ISTITUTO ITALIANO
DI TECNOLOGIA

ISTITUTO ITALIANO
DI TECNOLOGIA

Lifting Monocular Events to 3D Human Poses

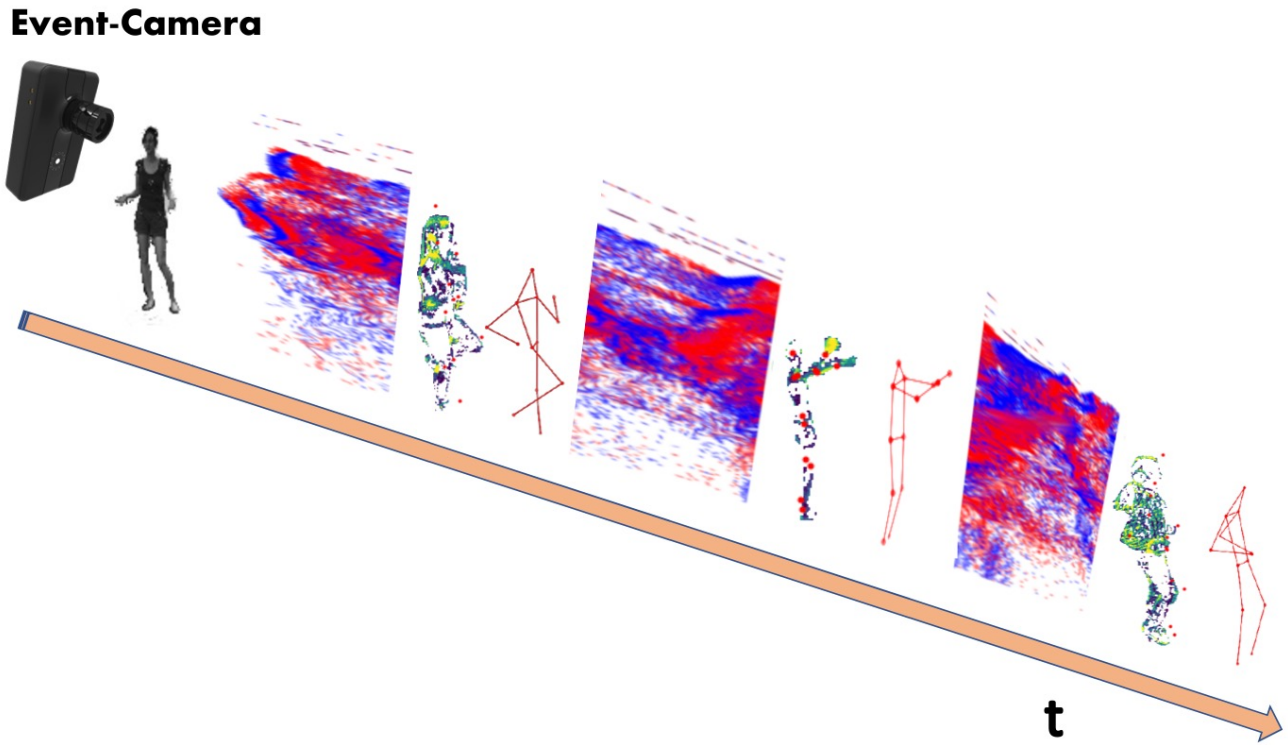
Gianluca Scarpellini

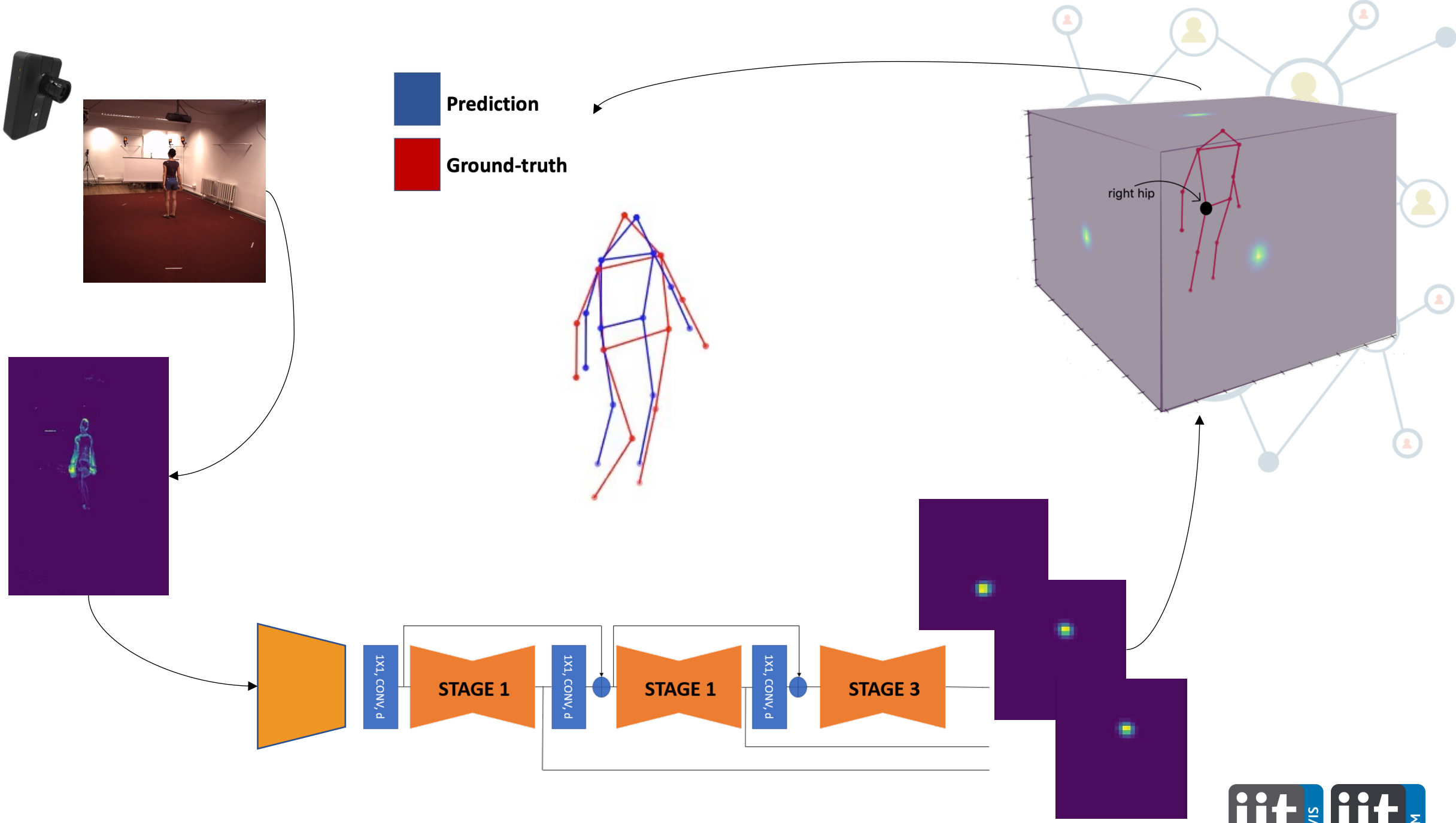
Pietro Morerio

Alessio Del Bue

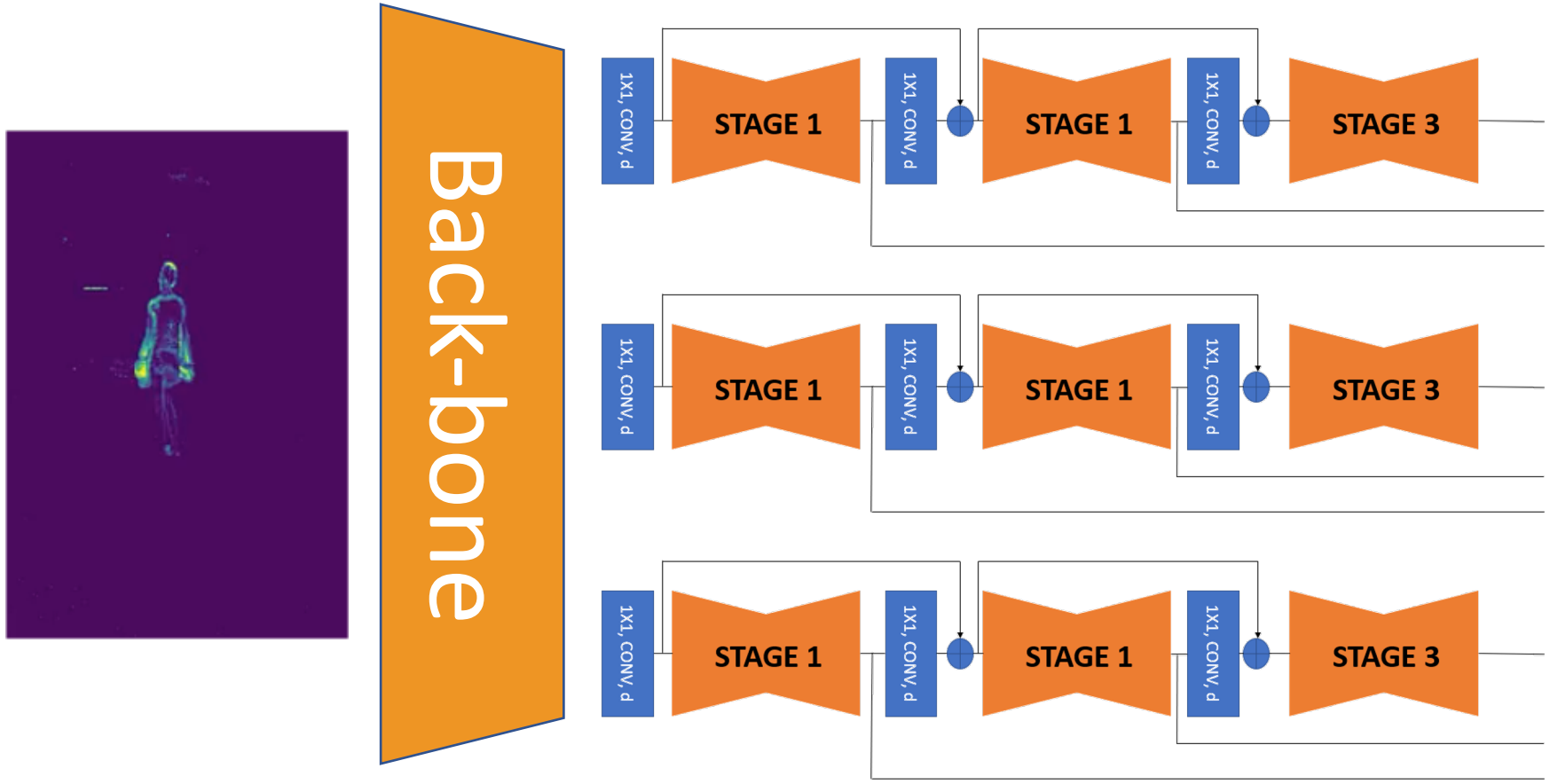
Our contribution

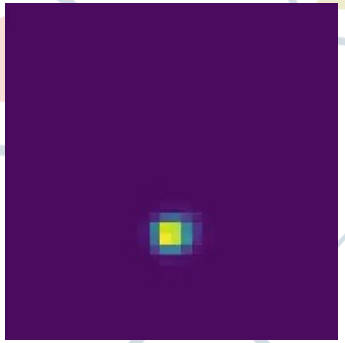
- First **events-only** monocular HPE approach
- Novel synthetic dataset for event-based human pose estimation
- Experiments for the best representation and backbone



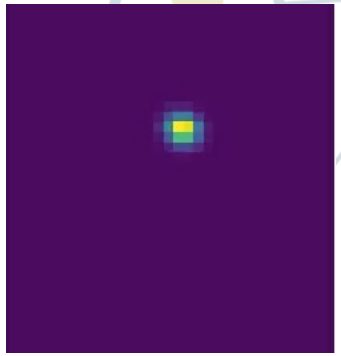

Prediction
Ground-truth
right hip
1X1, CONV, d
STAGE 1
1X1, CONV, d
STAGE 1
1X1, CONV, d
STAGE 3

Methodology: *marginal heatmaps*[1]

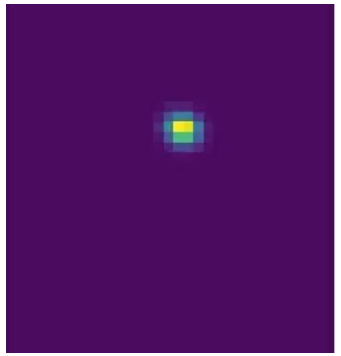




$\hat{H}^i_{xy}$



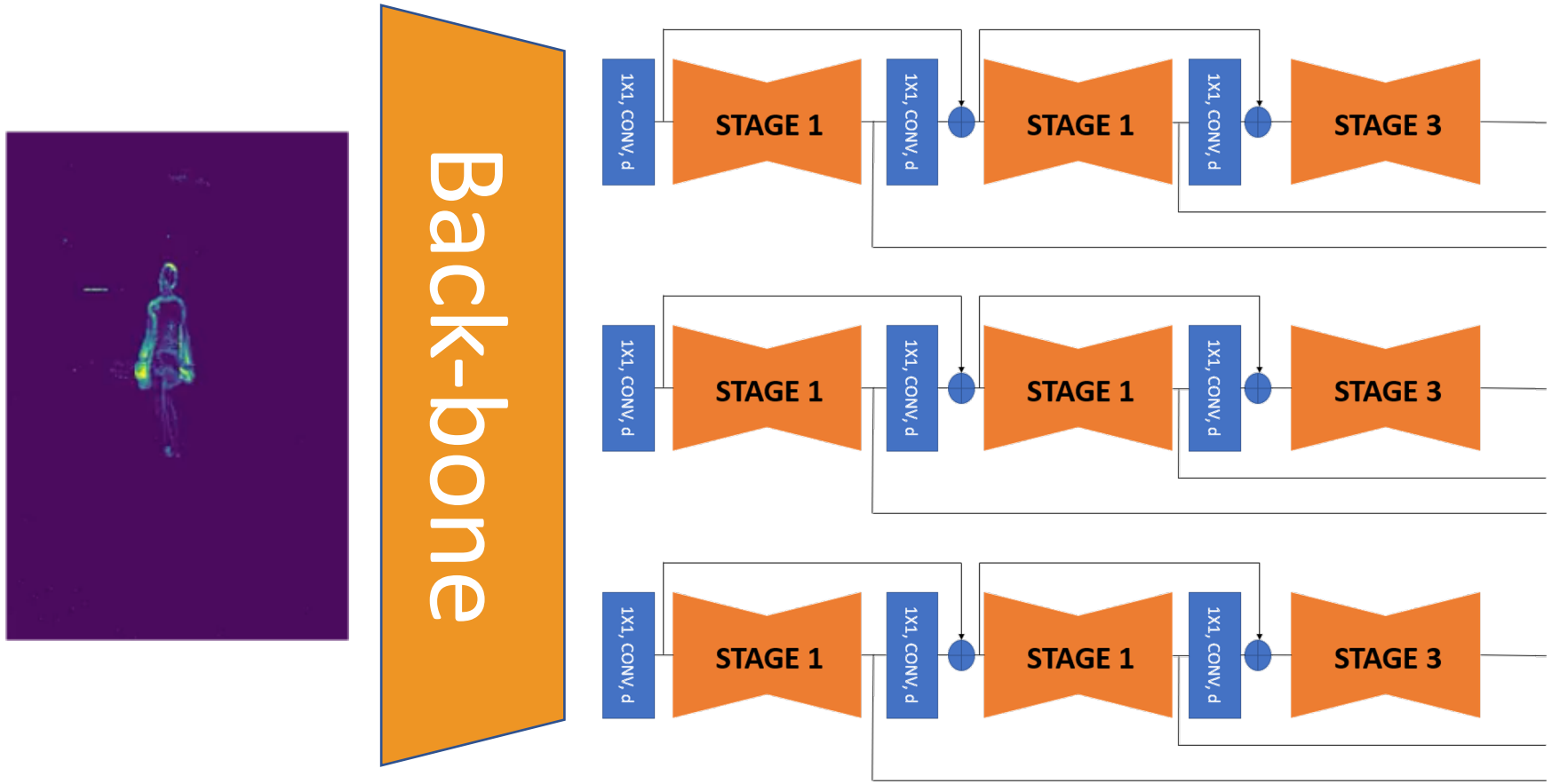
$\hat{H}^i_{zy}$

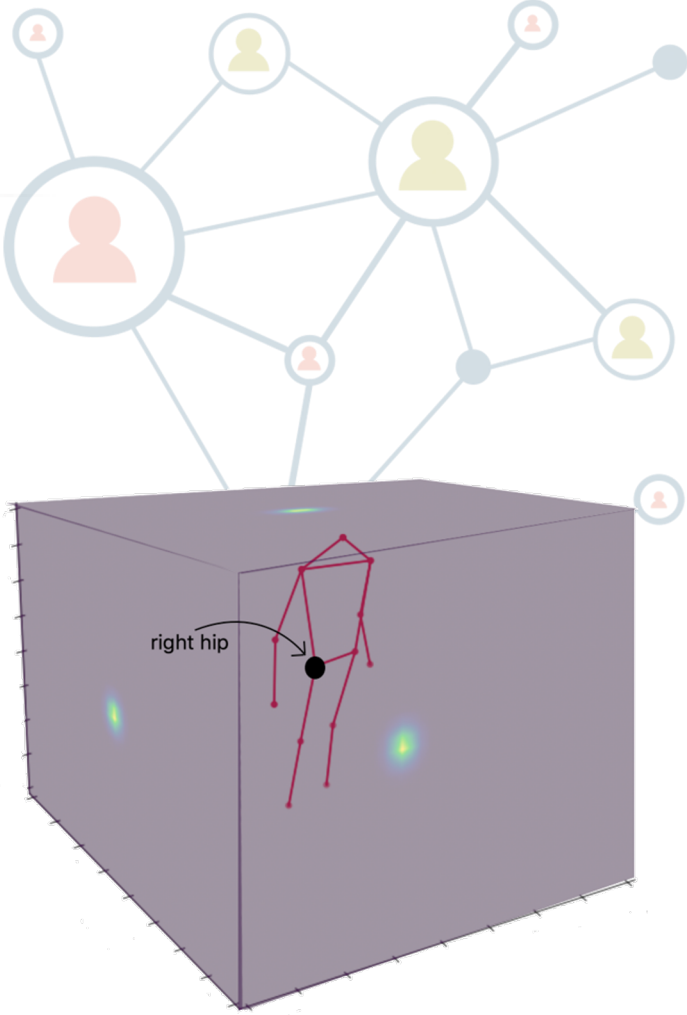


$\hat{H}^i_{xz}$

[1] Nibali, Aiden, et al. **3d human pose estimation with 2d marginal heatmaps.** *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019.

Methodology: *marginal heatmaps*



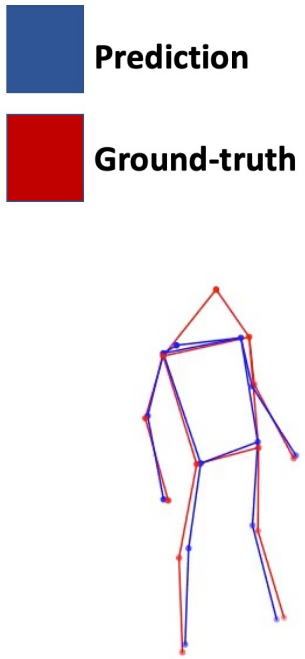


Experiments: DHP19

Method	input	MPJPE(mm)
Calabrese *et al.* [5]	stereo	79.63
Constant-count – stage 3	monocular	92.09
Voxel-grid – stage 3	monocular	95.51
Constant-count – stage 1	monocular	96.69
Voxel-grid – stage 1	monocular	105.24

MPJPE (lower is better)

Comparsion with stereo approach on DHP19[1]



[1] Calabrese, Enrico, et al. **Dhp19: Dynamic vision sensor 3d human pose dataset.** *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

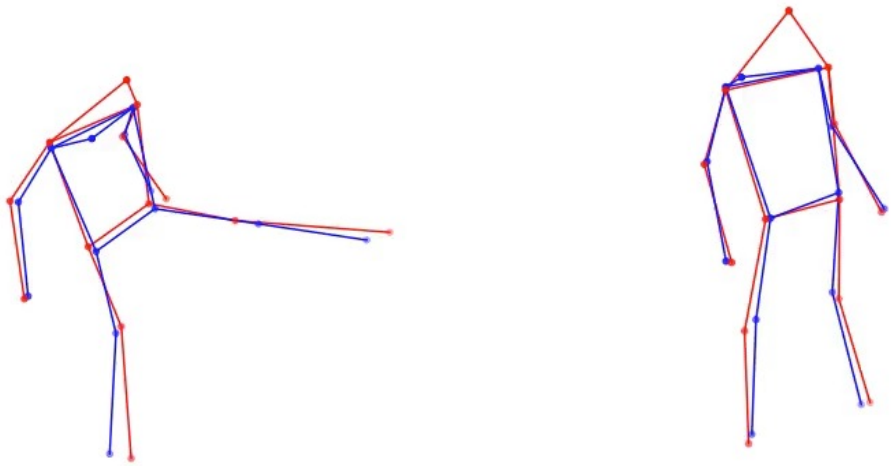
Experiments: DHP19 - *ablations*

Repr.	Model	Initialization	MPJPE (mm)
constant-count	ResNet-34	Random initialized	92.22
		Action recognition	95.19
		Reconstruction	98.89
		ImageNet	**92.09**
	ResNet-50	Random initialized	92.22
		Action recognition	92.26
		ImageNet	92.51
voxel-grid	ResNet-34	Random initialized	93.06
		Action recognition	95.26
		Reconstruction	105.44
		ImageNet	95.51
	ResNet-50	Random initialized	93.88
		Action recognition	93.54
		ImageNet	93.98

MPJPE (lower is better)

Prediction

Ground-truth



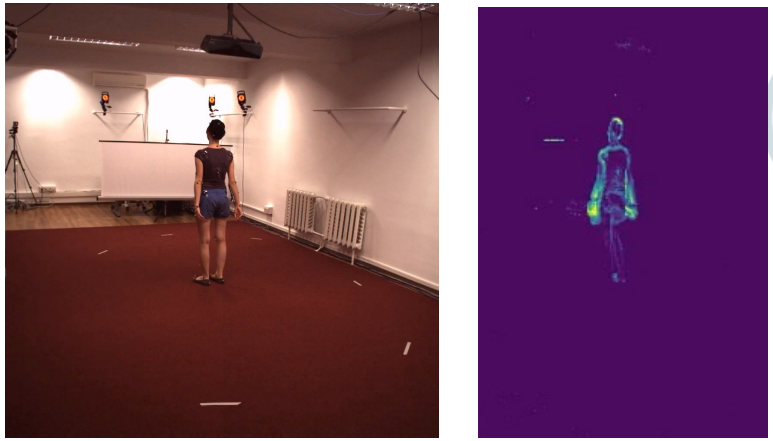
[1] Calabrese, Enrico, et al. **Dhp19: Dynamic vision sensor 3d human pose dataset.** *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

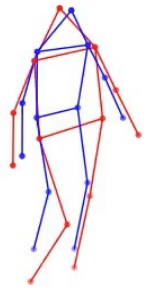
Experiments: Event-Human3.6m

Method	input	MPJPE(mm)
Metha *et al.* [38] (ResNet-50)	RGB	80.50
Kanazawa *et al.* [22]	RGB	88.00
Nibali *et al.* [43]	RGB	57.00
Pavlakos *et al.* [44]	RGB	71.90
Luvizon *et al.* [33]	RGB	53.20
Cheng *et al.* [9]	RGB	**40.10**
Spatio-temporal voxel-grid (**Ours**)	Events	119.18
Constant-count (**Ours**)	Events	116.40

MPJPE (lower is better)

Comparsion with RGB approaches on H3.6m[1]

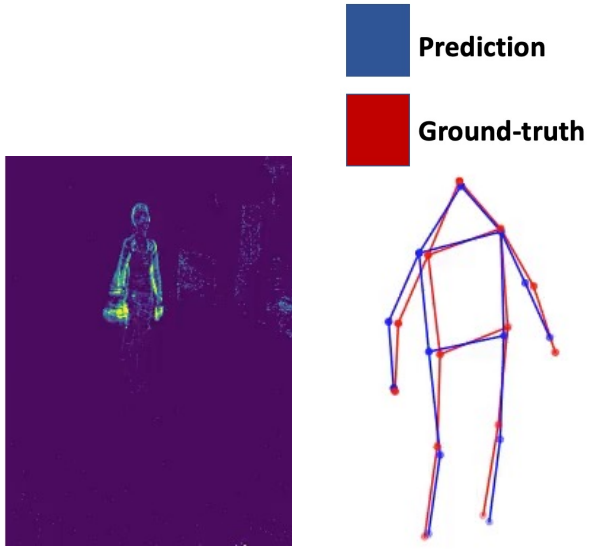




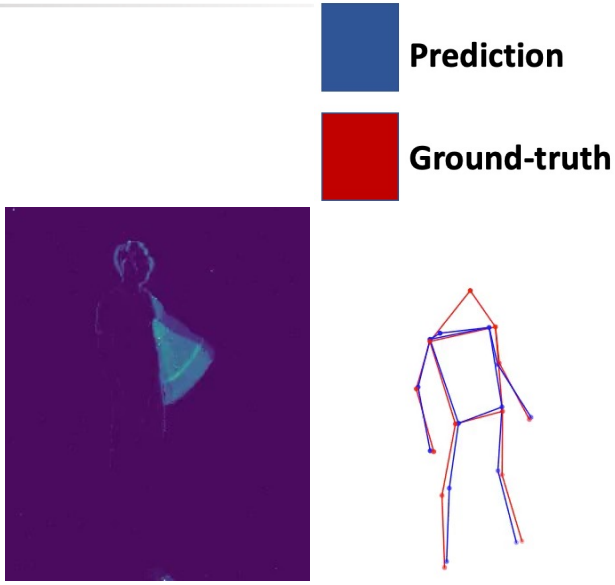
From RGB frames[1] to events

[1] Ionescu, Catalin, et al., **Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments.** *IEEE transactions on pattern analysis and machine intelligence*

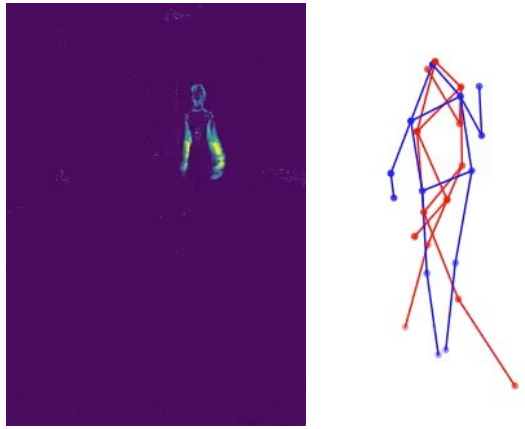
Visual results



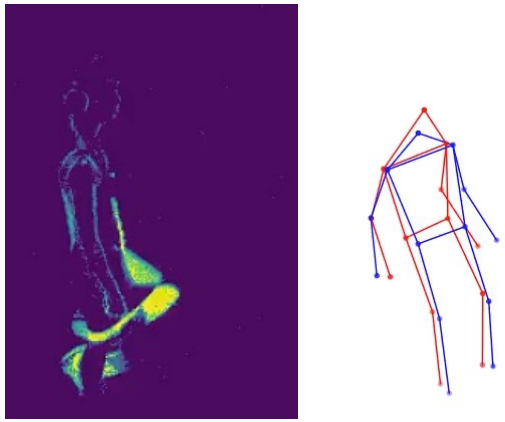
Discussion – Event-Human3.6m



Left-arm – DHP19



Smoking – Event-Human3.6m



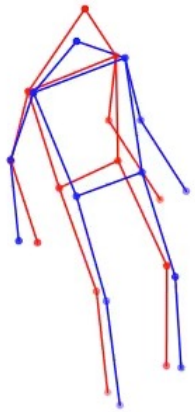
Side-kick – DHP19

Conclusion

- Main causes of **failure**
  - **Static movements**
  - **Occluded parts of the body**

- Constat-count representation works better than spatio-temporal voxel-grid

- ImageNet pretraining improves the results



ISTITUTO ITALIANO
DI TECNOLOGIA

ISTITUTO ITALIANO
DI TECNOLOGIA

Thank you!

@pavis_iit, @gianscarpellini,
@pmorerio, @ilpazuzu

GitHub

https://tinyurl.com/b3kwbrmy

arXiv.org

https://arxiv.org/abs/2104.10609