Precision Modeling of 3D Human Motion
(Behaviour and Performance Analysis)

Ajmal S Mian
Professor

Computer Science & Software Engineering

Personal website: http://ajmalsaeed.net/

University profile: https://research-repository.uwa.edu.au/en/persons/ajmal-mian

Overview

- Human motion analysis vs video retrieval

- Pose invariant human action recognition from trajectories
- Application to performance optimization in sports

Trajectories

- Video based human action recognition
- Full 3D mesh human pose recovery from monocular video
- Deep Affinity Network for multiple object tracking

Videos

Human Motion Analysis is Unique

Human Action Recognition Without Human

Yun He, Soma Shirakabe, Yutaka Satoh, and Hirokatsu Kataoka(✉)

National Institute of Advanced Industrial Science and Technology (AIST),
Tsukuba, Ibaraki, Japan
{yun.he,shirakabe-s,yu.satou,hirokatsu.kataoka}@aist.go.jp

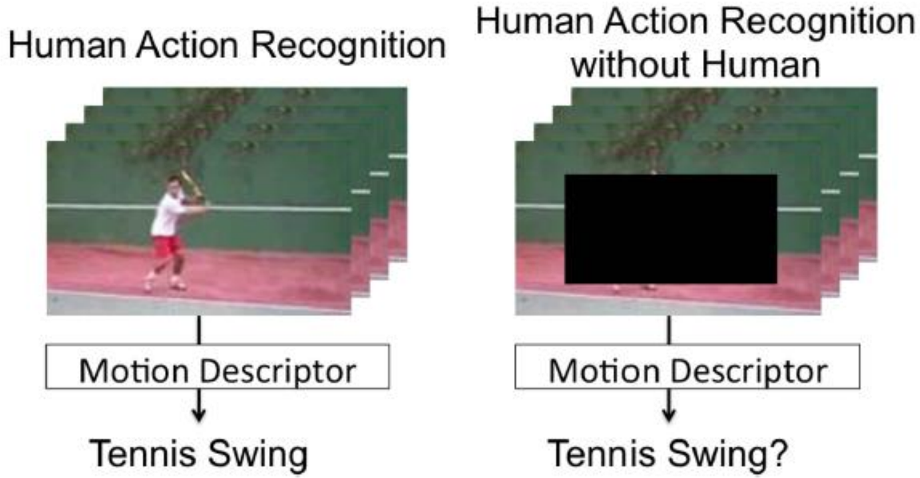


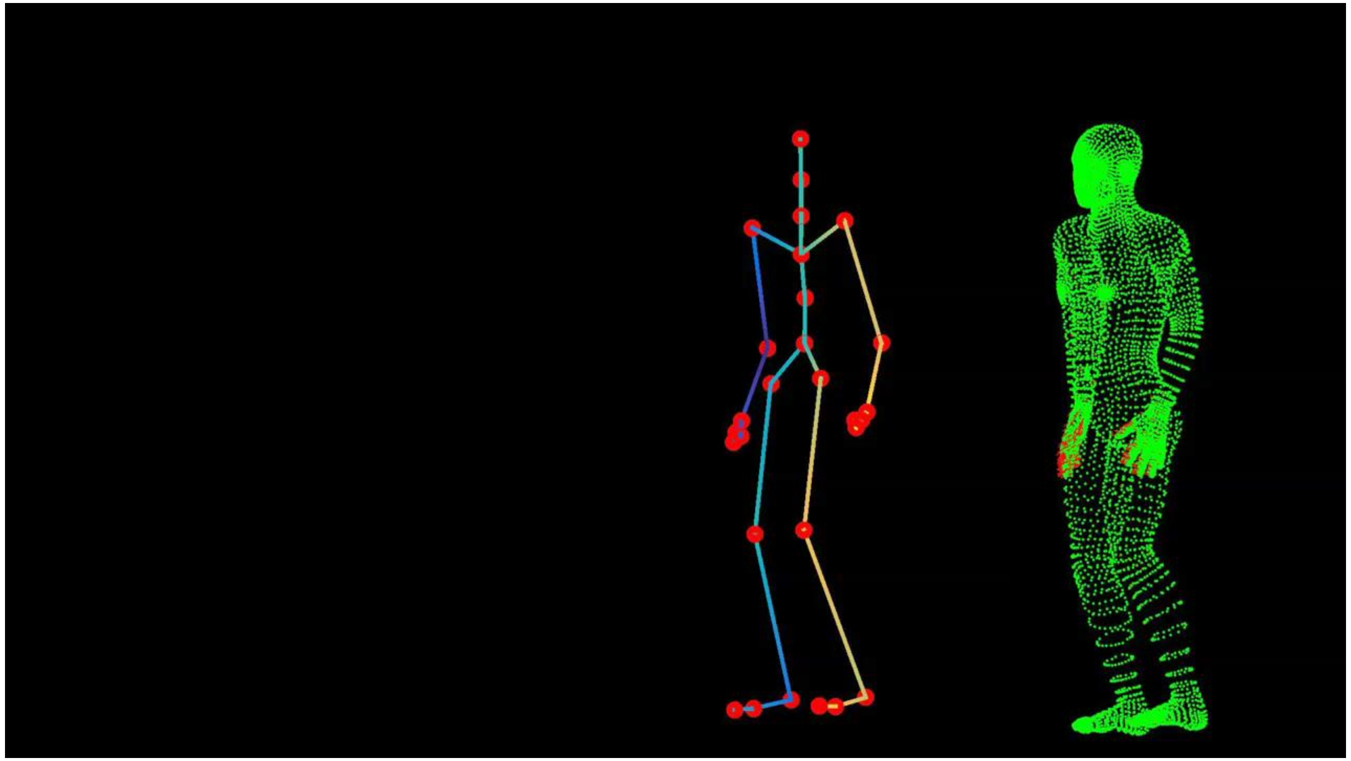
Table 2. Performance rate of human action recognition with or without a human

With or without a human	Stream	% on UCF101 (split 1)
With human	Spatial stream	51.26
	Temporal stream	40.50
	Two-stream	**56.91**
Without human	Spatial stream	45.33
	Temporal stream	26.80
	Two-stream	**47.42**

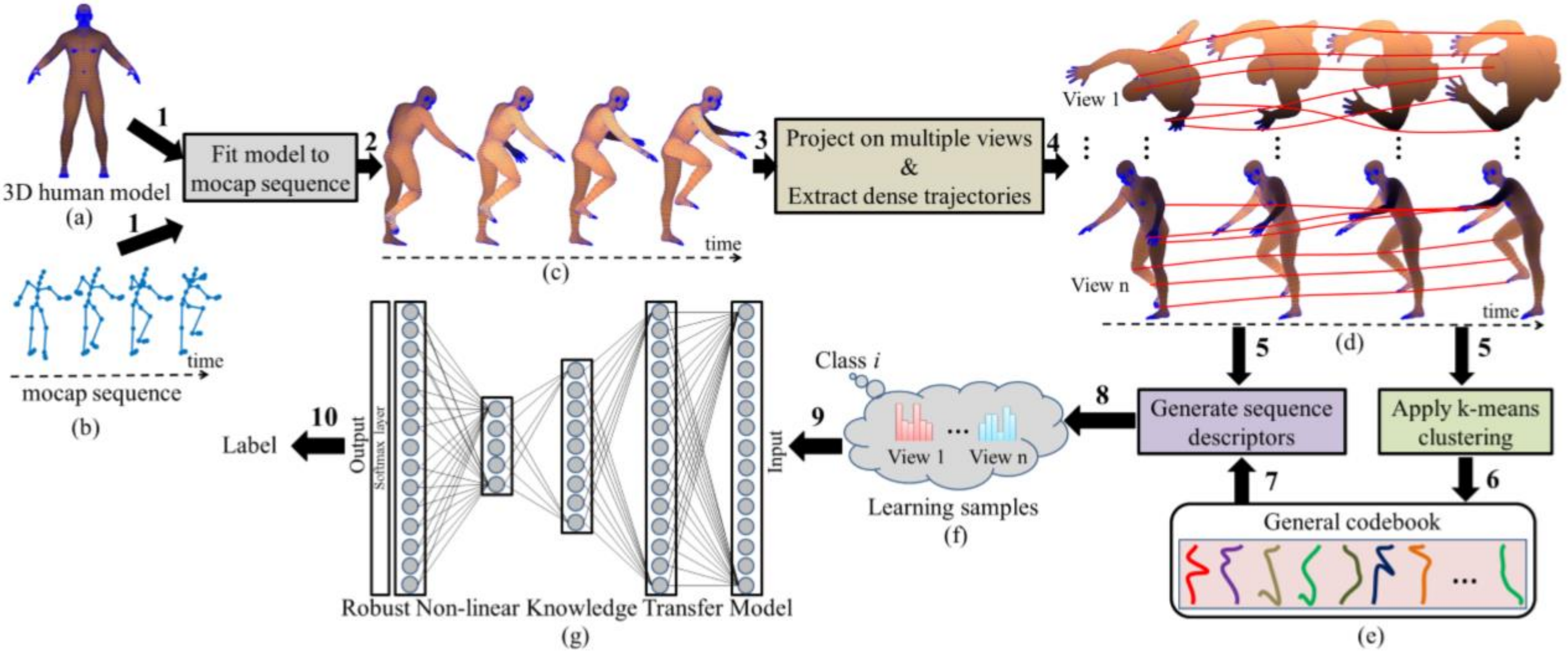
ECCV 2016 workshops

Learning from Synthetic Trajectories

- Trajectories represent motion – not the background or anything else
- Can be generated from MoCap (skeleton) data which is widely available e.g. CMU MoCap
- Fit synthetic 3D humans to MoCap data and generate motion trajectories
- The trajectories can be projected on different camera viewpoints (180)
- Use dummy action labels

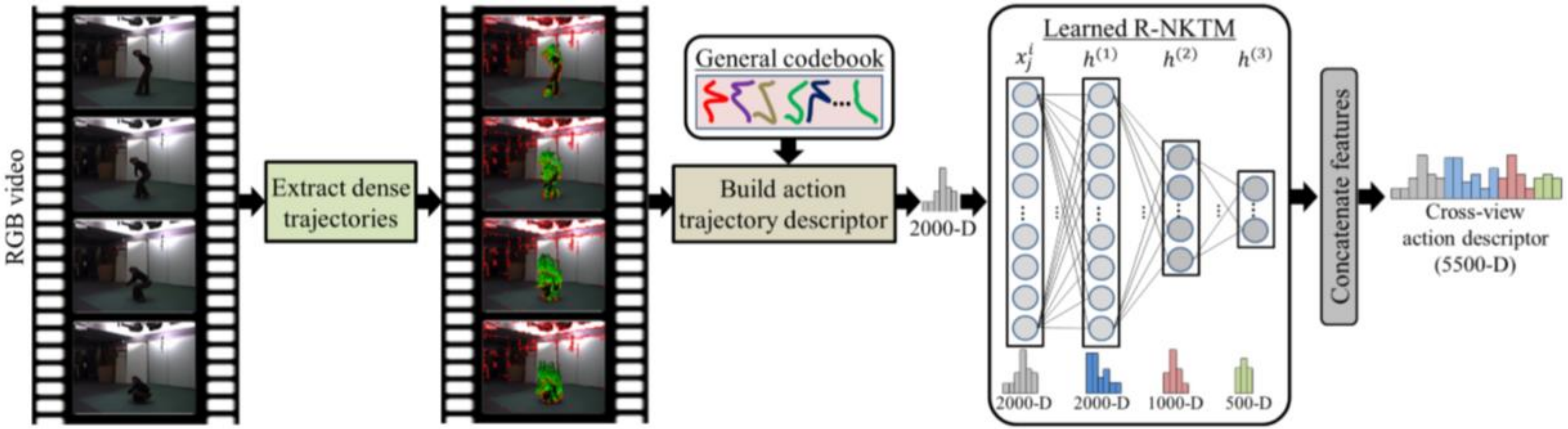


Non-linear Knowledge Transfer



H. Rahmani, **A. Mian** and M. Shah, ***Learning a deep model for human action recognition from novel viewpoints***, PAMI, 2018.

Action Recognition in Real Videos

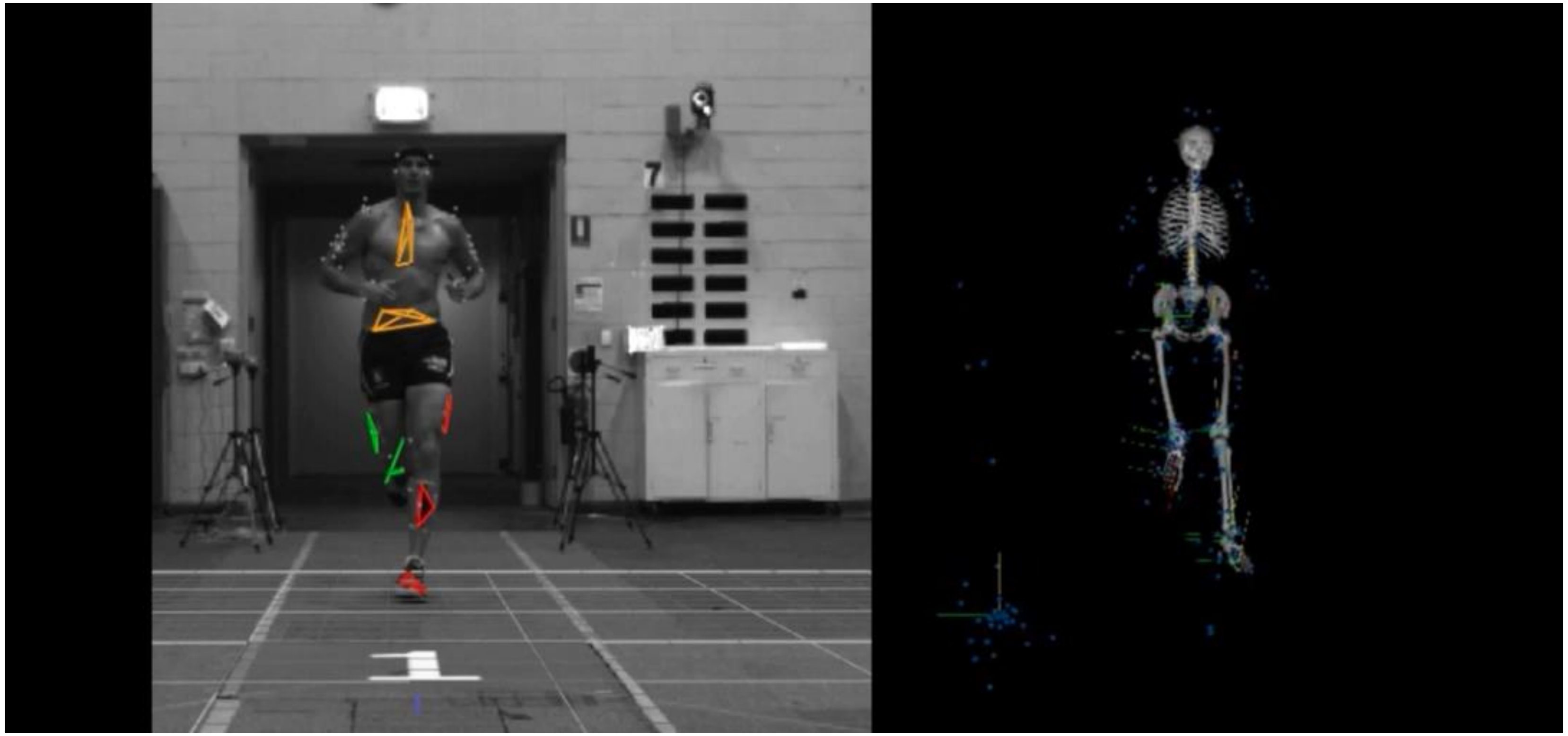


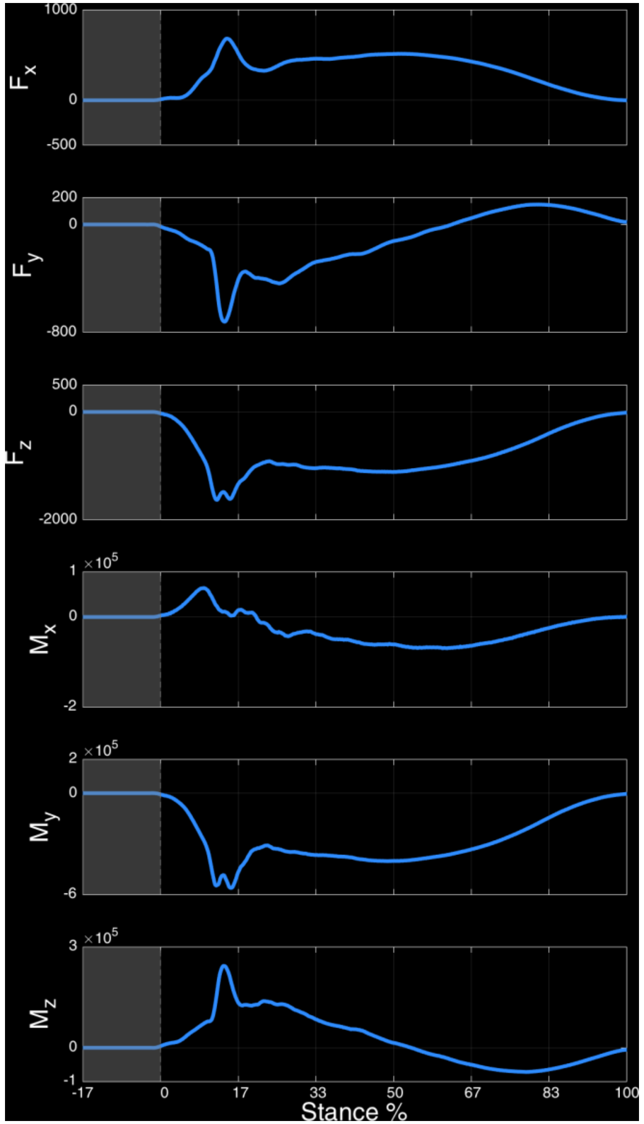
- Dense Trajectories are extracted from real videos and passed through the learned model
- The output of the model is used for action recognition

Human Performance Optimization

- I will show R-NKTM results later
- Let me first show some interesting applications of trajectory based motion analysis in sports

Human Motion Analysis



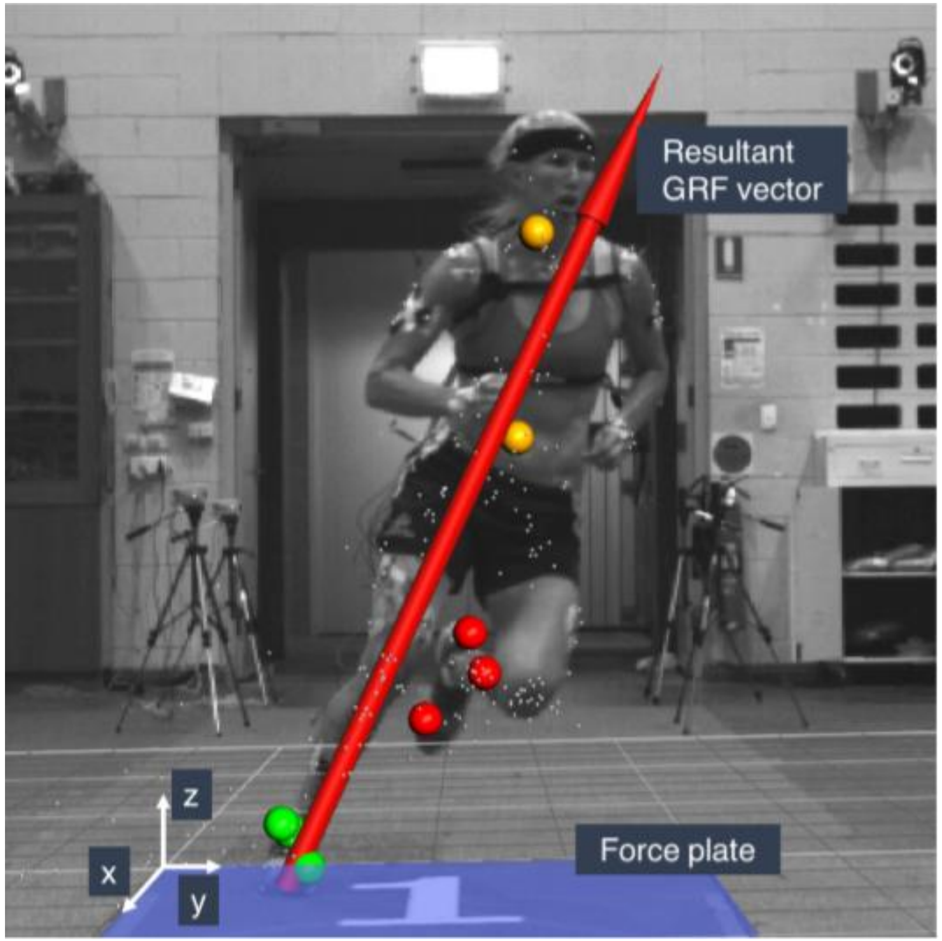


Ground Reaction Forces/Moments

- Estimating forces and moments exerted on the ground (& knee) are critical to
  - injury prevention
  - optimizing biological motion

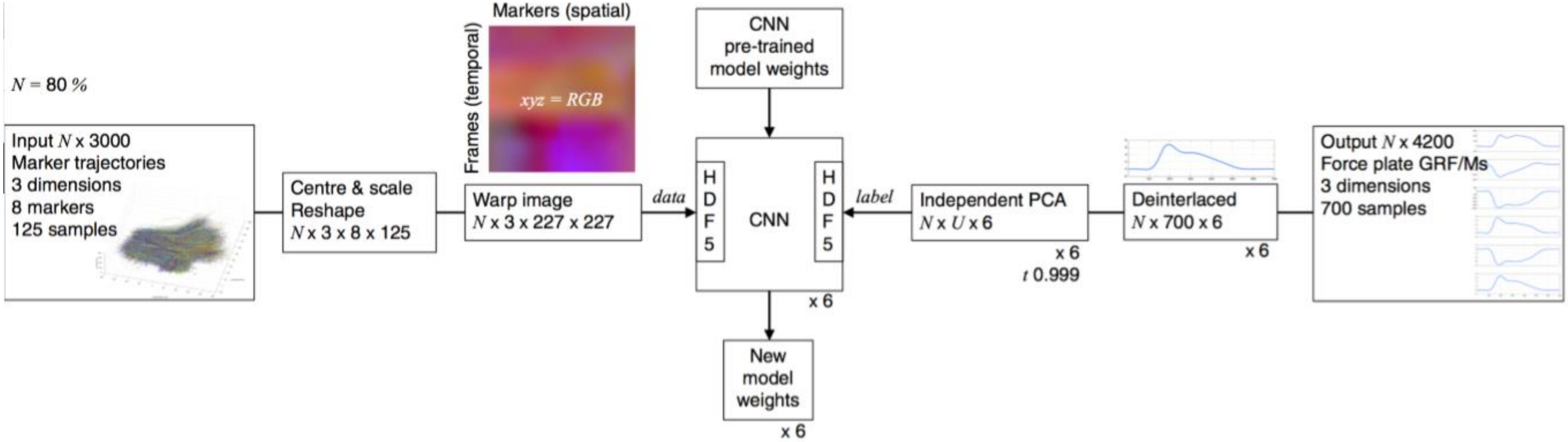
- These measurements can only be done inside a lab using expensive force plate

- To be able to perform these measurements without the force-plate is ground breaking because we can bring the capability outside the lab



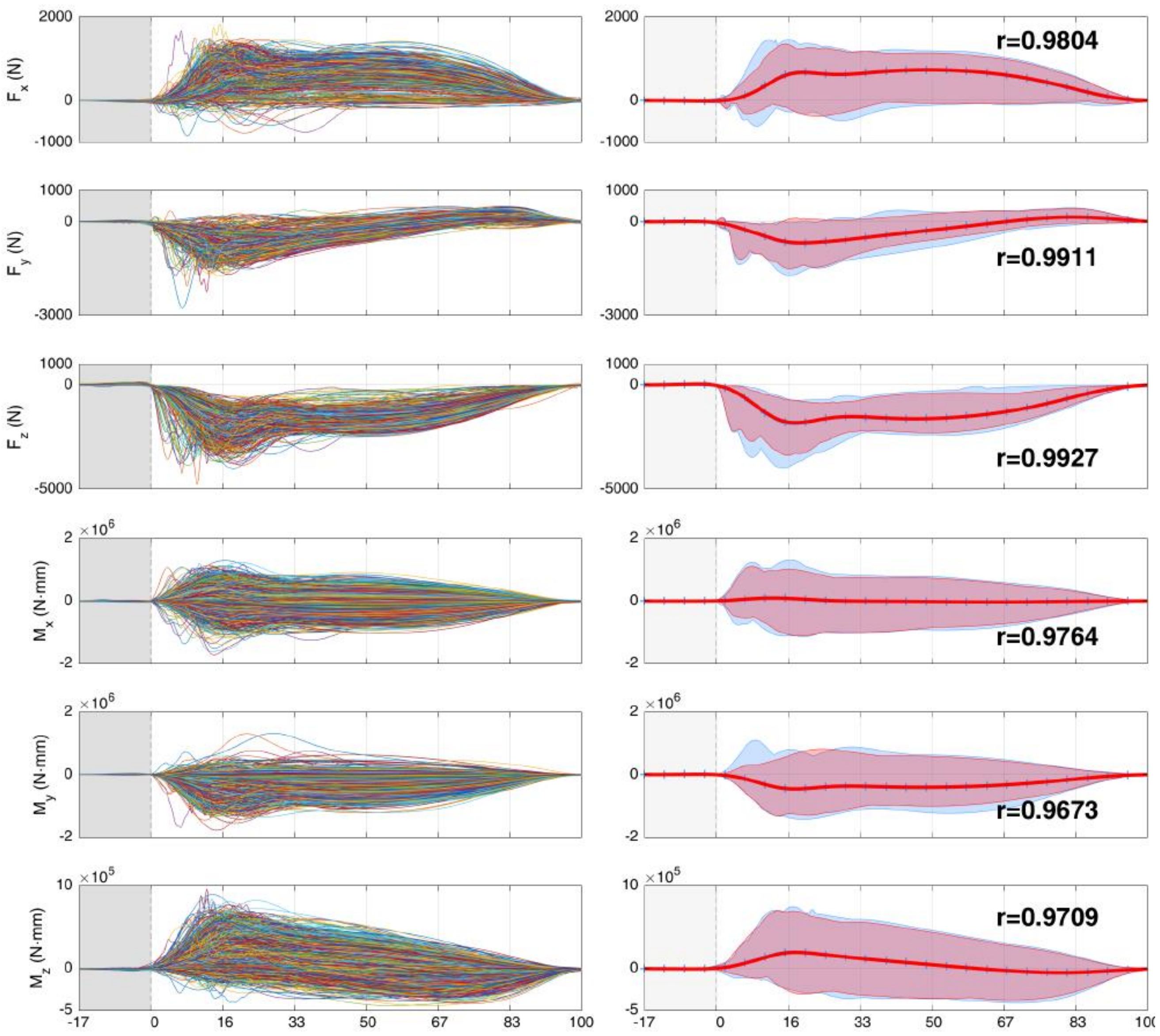
Marker Trajectory → GRF/Ms

By converting marker data to an image,
we can transfer CNNs (trained for image classification)
to estimate ground reaction forces and moments

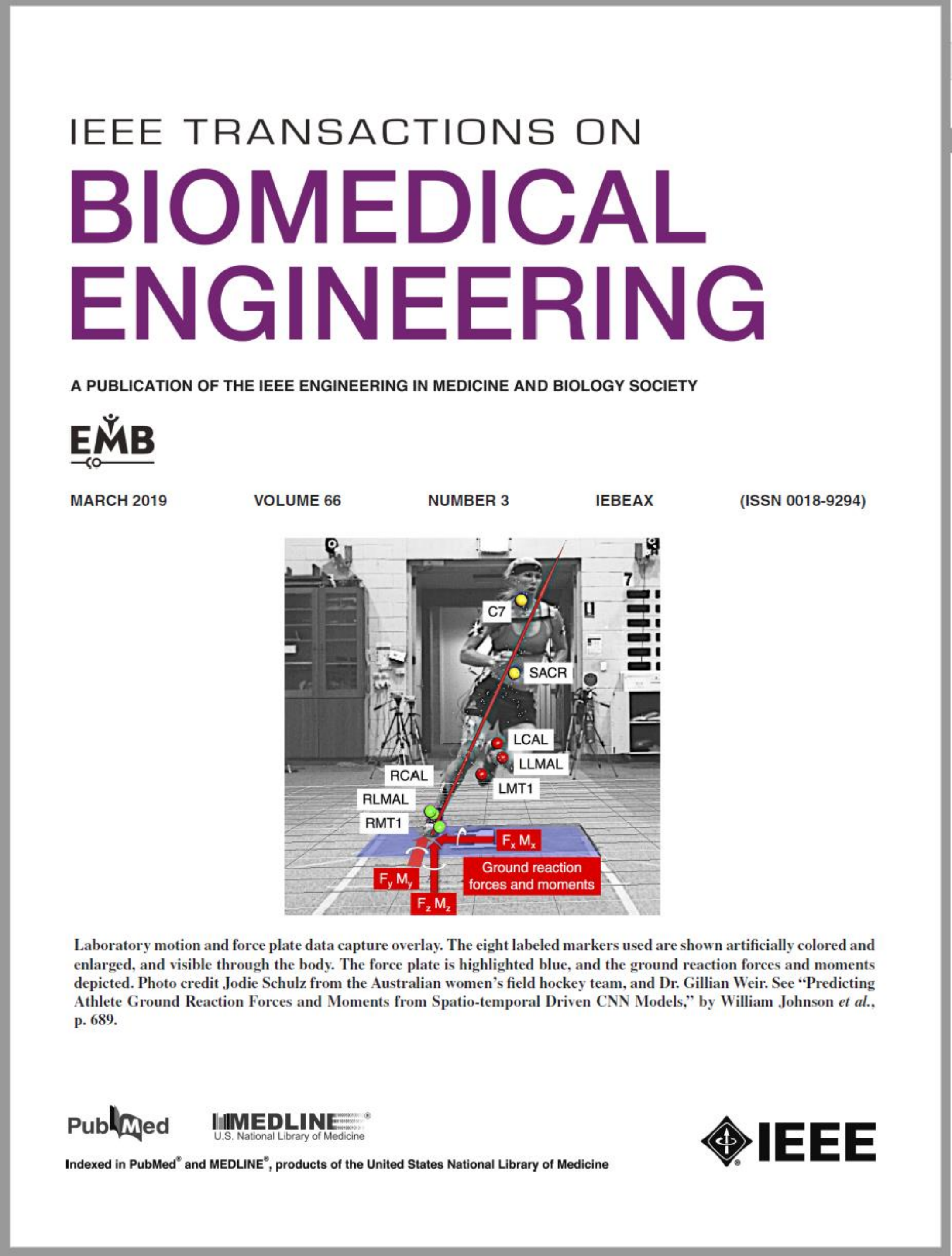


Predicting Athlete Ground Reaction Forces and Moments from Spatio-temporal Driven CNN Models
William Johnson, Jacqueline Alderson, David Lloyd and **Ajmal Mian,** IEEE Trans on Biomedical Engineering (TBME), Mar 2019.

GRF/Ms Estimation Results



Front Cover of March Issue



IEEE TRANSACTIONS ON
BIOMEDICAL ENGINEERING

A PUBLICATION OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY

EMB

MARCH 2019 VOLUME 66 NUMBER 3 IEBEAX (ISSN 0018-9294)

Laboratory motion and force plate data capture overlay. The eight labeled markers used are shown artificially colored and enlarged, and visible through the body. The force plate is highlighted blue, and the ground reaction forces and moments depicted. Photo credit Jodie Schulz from the Australian women's field hockey team, and Dr. Gillian Weir. See "Predicting Athlete Ground Reaction Forces and Moments from Spatio-temporal Driven CNN Models," by William Johnson *et al.*, p. 689.

PubMed MEDLINE U.S. National Library of Medicine

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

IEEE

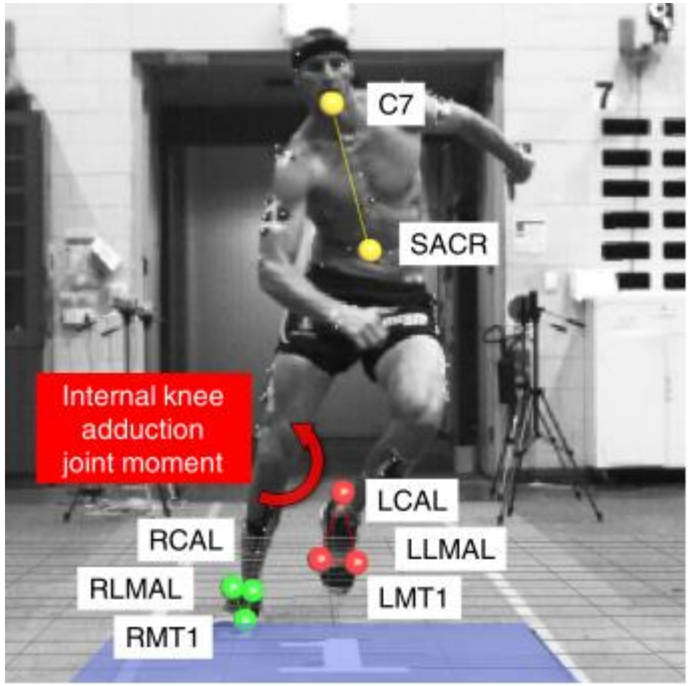
Full Citation

William Johnson, Jacqueline Alderson, David Lloyd, Ajmal Mian, "Predicting Athlete Ground Reaction Forces and Moments From Spatio-Temporal Driven CNN Models", IEEE TBME, vol. 66(3): 689 – 694, 2019.

Follow up Work in Biomechanics

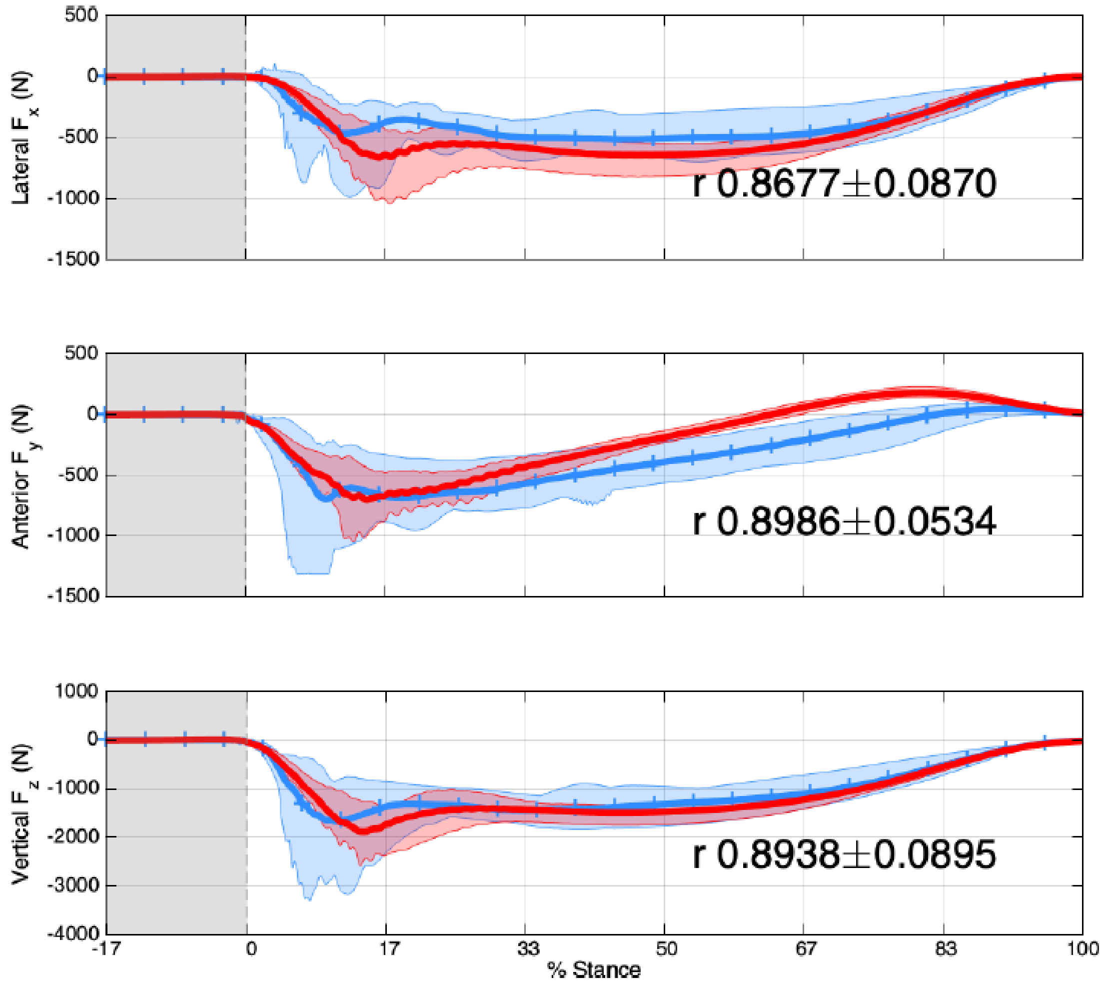
"On-field player workload exposure and knee injury risk monitoring via deep learning", William Johnson, Ajmal Mian, David Lloyd, Jacqueline Alderson, arXiv:1809.08016 , **2019**

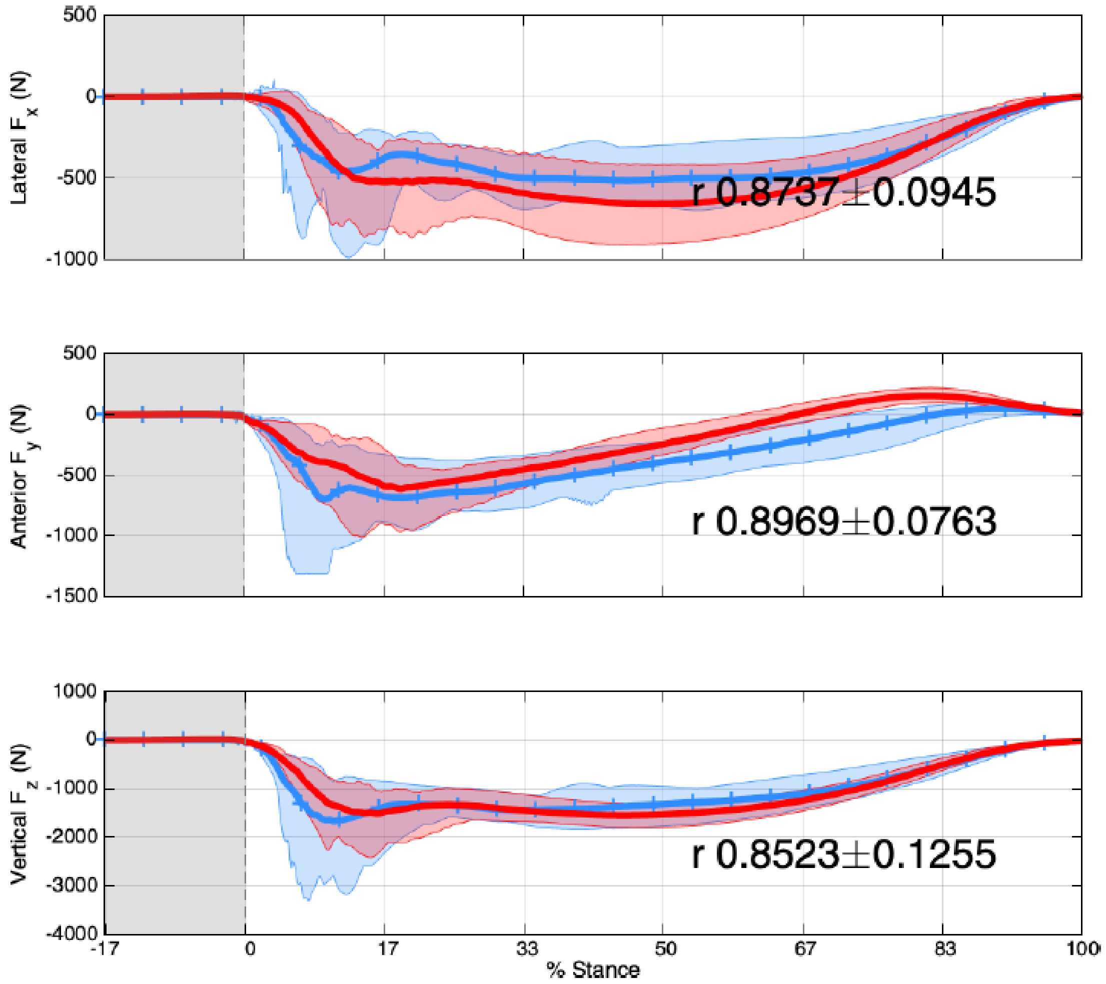
Can we estimate GRF/Ms from IMUs?



Multidimensional ground reaction forces and moments from wearable sensor accelerations via deep learning
W.R. Johnson, A. Mian, M. Robinson, J. Verheul, D. Lloyd, J. Alderson, arXiv:1903.07221 **, 2019**

Correlations drop but still good





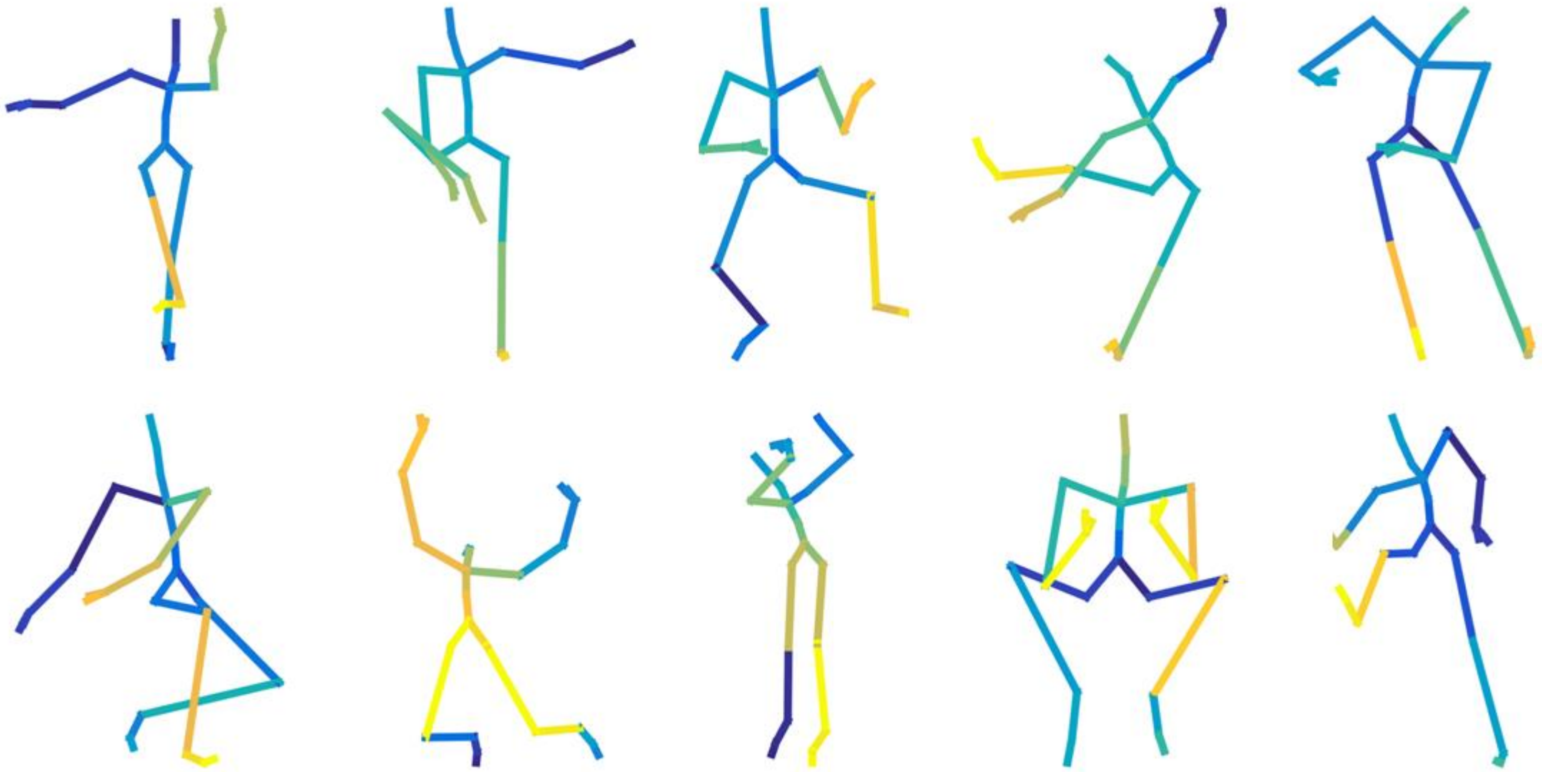
Video-Based

Video based human action recognition

Full 3D mesh human pose recovery from monocular video

Deep Affinity Network for multiple object tracking in video

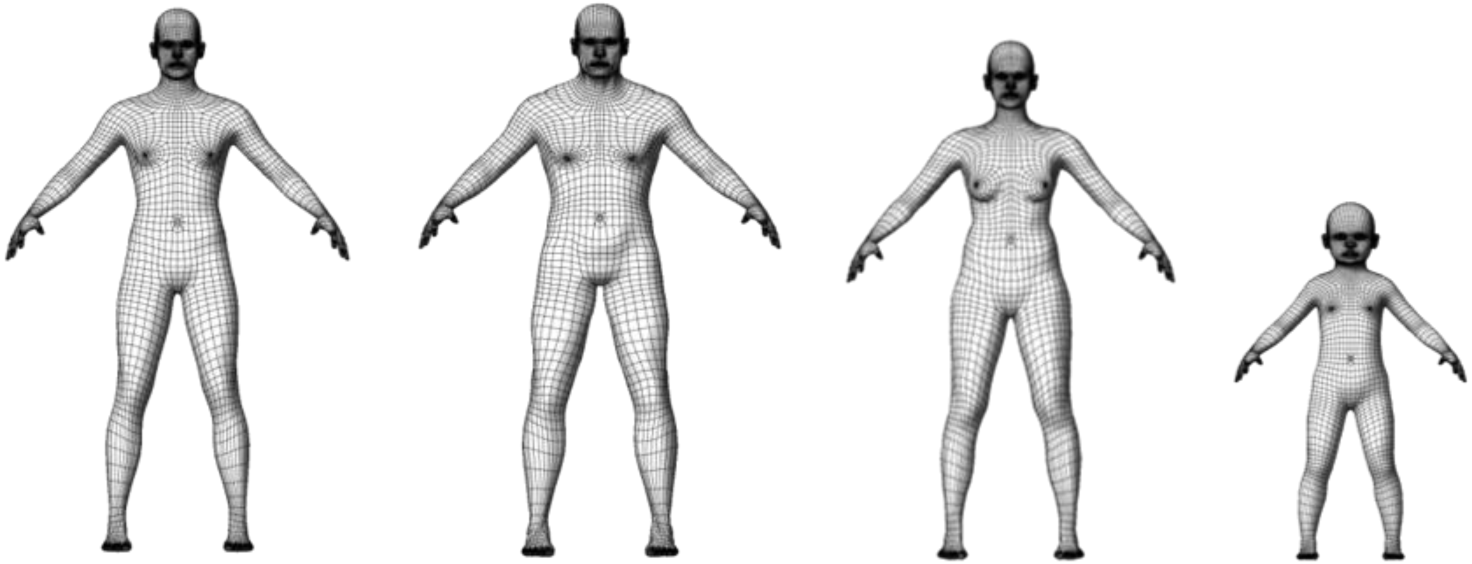
Learning Human Pose Models



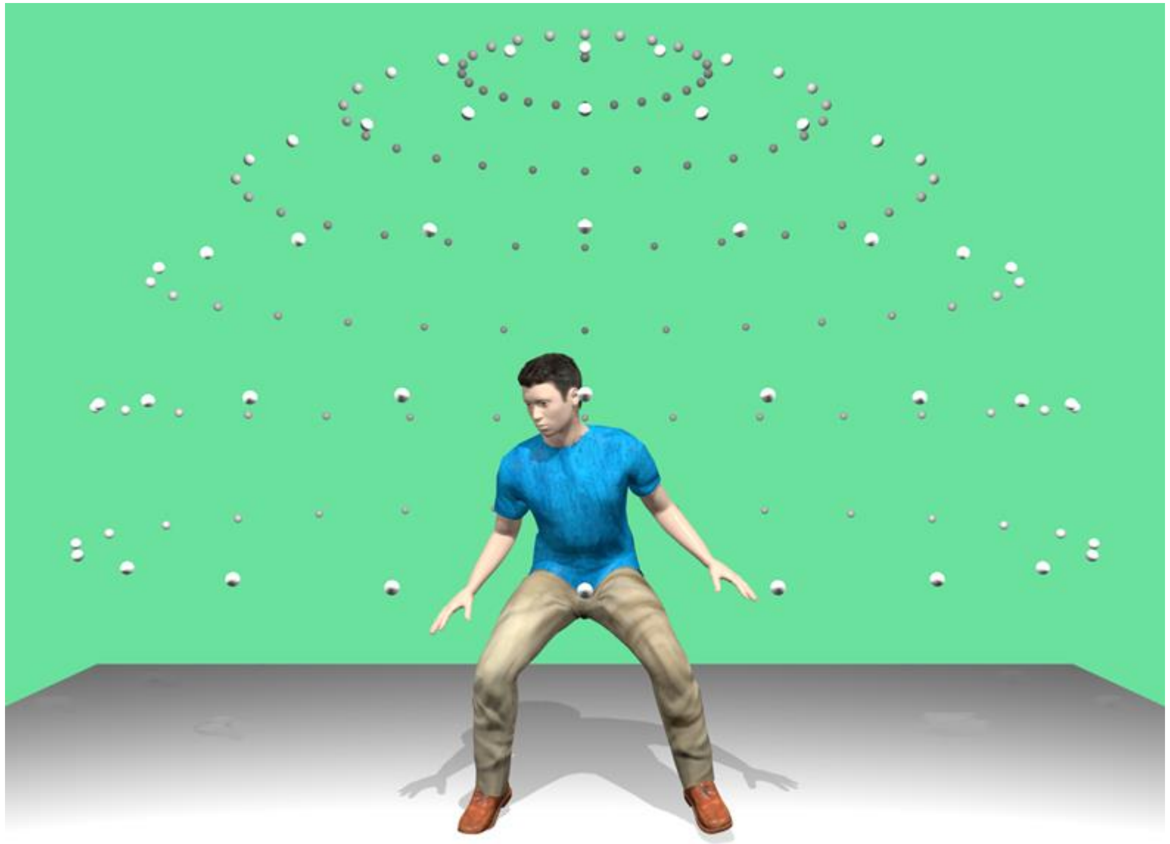
- Learning a CNN that will map an input image to its corresponding pose (discrete mapping i.e. classification)
- Using the CMU MoCap data again
- Cluster with a skeleton distance metric to select N representative human poses

3D Humans

- MakeHuman software
- Size
- Gender
- Age
- Basic clothing



Multiple Camera Viewpoints + Clothes



Complete Pipeline

- Generate realistic images and videos in Blender

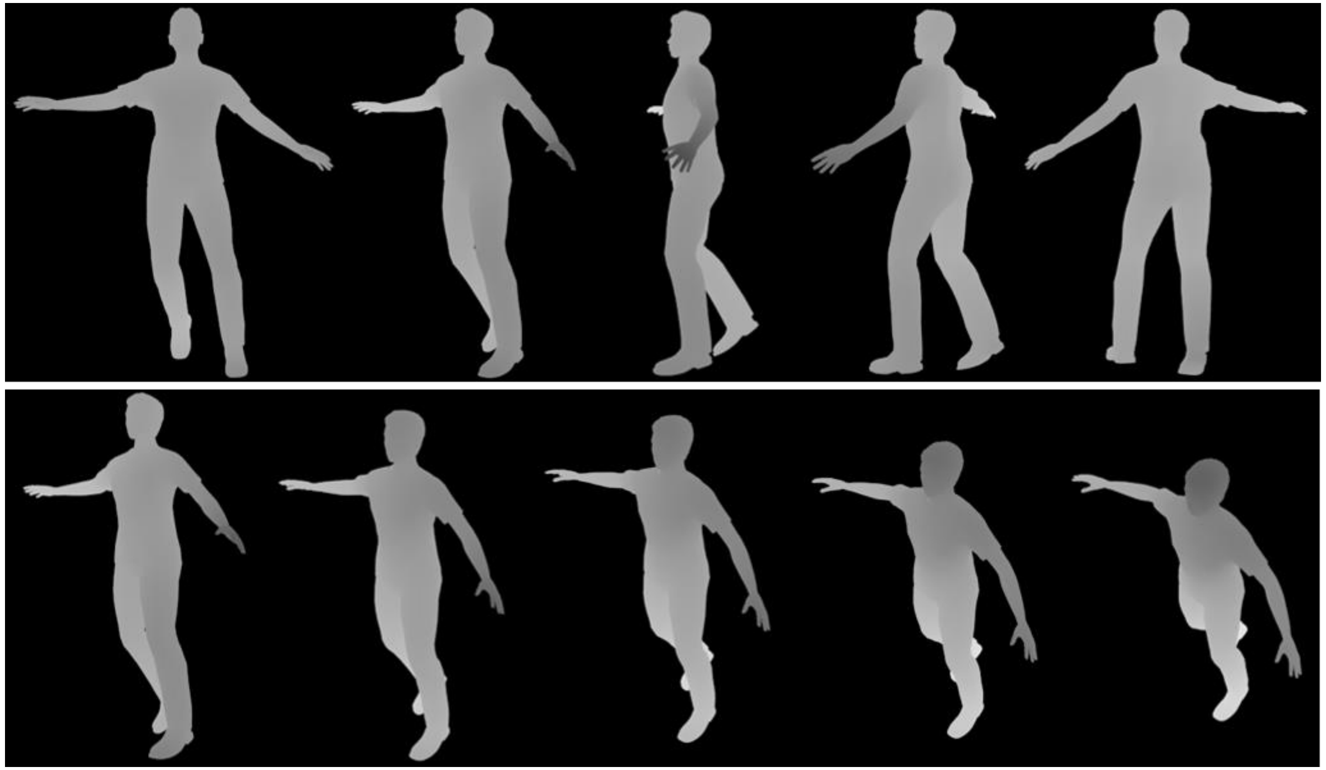
Learning Human Pose Models from Synthesized Data for Robust RGB-D Action Recognition
Jian Liu, Hossein Rahmani, Naveed Akhtar, Ajmal Mian, IJCV 2019 https://arxiv.org/abs/1707.00823

Sample Images with Variations

- 180 cameras x 5 human models x 262 shirts x 183 trousers x 2000 backgrounds x (2 light directions x random intensities)
- Millions of training images with known ground truth

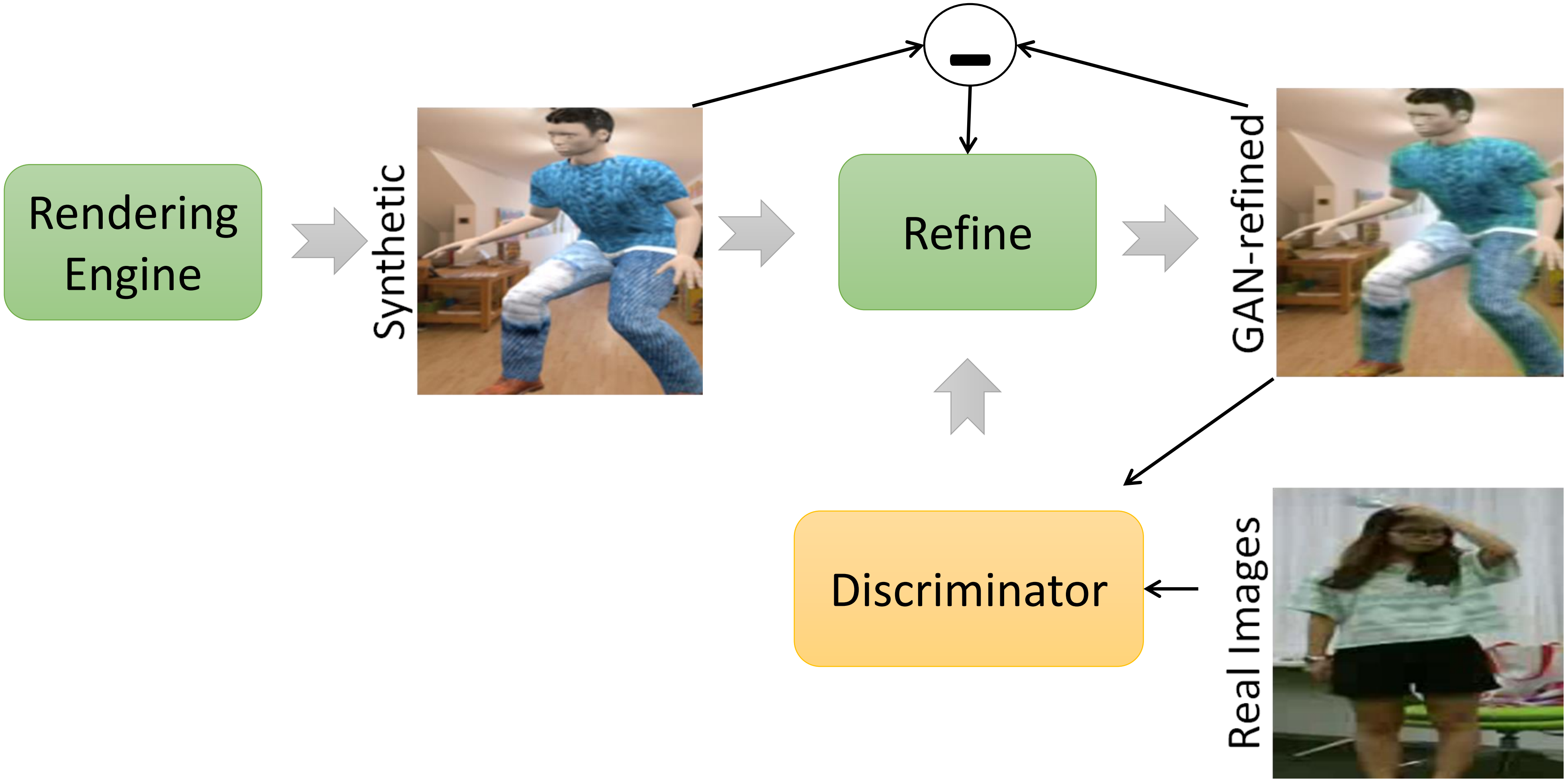
Depth Images

- We also get depth images from Blender

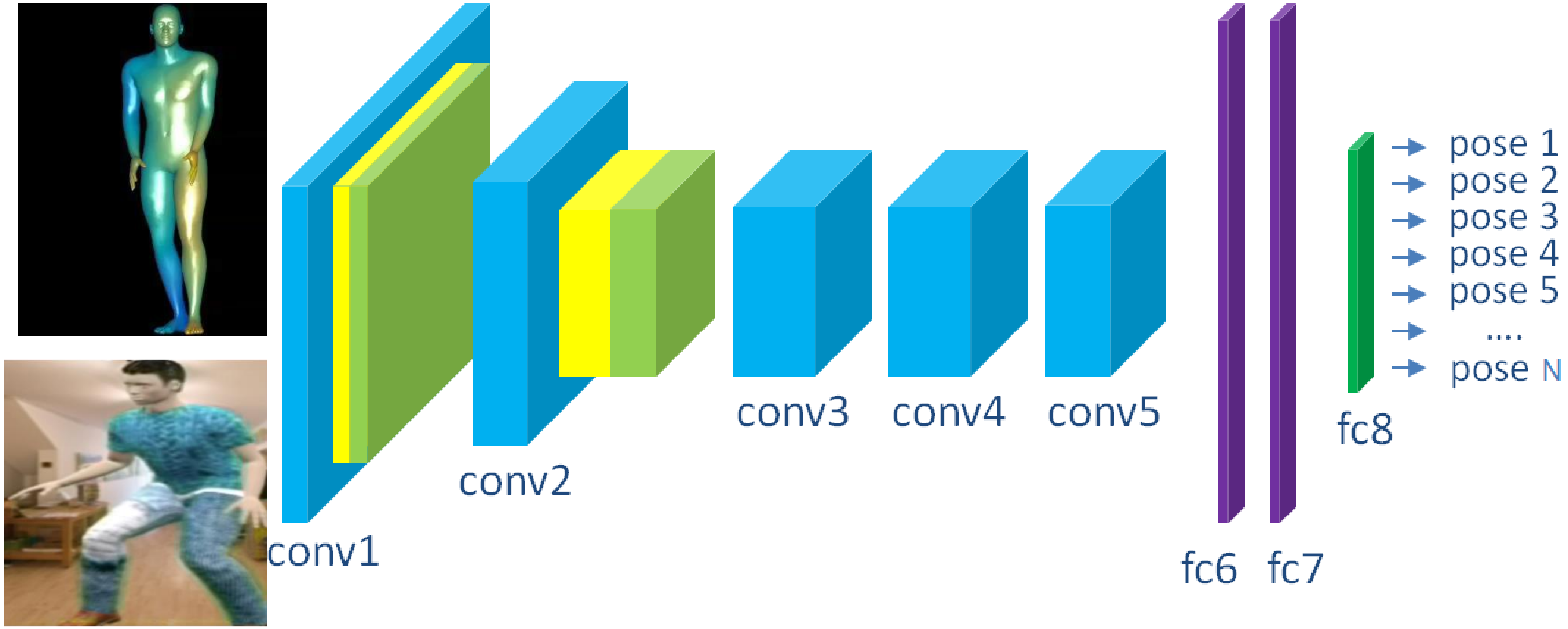


Generative Adversarial Training

- Need to minimize distribution gap between real and synthetic images

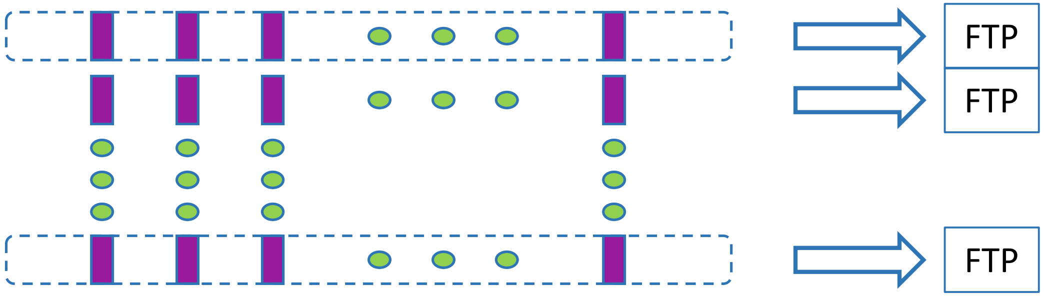


CNN to Map RGB-D Images to N Poses

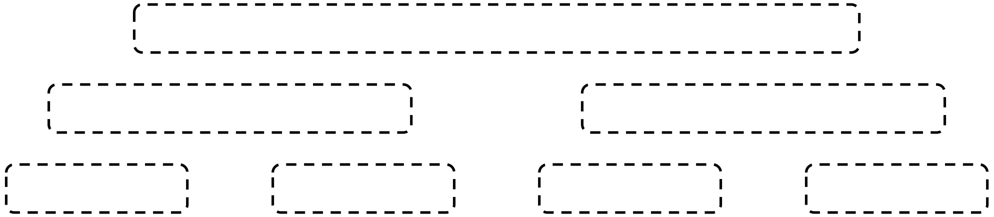


Temporal Modeling

- Apply Fourier analysis over CNN output features



- We use a three layer Fourier Temporal Pyramid



Select a CNN Architecture

- Find the model with highest accuracy + lowest output feature dimensionality
- GoogleNet (inception v-1) used for remaining experiments

Network	Layer	Dimension	HPM_{RGB}	HPM_{3D}
UWA3D Multiview Activity-II				
AlexNet	fc7	4096	61.2	72.1
ResNet-50	pool5	2048	**65.4**	74.0
GoogLeNet	pool5	**1024**	64.7	**74.1**
Northwestern-UCLA Multiview				
AlexNet	fc7	4096	69.9	78.7
ResNet-50	pool5	2048	75.7	77.3
GoogLeNet	pool5	**1024**	**76.4**	**79.8**

Does GAN-Refinement Help?

- Raw synthetic images already perform quite well
- GAN-refinement the accuracy further improves
- Improvement is more for RGB

Training Data	HPM_{RGB}	HPM_{3D}
UWA3D Multiview Activity-II		
Raw synthetic images	64.7	73.8
GAN-refined synthetic images	**68.0**	**74.8**
Northwestern-UCLA Multiview		
Raw synthetic images	76.4	78.4
GAN-refined synthetic images	**77.8**	**79.7**

Comparison with SURREAL Data

Method	Training Data	$V^{\text{challenge}}$	Mean
UWA3D Multiview Activity-II			
$\mathrm{HPM}_{\mathrm{RGB}}$	SURREAL	61.6	67.4
$\mathrm{HPM}_{\mathrm{RGB}}$	Proposed data	**69.0**	**68.0**
$\mathrm{HPM}_{\mathrm{3D}}$	SURREAL	65.8	72.1
$\mathrm{HPM}_{\mathrm{3D}}$	Proposed data	**74.7**	**74.8**

Method	Training	UWA3D
GoogLeNet	without synthetic data	62.8
$\mathrm{HPM}_{\mathrm{RGB}}$	with synthetic data	**68.0**
C3D†	with synthetic data	↑2.3
LRCN†	with synthetic data	↑3.5

Results on UWA-Multiview-II Dataset

- 20 subjects
- 30 actions
- 4 viewpoints
- RGB-D videos (Kinect-1)
- 640×480 RGB resolution
- 320×240 Depth resolution

Method	Data	$V_{1,2}^{3}$	$V_{1,2}^{4}$	$V_{1,3}^{2}$	$V_{1,3}^{4}$	$V_{1,4}^{2}$	$V_{1,4}^{3}$	$V_{2,3}^{1}$	$V_{2,3}^{4}$	$V_{2,4}^{1}$	$V_{2,4}^{3}$	$V_{3,4}^{1}$	$V_{3,4}^{2}$	Mean
Baseline														
DVV [26]	Depth	35.4	33.1	30.3	40.0	31.7	30.9	30.0	36.2	31.1	32.5	40.6	32.0	33.7
Action Tube [10]	RGB	49.1	18.2	39.6	17.8	35.1	39.0	52.0	15.2	47.2	44.6	49.1	36.9	37.0
CVP [70]	Depth	36.0	34.7	35.0	43.5	33.9	35.2	40.4	36.3	36.3	38.0	40.6	37.7	37.3
LRCN [3]	RGB	53.9	20.6	43.6	18.6	37.2	43.6	56.0	20.0	50.5	44.8	53.3	41.6	40.3
AOG [59]	RGB	47.3	39.7	43.0	30.5	35.0	42.2	50.7	28.6	51.0	43.2	51.6	44.2	42.3
Hankelets [25]	RGB	46.0	51.5	50.2	59.8	41.9	48.1	66.6	51.3	61.3	38.4	57.8	48.9	51.8
JOULE [16]	RGB-D	43.6	67.1	53.6	64.4	56.4	49.1	65.7	48.2	76.2	33.5	79.8	46.4	57.0
Two-stream [47]	RGB	63.0	47.1	55.8	60.6	53.4	54.2	66.0	50.9	65.3	55.5	68.0	51.9	57.6
DT [55]	RGB	57.1	59.9	54.1	60.6	61.2	60.8	71.0	59.5	68.4	51.1	69.5	51.5	60.4
C3D [18]	RGB	59.5	59.6	56.6	64.0	59.5	60.8	71.7	60.0	69.5	53.5	67.1	50.4	61.0
nCTE [13]	RGB	55.6	60.6	56.7	62.5	61.9	60.4	69.9	56.1	70.3	54.9	71.7	54.1	61.2
NKTM [38]	RGB	60.1	61.3	57.1	65.1	61.6	66.8	70.6	59.5	73.2	59.3	72.5	54.5	63.5
R-NKTM [40]	RGB	64.9	67.7	61.2	68.4	64.9	70.1	73.6	66.5	73.6	60.8	75.5	61.2	**67.4**
Proposed														
HPM_{RGB}	RGB	72.4	73.4	64.3	71.9	50.8	62.3	69.9	61.8	75.5	69.4	78.4	66.2	68.0
HPM_{RGB}+Traj	RGB	81.0	78.3	72.9	76.8	67.7	75.7	79.9	67.0	85.1	77.2	85.5	69.9	76.4
HPM_{3D}	Depth	80.2	80.1	75.6	78.7	59.0	69.0	72.1	65.2	84.8	79.1	82.5	71.1	74.8
HPM_{RGB}+HPM_{3D}	RGB-D	79.9	83.9	76.3	84.6	61.3	71.3	77.0	68.9	85.1	78.7	87.0	74.8	77.4
HPM_{RGB}+HPM_{3D}+Traj	RGB-D	85.8	89.9	79.3	85.4	74.4	78.0	83.3	73.0	91.1	82.1	90.3	80.5	**82.8**

Table 3. Action recognition accuracy (%) on the UWA3D Multiview-II dataset. $V_{1,2}^{3}$ means that view 1 and 2 were used for training and view 3 alone was used for testing

NTU RGB-D Dataset

- 56,880 videos (Kinect v2)
- 40 human subjects
- 60 actions including 10 multi-person actions
- Changes in viewpoint, sensor height/distance

- We were the first to report RGB only results on this challenging dataset

- Our RGB only method ($HPM_{RGB} + Traj$) achieves higher accuracy than RGB-D methods then

- Our method achieves the highest RGB-D action recognition accuracy

Method	Data type	Cross Subject	Cross View
Baseline			
HON4D (Oreifej and Liu, 2013)	Depth	30.6	7.3
SNV (Yang and Tian, 2014)	Depth	31.8	13.6
HOG-2 (Ohn-Bar and Trivedi, 2013)	Depth	32.4	22.3
Skeletal Quads (Evangelidis et al, 2014)	Joints	38.6	41.4
Lie Group (Vemulapalli et al, 2014)	Joints	50.1	52.8
Deep RNN (Shahroudy et al, 2016a)	Joints	56.3	64.1
HBRNN-L (Du et al, 2015)	Joints	59.1	64.0
Dynamic Skeletons (Hu et al, 2015)	Joints	60.2	65.2
Deep LSTM (Shahroudy et al, 2016a)	Joints	60.7	67.3
LieNet (Huang et al, 2016)	Joints	61.4	67.0
P-LSTM (Shahroudy et al, 2016a)	Joints	62.9	70.3
LTMD (Luo et al, 2017)	Depth	66.2	-
ST-LSTM (Liu et al, 2016)	Joints	69.2	**77.7**
DSSCA-SSLM (Shahroudy et al, 2017)	RGB-D	**74.9**	-
Proposed			
HPM_{RGB}	RGB	68.5	72.9
HPM_{RGB}+Traj	RGB	75.8	83.2
HPM_{3D}	Depth	71.5	70.5
HPM_{RGB}+HPM_{3D}	RGB-D	75.8	78.1
HPM_{RGB}+HPM_{3D}+Traj	RGB-D	**80.9**	**86.1**

Full 3D Human Pose Recovery

Monocular
2D Video

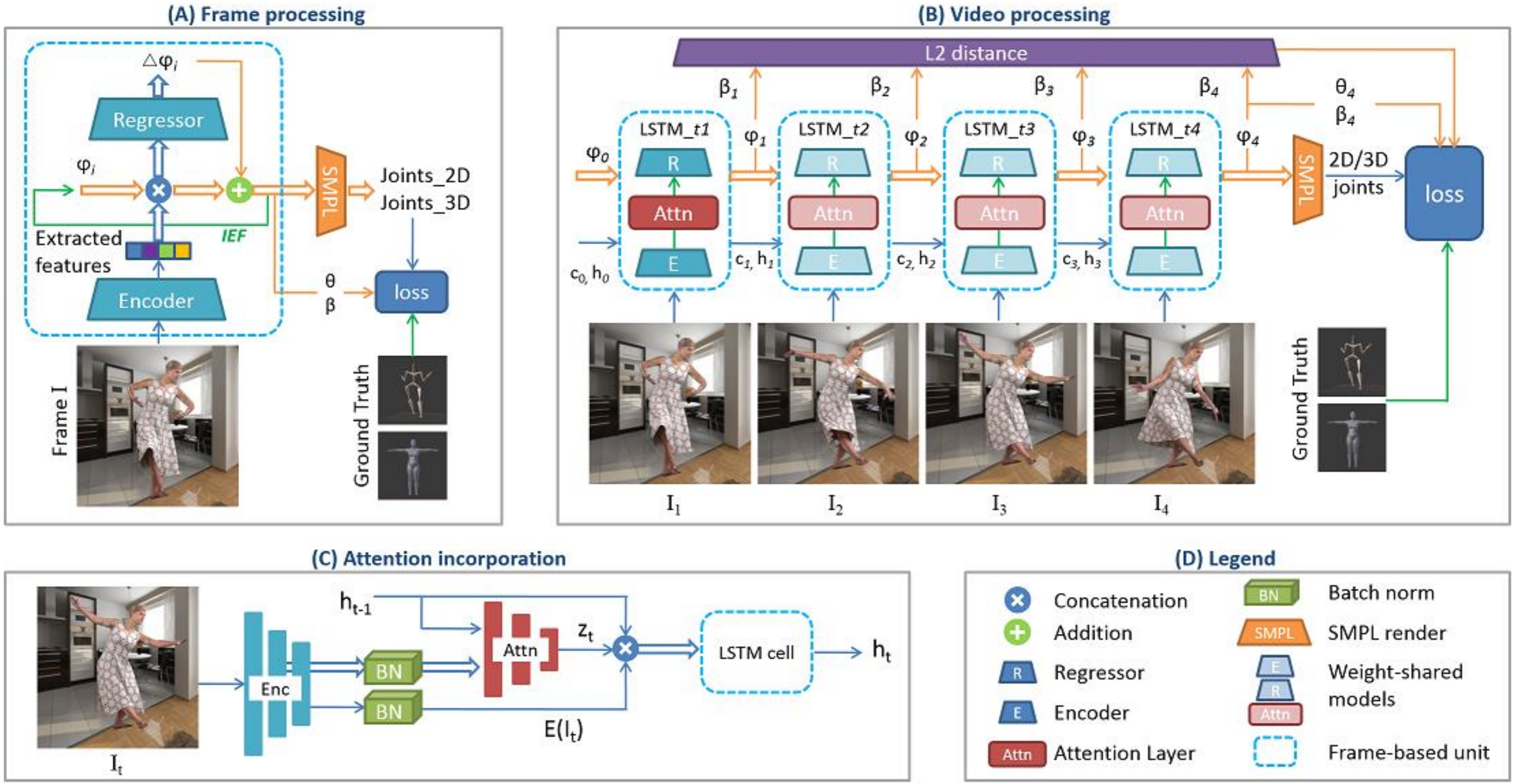
Algorithm X

3D Video
360^o human

For Every Frame

Human shape parameters β
Human pose parameters θ
Global 3D rotation R
2D Translation in image t
Scale s

Architecture for 3D Human Pose Recovery



Network Losses (frame-wise)

2D joint locations loss

$$\mathcal{L}_{proj} = \sum_i ||\chi_i ({}_{2D}\mathbf{J}_i - {}_{2D}\hat{\mathbf{J}}_i)||_1$$

3D joint locations loss

$$\mathcal{L}_{3Djoint} = \sum_i ||{}_{3D}\mathbf{J}_i - {}_{3D}\hat{\mathbf{J}}_i||_2^2$$

SMPL parameters loss

$$\mathcal{L}_{smpl} = \sum_i ||[\boldsymbol{\beta}_i, \boldsymbol{\theta}_i] - [\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\theta}}_i]||_2^2$$

Overall Video-based Loss

- Minimize loss over T frames

$$\mathcal{L}_{shape} = \sum_{t=1}^{T-1} ||\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t||_2^2$$

- Body shape parameters β should not change from frame to frame

$$\mathcal{L} = \sum_{t=1}^{T} \lambda((\mathcal{L}_{proj})_t + \delta(L_{3D})_t) + \mathcal{L}_{shape}$$

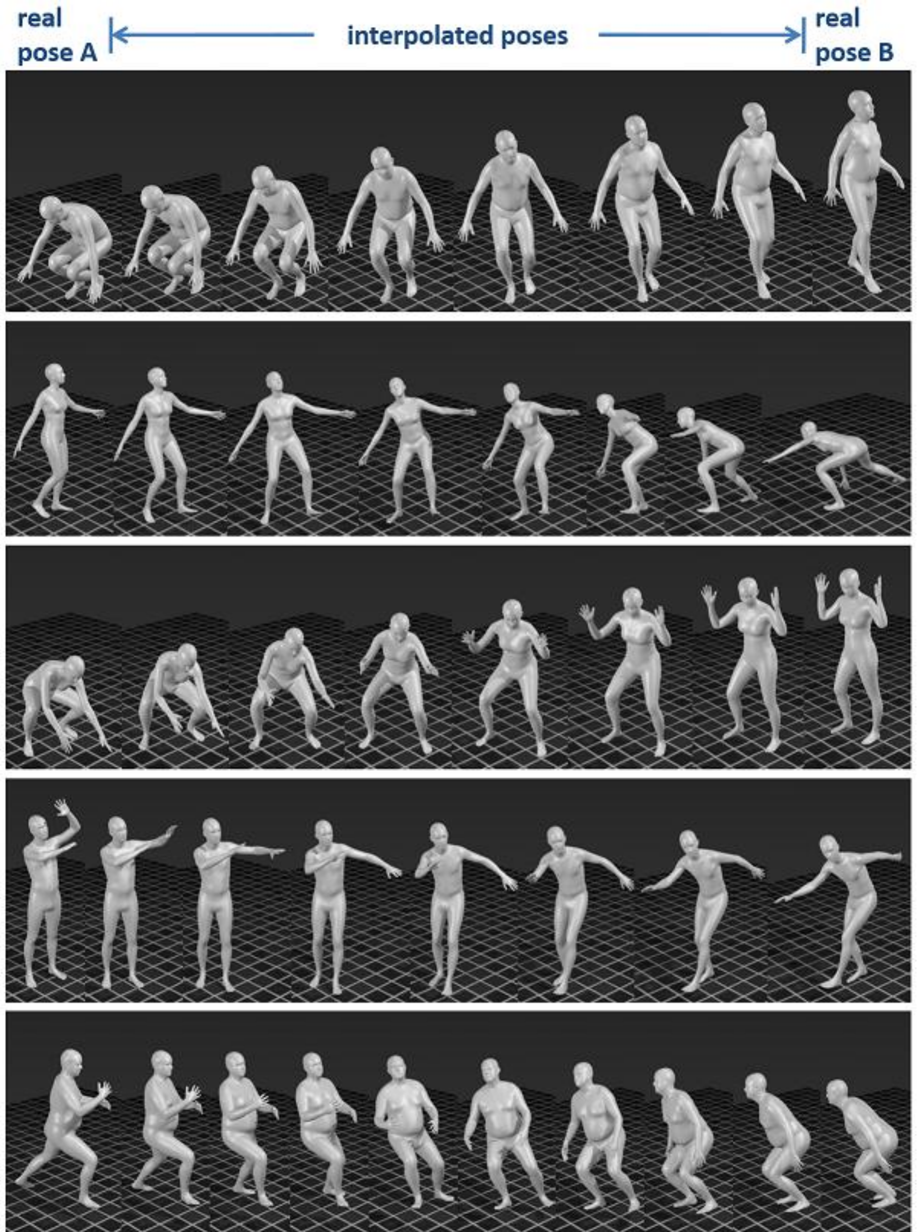
where $\mathcal{L}_{3D} = \mathcal{L}_{3Djoint} + \mathcal{L}_{\theta}$

Training Data ???

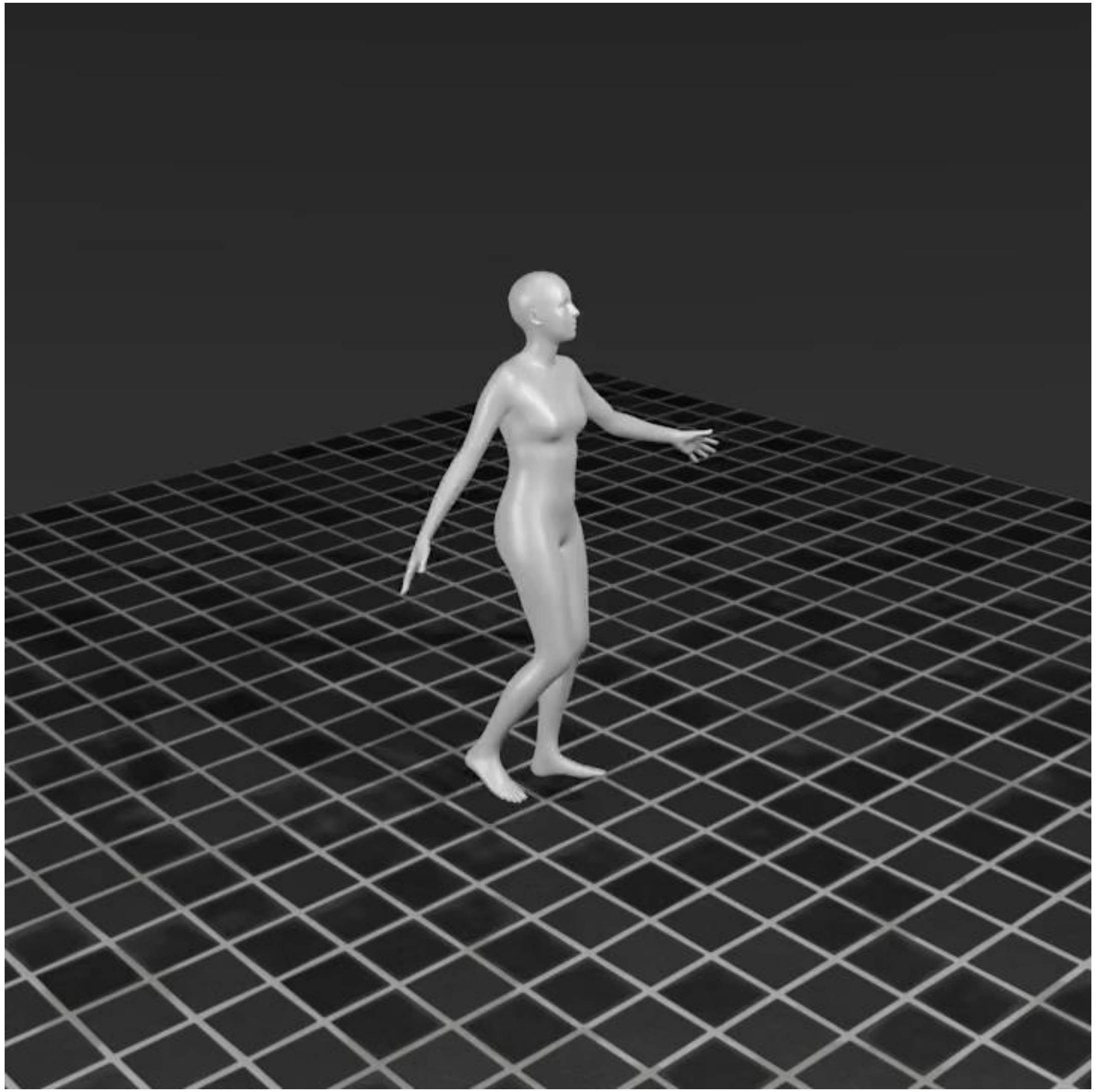
- Use SMPL model to generate humans for varying shapes
  - Linear combinations of shapes
- Use MoCap data to generate varying poses
  - Interpolate poses to generate novel motions
- Clothes??
  - Pasting texture on bodies – unrealistic
  - Rigid clothes – unrealistic
  - Design real clothes and apply a Physics engine to model cloth deformations with human motion and gravity – now you are talking!

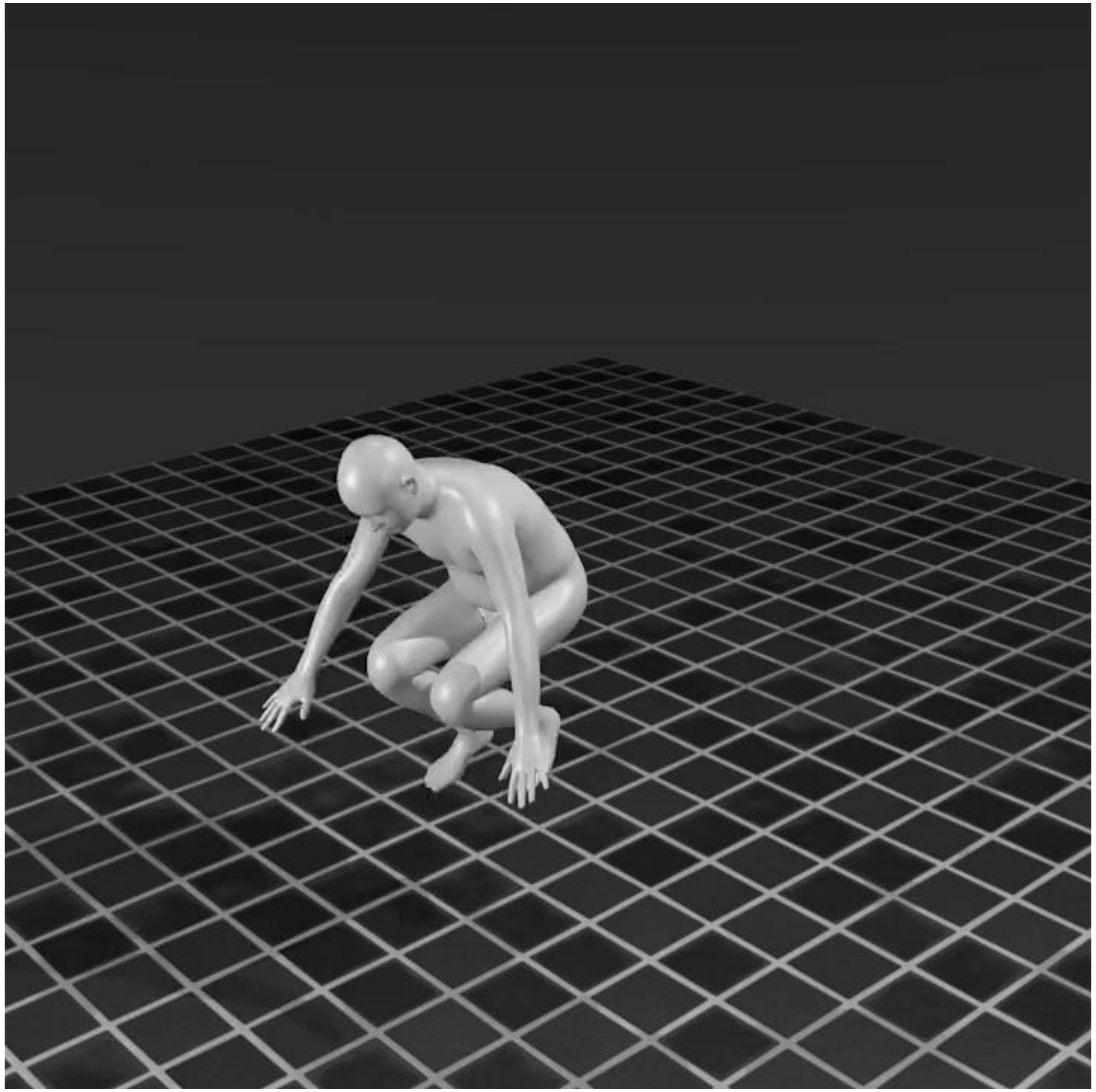
Pose Interpolation

- We can generate novel human motions using Quaternion interpolation between the skeletons of very different poses
- On the rights side are motions that were never performed by anyone
- The transitions are smooth giving realistic motions



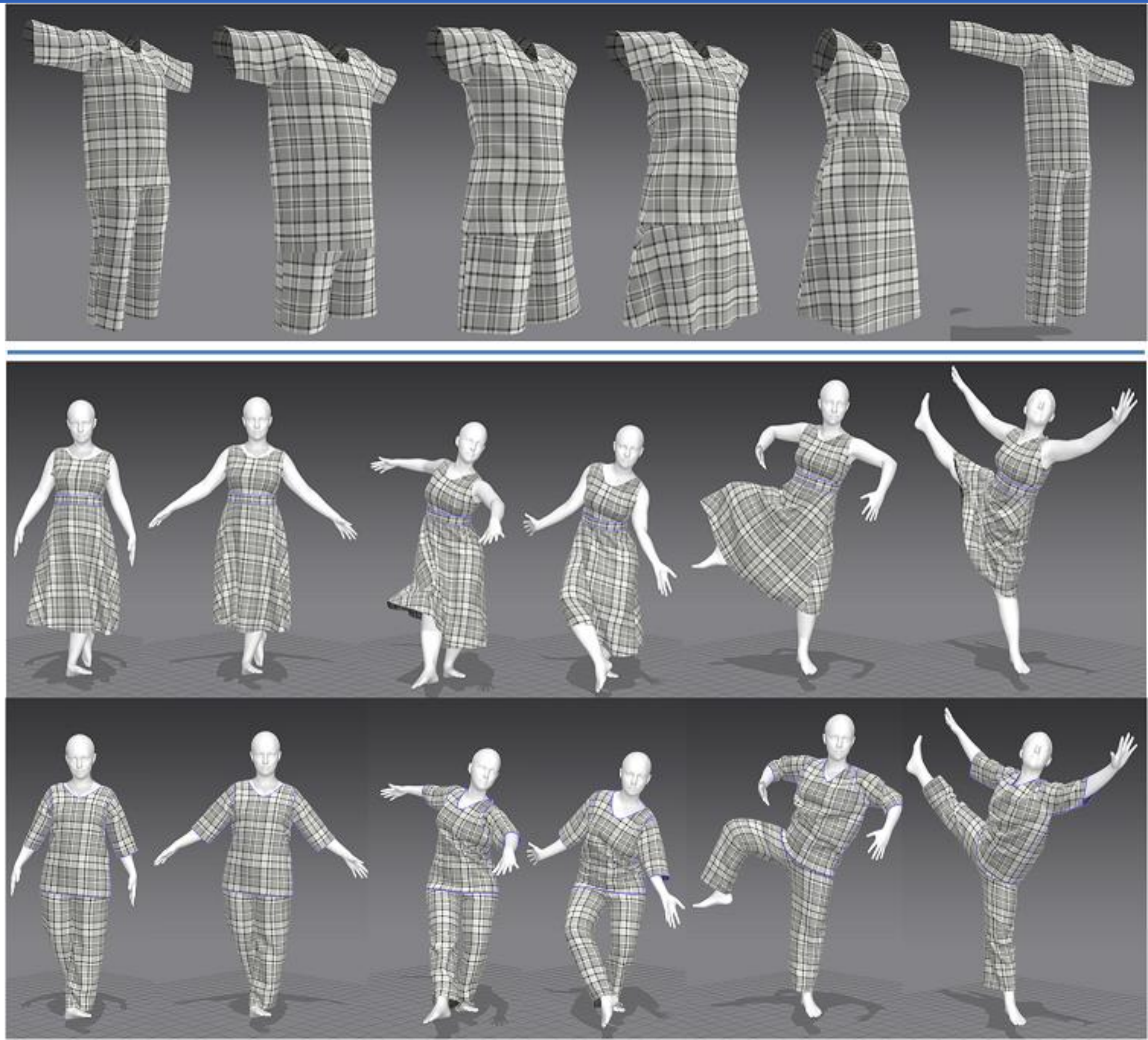
Pose Interpolation Demo Videos





Designing and Rendering Clothes

- MarvelousDesigner (MD7) software is used to design clothes from scratch
- Cloth cutting, sewing and its physics based redering is performed a given MoShed sequence
- Notice how the loose clothing moves with the avatar



Clothing Textures, Lighting & Background

MoShed sequences are rendered in Blender while varying

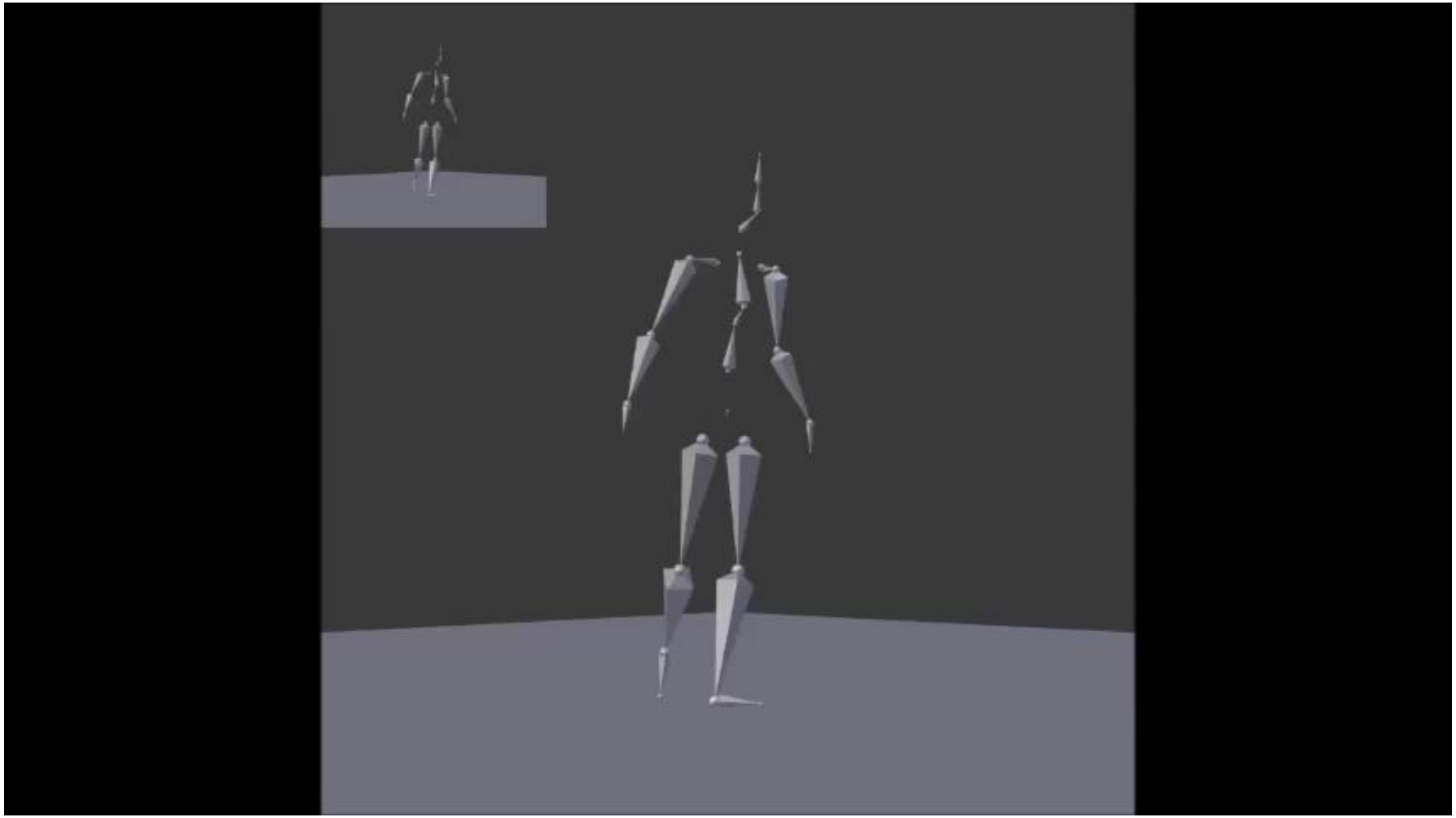
1. Clothing textures
2. Backgrounds
3. Light sources
4. Camera viewpoints

<u>GigaVision</u>

We can also simulate multi-camera, multi-resolution etc

Sample Videos

We show high resolution videos (720x720) for better visualization. Our network requires only 250x250 resolution.



Sample data and code to generate more data is available on GitHub https://github.com/liujianee/MVIPER

Results on Our Generated Data

Procrustes Analysis (PA) Mean Per Joint Position Error (MPJPE)

Method	MPJPE	PA-MPJPE	MPVPE	MRSV_1	MRSV_2
SMPLify [36]	152.1	109.3	1426.9	0.85	0.41
HMR [6]	133.2	81.3	1056.5	0.82	0.36
HMR†	125.6	77.6	923.7	0.76	0.32
MVIPER (ours)	**93.2**	**60.5**	**692.7**	**0.51**	**0.29**

Mean Per Vertex Position Error

$$\mathrm{MPVPE} = \frac{1}{M}\sum_{j=1}^{M}\left(\sum_{i=1}^{N}||\hat{\mathbf{v}}_i - \mathbf{v}_i||_2\right)$$

Mean Running Shape Variation

$$\mathrm{MRSV} = \frac{1}{M}\sum_{i=1}^{M-1}\left(||\hat{\beta}_{i+1} - \hat{\beta}_i||_p\right)$$

[6] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” CVPR, 2018.

[36] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, M. Black, “Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image,” ECCV’16.

Results on Human 3.6M Dataset

- Only 2D joints are available

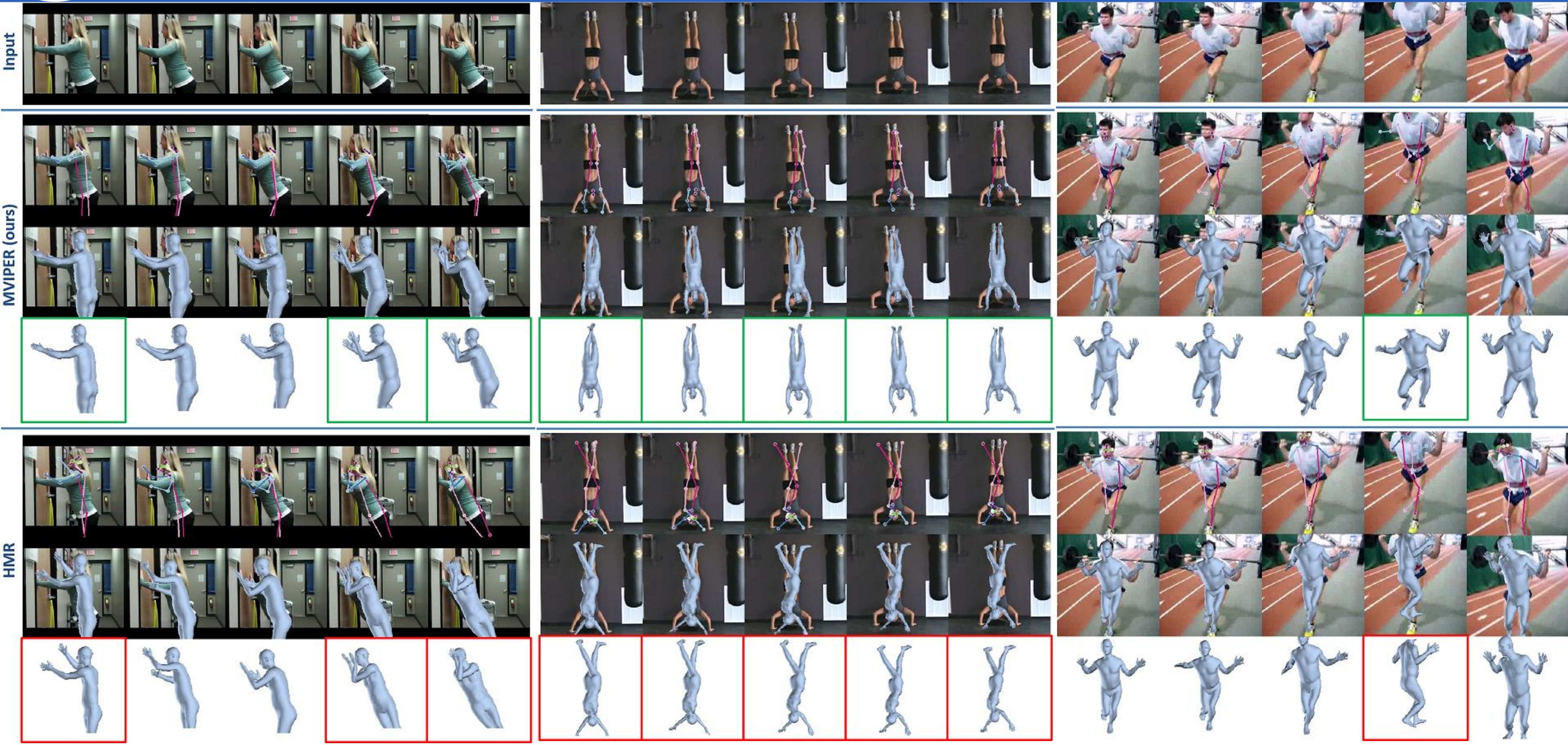
Procrustes Analysis Mean Per Joint Position Error PA-MPJPE (mm)

Protocol-1	Direc.	Discu.	Eat	Greet	Phone	Photo	Pose	Purchase	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Mean
HMR [6]	52.3	54.7	54.3	57.1	60.9	70.4	51.6	49.9	65.7	76.0	58.6	52.5	60.2	**45.2**	**53.6**	57.5
HMR†	51.2	52.4	53.8	56.9	59.9	65.0	50.4	49.2	66.3	73.1	59.2	52.6	60.0	46.6	53.9	56.7
MVIPER (ours)	**48.1**	**48.8**	**49.6**	**55.3**	**53.8**	**63.4**	**49.4**	**48.0**	**58.5**	**67.4**	**54.4**	**52.2**	**59.3**	47.3	54.3	**54.0**

Protocol-2	Direc.	Discu.	Eat	Greet	Phone	Photo	Pose	Purchase	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Mean
SMPLify [36]	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3	137.3	83.4	77.3	79.7	86.8	81.7	82.3
HMR [6]	53.2	56.8	50.4	62.4	54.0	72.9	49.4	51.4	57.8	73.7	54.4	50.0	62.6	47.1	**55.0**	56.7
HMR†	52.1	53.9	51.4	61.1	54.4	66.1	49.6	48.7	58.3	69.9	54.6	**50.0**	60.6	49.3	55.5	55.7
MVIPER (ours)	**48.0**	**46.0**	**46.0**	**57.1**	**48.6**	**61.3**	**47.7**	**46.8**	**54.1**	**67.1**	**48.9**	50.1	**59.1**	**47.8**	56.1	**52.3**

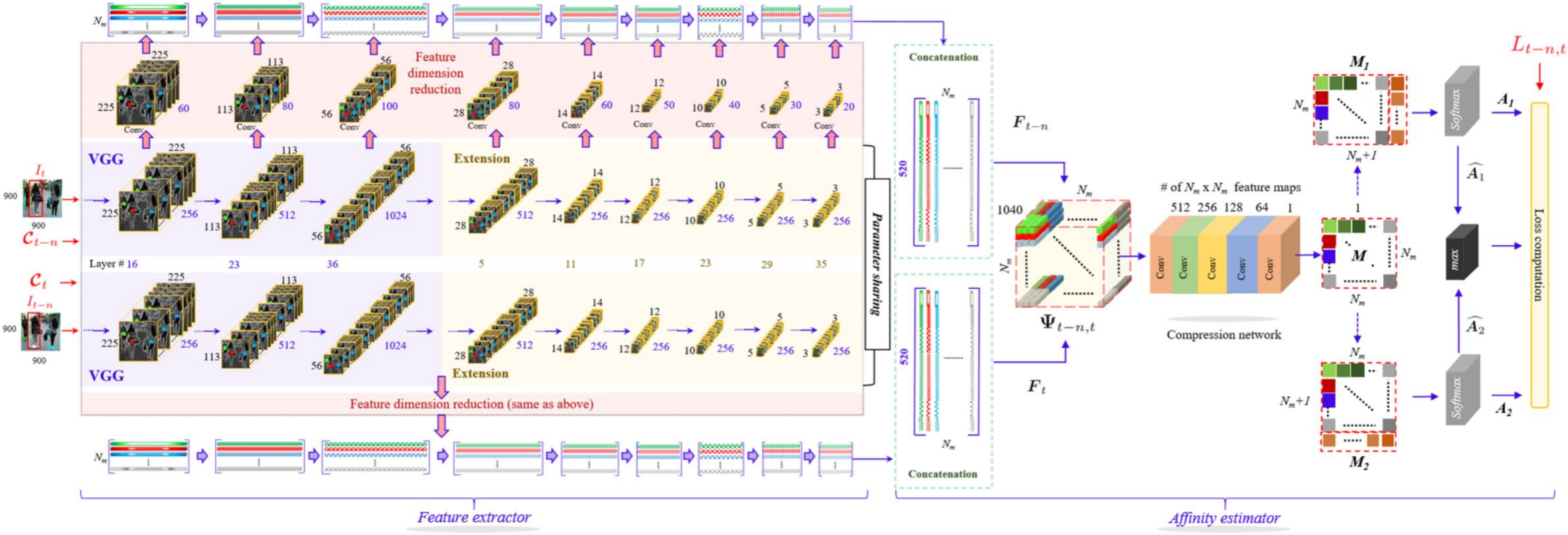
Protocol-1: uses samples from all four provided viewpoints for testing,
Protocol-2: uses only the frontal viewpoint samples.

Qualitative Comparison



Deep Affinity Network

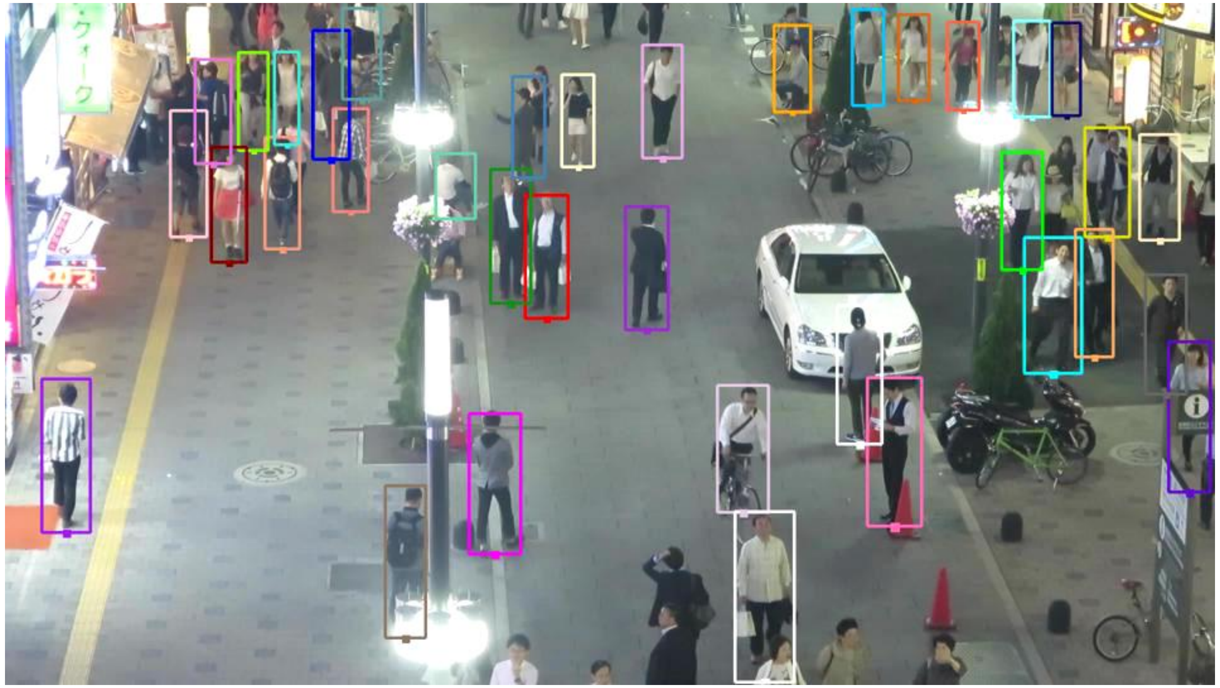
Real time computation of affinity matrix for multiple object tracking in videos

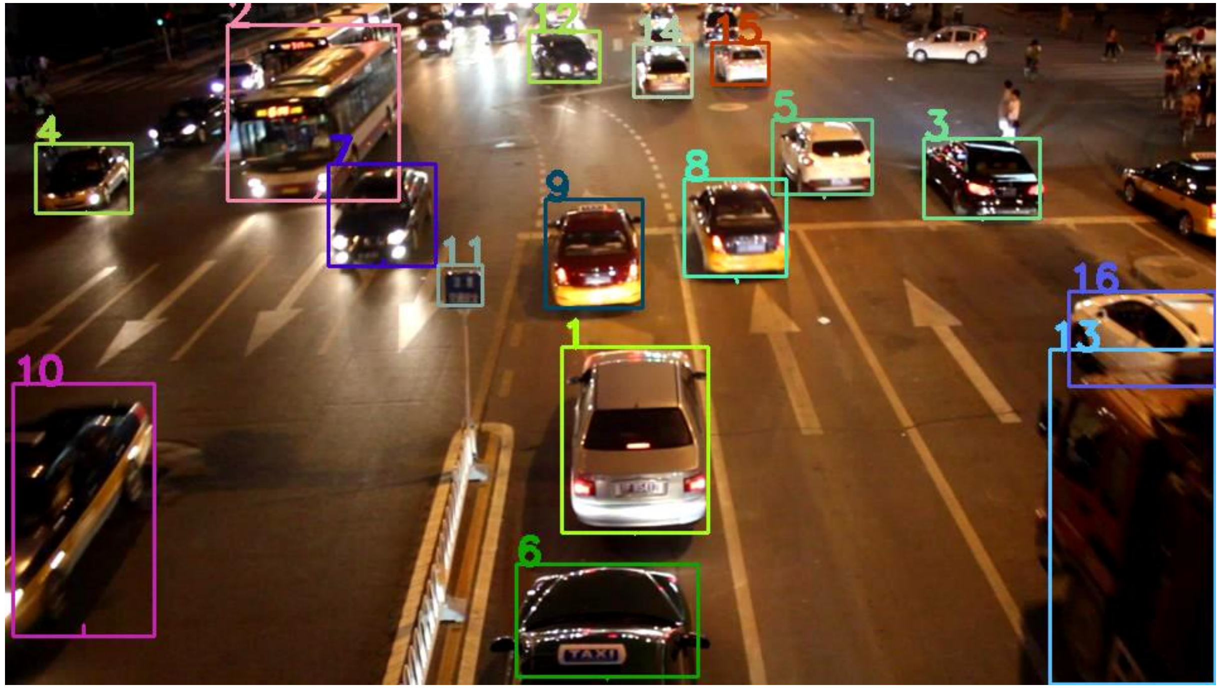


Deep Affinity Network for Multiple Object Tracking

ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, Mubarak Shah, IEEE TPAMI 2019 [arXiv:1810.11780](https://arxiv.org/abs/1810.11780)

MOT: Sample Video Results





Deep Affinity Network is available on GitHub (also called Single Shot Tracker SST)
https://github.com/shijieS/SST

Conclusions and Future Work

- Human motion should be studied differently from content based video retrieval
- Synthetic data is useful
  - When real annotated data is not available
  - Even when real data is available – improves performance of models trained on real data
- Simple techniques from machine learning and computer vision can have a great impact on human performance analysis
- Future/Current work
  - Predict ground reaction forces and moments from monocular video
  - investigate the quality of walking gait in response to Total and Unicompartmental knee arthroplasty surgery

Main Contributors





Datasets and code : http://staffhome.ecm.uwa.edu.au/~00053650/

Sample video data and code to generate more https://github.com/liujianee/MVIPER

Deep Affinity Network (also called Single Shot Tracker SST) https://github.com/shijieS/SST