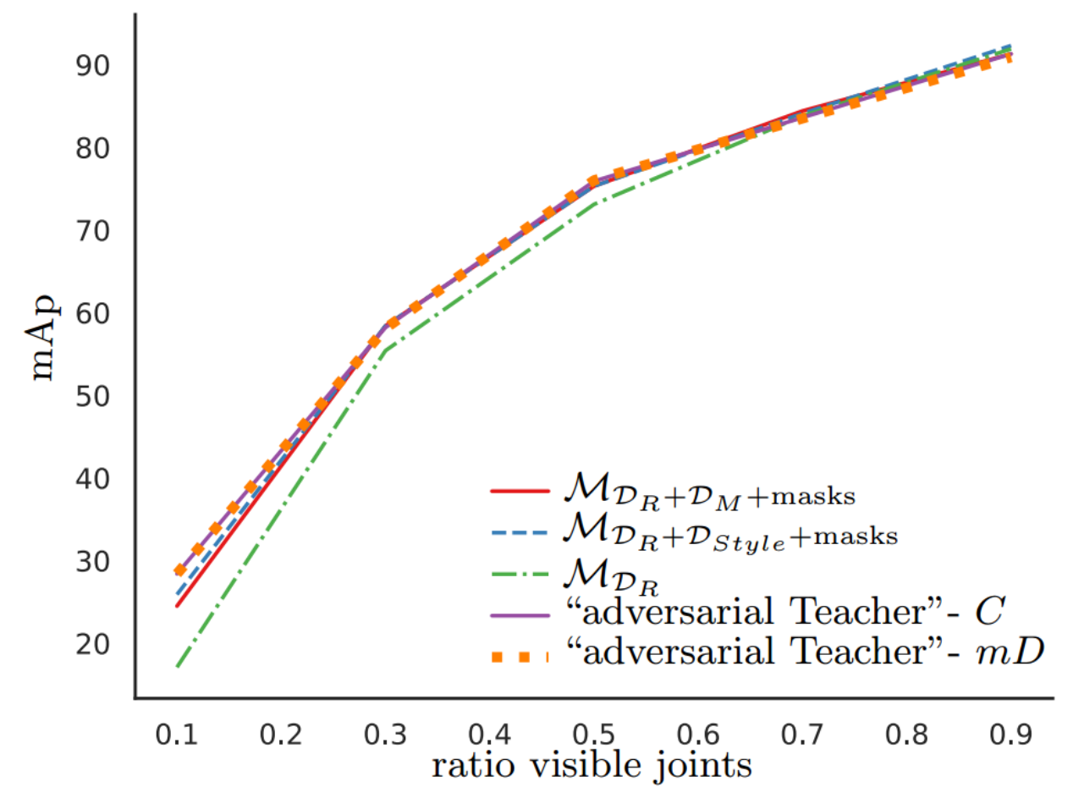
and update by

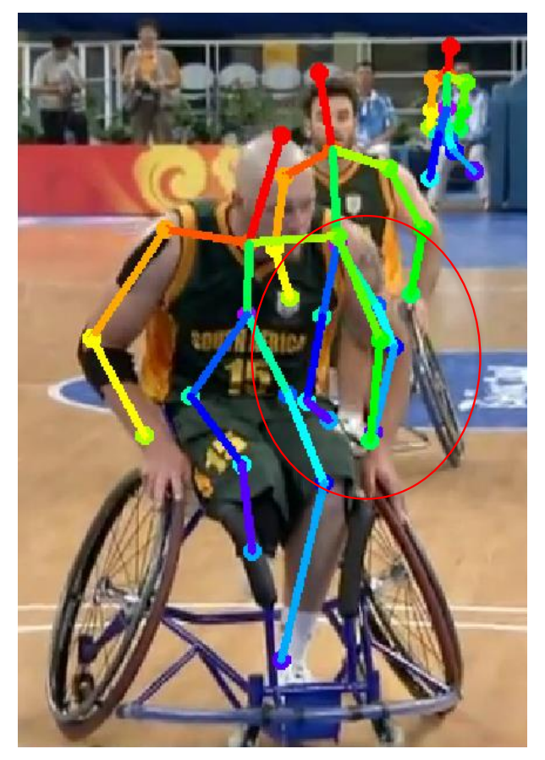
$$P_i = \tilde{P}_i + \delta\alpha\tilde{P}_i$$

$$P_j = \tilde{P}_j - \delta\frac{\alpha\tilde{P}_i}{|g|-1}.$$

Results Teacher

Model	Grouping	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
$\mathcal{M}_{\mathcal{D}_R}$		91.3	89.1	79.2	70.4	75.9	71.5	66.7	77.7
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_S}$		91.0	89.4	80.4	71.2	75.3	73.3	**68.1**	78.4
“adversarial Teacher”	C	**91.7**	90.0	**80.9**	71.2	**77.1**	**73.6**	67.7	**78.9**
“adversarial Teacher”	mD	91.5	**90.4**	80.5	**72.2**	75.8	73.1	67.6	78.7

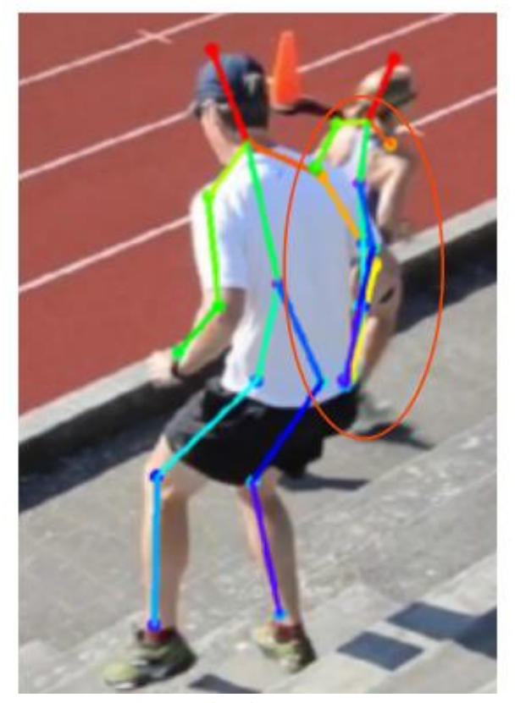




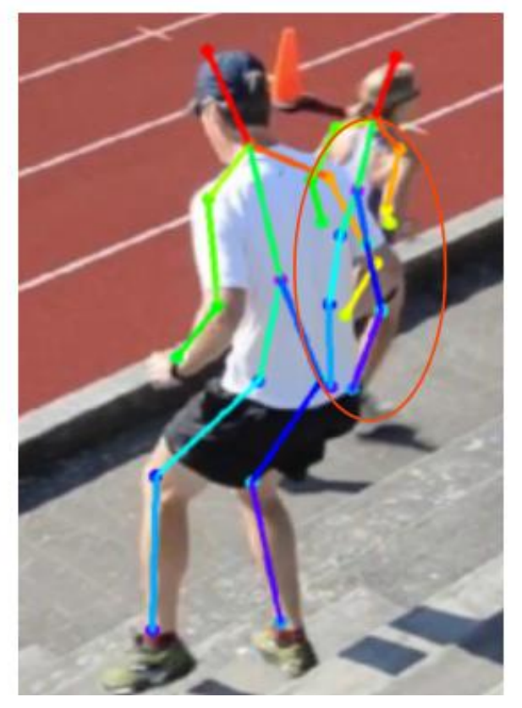
adversarial teacher C

adversarial teacher C

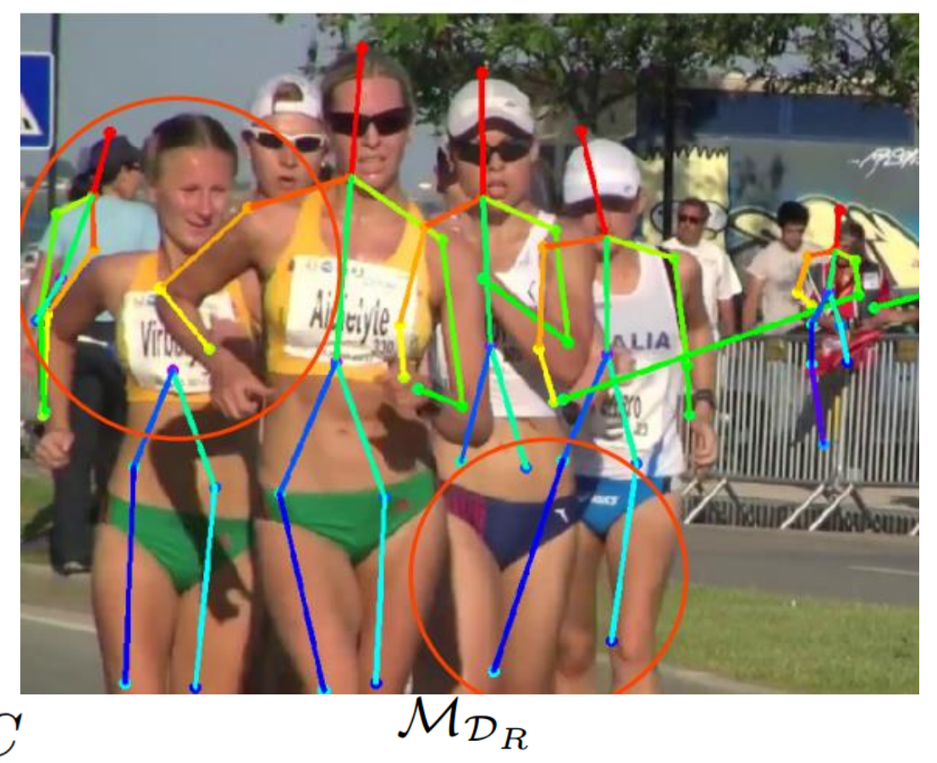
Qualitative Results and Conclusion



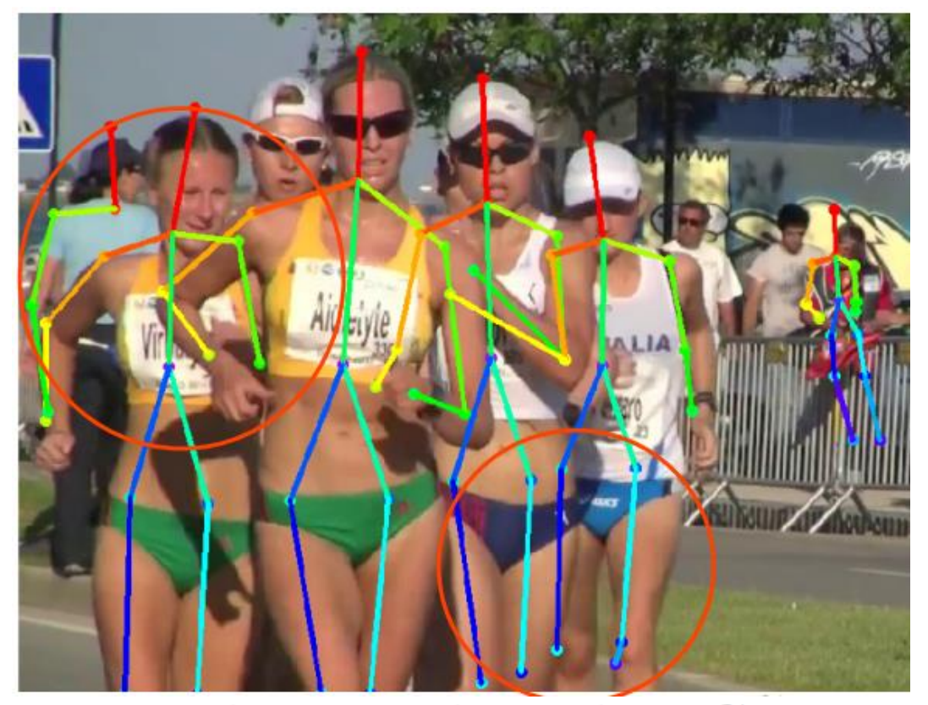
$\mathcal{M}_{\mathcal{D}_R}$



adversarial teacher C



$\mathcal{M}_{\mathcal{D}_R}$



adversarial teacher C

- Training with synthetic data improves multi-person pose estimation
- Augmenting real data with synthetic humans helps
- For the mixed dataset most of the improvement is due to more occlusion (masking out synthetic humans)
- Stylization helps only when synthetic humans are masked out
- Informed sampling enables more effective use of synthetic data