Multi-Person Extreme Motion Prediction: Supplementary Material

Wen Guo[1,2]*, Xiaoyu Bie[1]*, Xavier Alameda-Pineda[1], Francesc Moreno-Noguer[2]
[1]Inria, Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[2]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[1]{wen.guo,xiaoyu.bie,xavier.alameda-pineda}@inria.fr, [2]fmoreno@iri.upc.edu

A. Personal data/Human subjects

Our data collection strategy went through an Ethics Review Board, and the recordings where authorised, together with the associated Consent Form. Our data does not contain any personally identifiable information beyond the images themselves. The data will be shared respecting all national and international regulations, as authorised by COERLE, the Ethics Review Board at INRIA.

B. More information about the dataset

B.1. Data Post-processing

As introduced in the main paper, it is a common phenomena in lab-based interaction Mocap datasets that many points are missing due to occlusions or tracking loss. This is even worse when dealing with extreme poses. To overcome this we have designed and implemented a 3D hand labelling toolbox.

For each missed value, we choose two orthogonal views among the several viewpoints, and label the missed keypoints by hand on these two frames to get two image coordinates. We then use the camera calibration to back project these two image coordinates into the 3D world coordinate, obtaining two straight lines. Ideally, the intersection of these two lines is the world coordinate of this missing point. Since these two lines do not always intersect, we find the nearest point, in the least-squares sense, to these two lines to approximate the intersection.

In this procedure we did not use the distortion parameters, since we observed that the distortion error is negligible on the views we choose for labeling. The intersection is projected into 3D and various 2D images to confirm the quality of the approximation by visual inspection. Figure A shows an example of labeling the missing joints.

B.2. Action names and joint order

Table A shows the name of the 16 actions performed by the couples of actors in ExPI. In the video of the supplementary material, we include example videos for each of the 16 actions. In the ExPI dataset, the pose of each person is an-

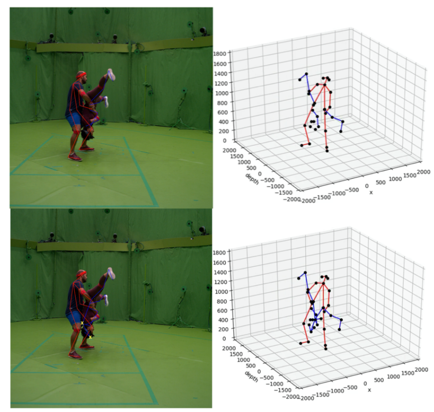


Figure A. Data-cleaning. **Top:**Data before cleaning. The two joints 'F-back' and 'F-fhead' are missed. **Bottom:** Data after cleaning. The yellow marks indicates the two relabeled joints.

Table A. Composition of the ExPI Dataset. The seven first actions are performed by both couples. Six more actions are performed by Couple 1, while three others by Couple 2.

Action	Name	Couple 1	Couple 2
A_1	A-frame	✓	✓
A_2	Around the back	✓	✓
A_3	Coochie	✓	✓
A_4	Frog classic	✓	✓
A_5	Noser	✓	✓
A_6	Toss out	✓	✓
A_7	Cartwheel	✓	✓
A_8	Back flip	✓	
A_9	Big ben	✓	
A_{10}	Chandelle	✓	
A_{11}	Check the change	✓	
A_{12}	Frog-turn	✓	
A_{13}	Twisted toss	✓	
A_{14}	Crunch-toast		✓
A_{15}	Frog-kick		✓
A_{16}	Ninja-kick		✓

*Equal contribution.

Table B. Comparison of ExPI with other publicly available datasets commonly used for 3D human tasks.

Dataset	AMASS [3]	H3.6m [2]	3DPW [8]	MuPoTS [7]	ExPI
3D joints	✓	✓	✓	✓	✓
Video	✓	✓	✓	✓	✓
Shape	✓	✓	✓		✓
Multi-person			✓	✓	✓
Extreme poses	✓				✓
Multi-view					✓

notated with 18 keypoints, so we have 36 keypoints for both actors. The order of the keypoints is as follows, where "F" and "L" denote the Follower and the Leader respectively, and "f", "l" and "r" denote "forward", "left" and "right":

(0) 'L-fhead'	(1) 'L-lhead'	(2) 'L-rhead'
(3) 'L-back'	(4) 'L-lshoulder'	(5) 'L-rshoulder'
(6) 'L-lelbow'	(7) 'L-relbow'	(8) 'L-lwrist'
(9) 'L-rwrist'	(10) 'L-lhip'	(11) 'L-rhip'
(12) 'L-lknee'	(13) 'L-rknee'	(14) 'L-lheel'
(15) 'L-rheel'	(16) 'L-ltoes'	(17) 'L-rtoes'
(18) 'F-fhead'	(19) 'F-lhead'	(20) 'F-rhead'
(21) 'F-back'	(22) 'F-lshoulder'	(23) 'F-rshoulder'
(24) 'F-lelbow'	(25) 'F-relbow'	(26) 'F-lwrist'
(27) 'F-rwrist'	(28) 'F-lhip'	(29) 'F-rhip'
(30) 'F-lknee'	(31) 'F-rknee'	(32) 'F-lheel'
(33) 'F-rheel'	(34) 'F-ltoes'	(35) 'F-rtoes'

B.3. Comparison with other datasets

Table B compares our dataset with several other public available 3D human datasets that are widely used in recent work [1, 4–6]. From this table, we can see that our dataset is eminently suitable for the task of multi-person extreme motion prediction, and it is also able to be used in human pose estimation in rare condition and challenging human shape estimation.

C. More Qualitative results

More qualitative results could be found at the end of this file. We compare our model with models that independently predict the motion of each person, i.e. Res-RNN [6], LTD [5], Hisrep [4] and MSR [1]. Our results are much closer to the ground truth, and it works well even on some extreme actions where other methods totally fail.

References

[1] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11467–11476, October 2021. B.3, C

[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. B

[3] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019. B

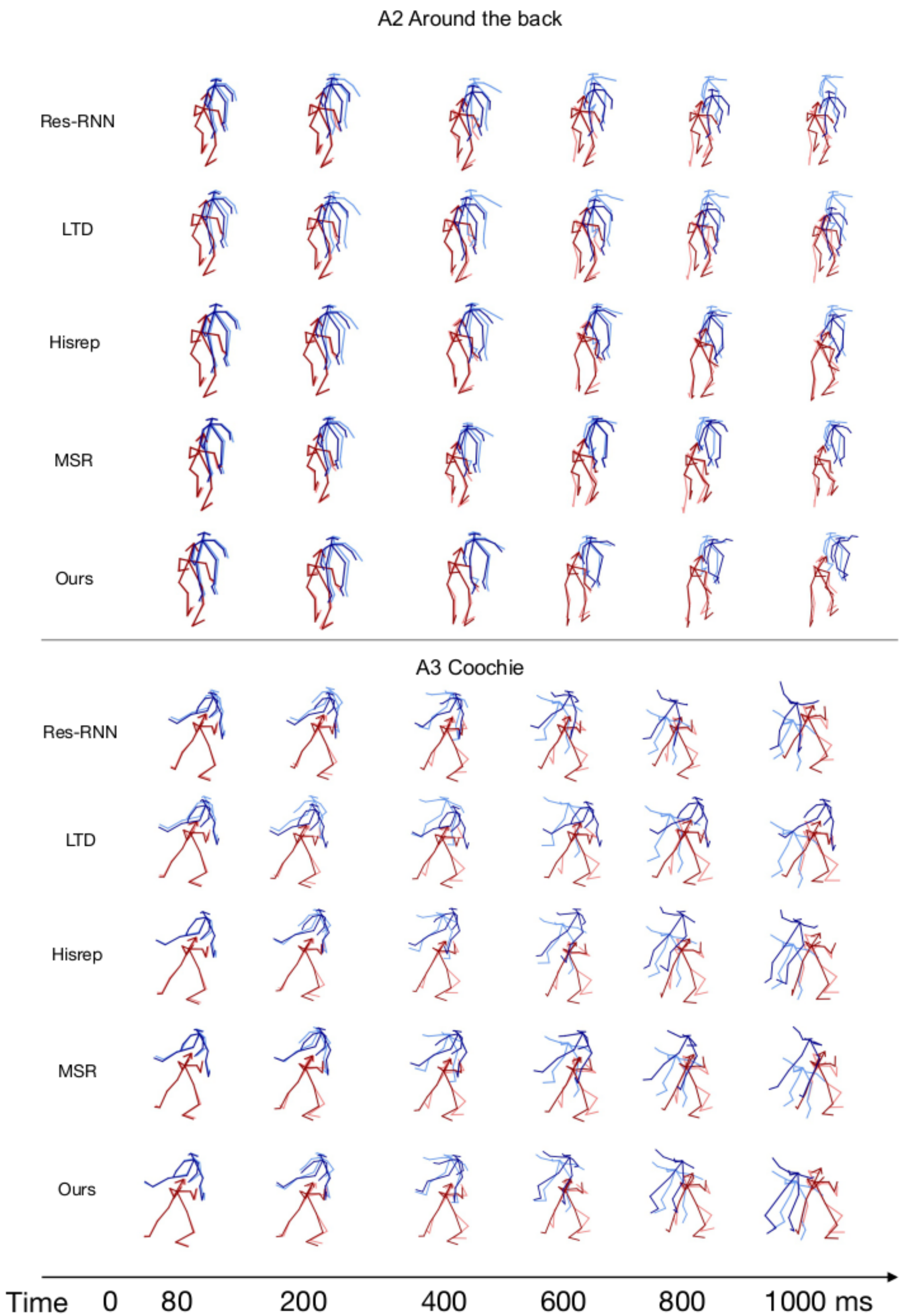
[4] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020. B.3, C

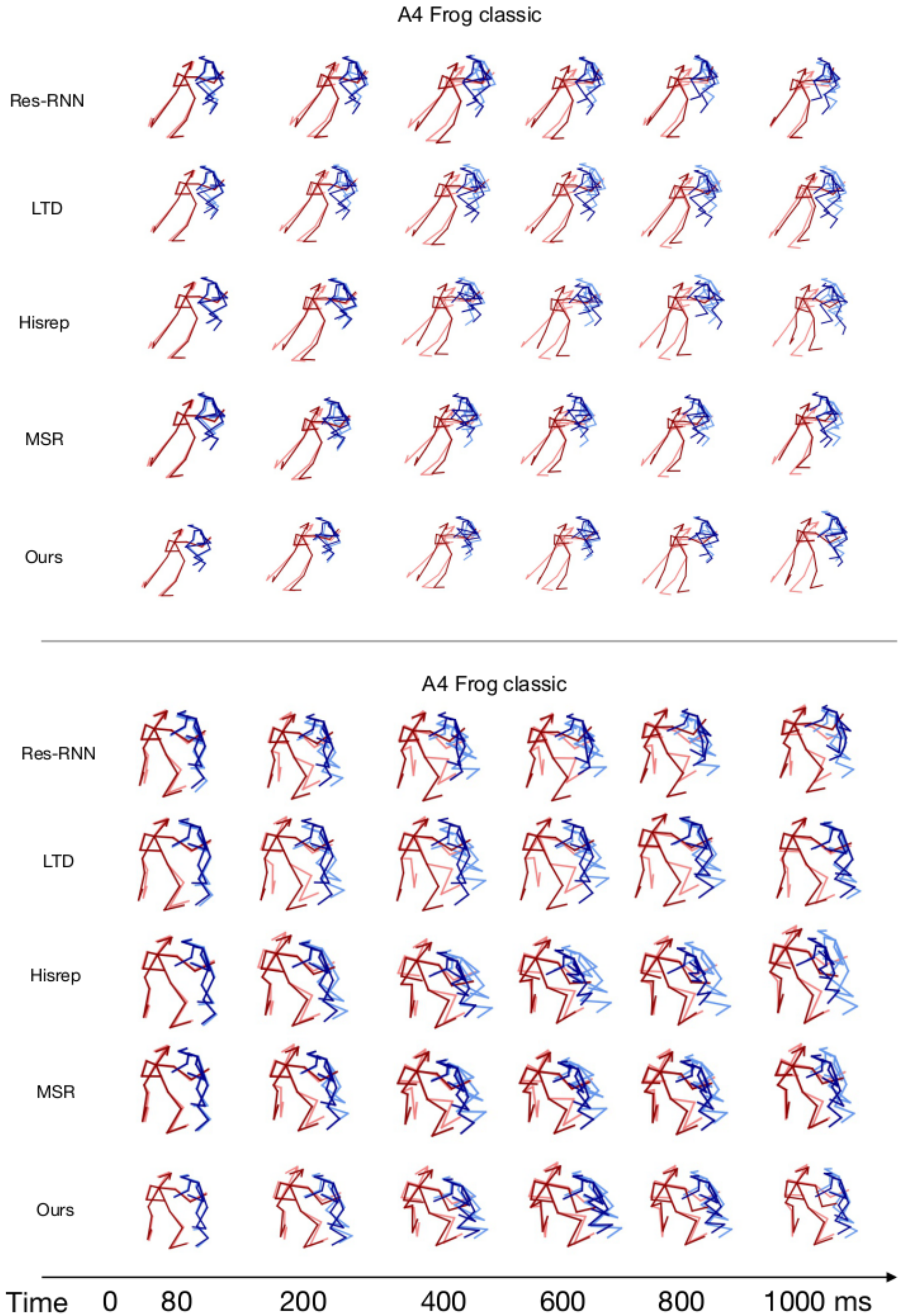
[5] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019. B.3, C

[6] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017. B.3, C

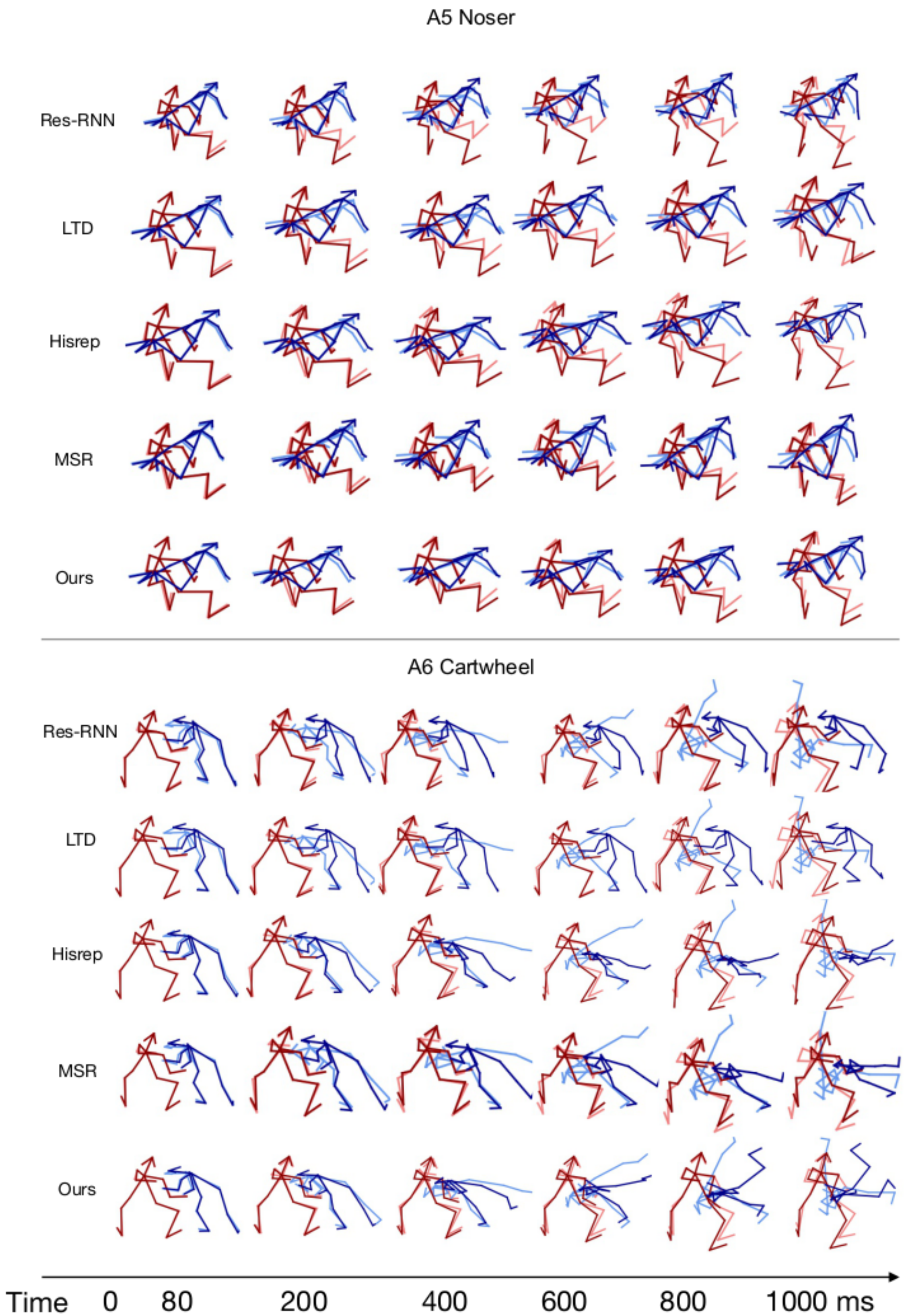
[7] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE, sep 2018. B

[8] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. B


A2 Around the back
Res-RNN
LTD
Hisrep
MSR
Ours
A3 Coochie
Res-RNN
LTD
Hisrep
MSR
Ours
Time 0 80 200 400 600 800 1000 ms



A4 Frog classic
Res-RNN
LTD
Hisrep
MSR
Ours
A4 Frog classic
Res-RNN
LTD
Hisrep
MSR
Ours
Time
0
80
200
400
600
800
1000 ms


A5 Noser
Res-RNN
LTD
Hisrep
MSR
Ours
A6 Cartwheel
Res-RNN
LTD
Hisrep
MSR
Ours
Time
0
80
200
400
600
800
1000 ms