Lab CudaVision
Learning Vision Systems on Graphics Cards (MA-INF 4308)

CudaLab Project

05.07.2022

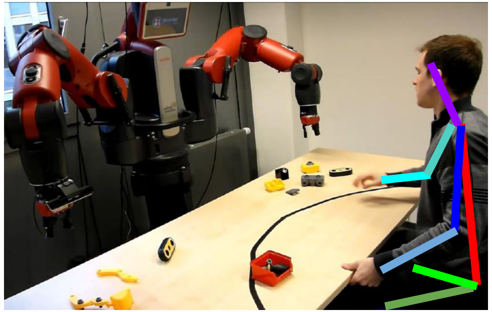
PROF. SVEN BEHNKE, ANGEL VILLAR-CORRALES

Contact: villar@ais.uni-bonn.de

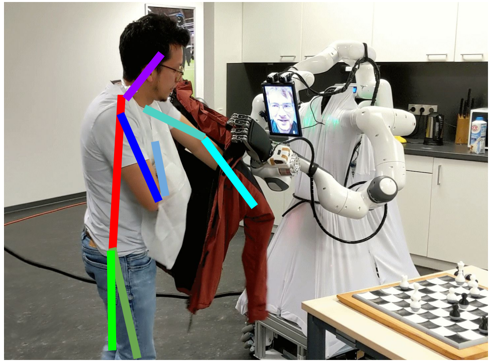
Human Pose Forecasting

Motivation

- Human-robot collaboration is a challenging task
  - Perception of the environment
  - Planning capabilities
  - ➢ **Predicting actions and behavior of nearby agents**



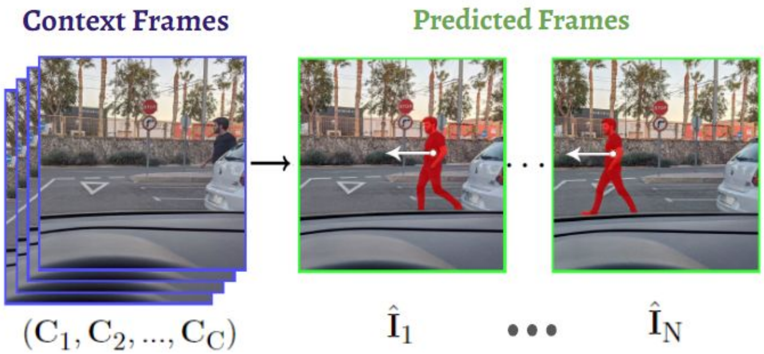
- Human pose is a good representation for:
  - Action recognition and prediction
  - Motion estimation
  - Planning and navigation

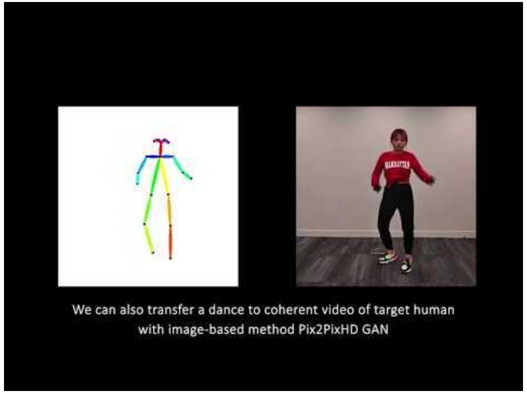


Human Pose Forecasting

- Given a sequence of C seed poses, generate next N plausible poses
  - Predictions must be temporally consistent
  - Incorporate human motion dynamics

- Multiple applications: sports , anticipating human behavior, motion transfer, ...

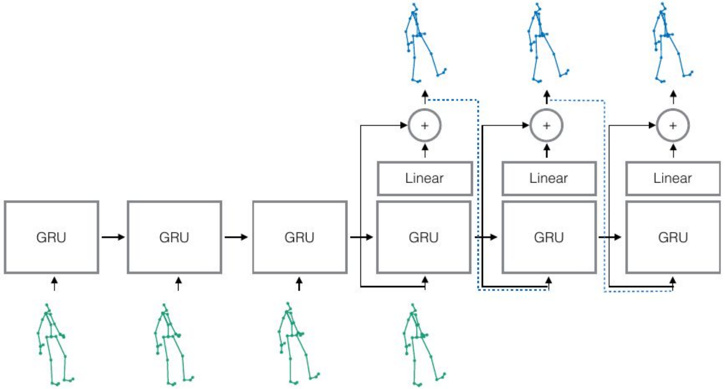




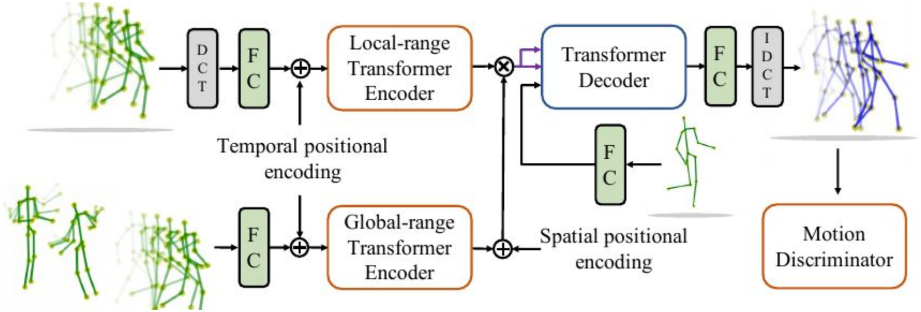
Related Work

- Human motion prediction is an ongoing research topic (2015 - …)

Recurrent neural networks



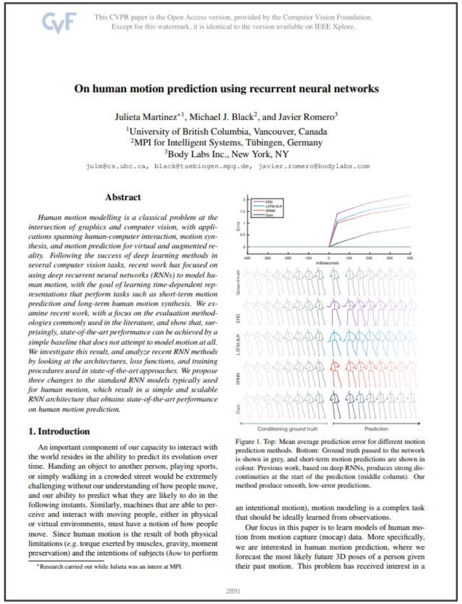
Transformers

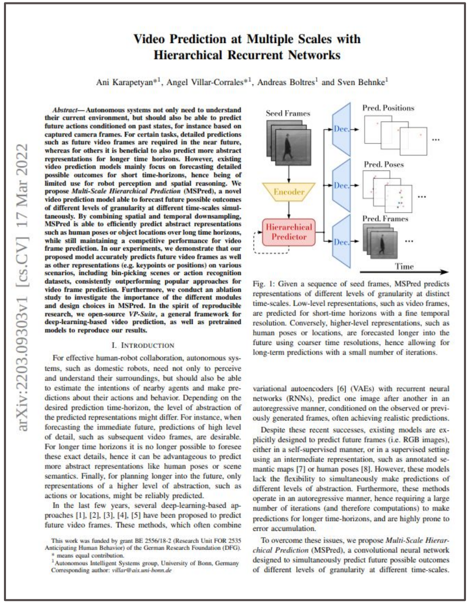


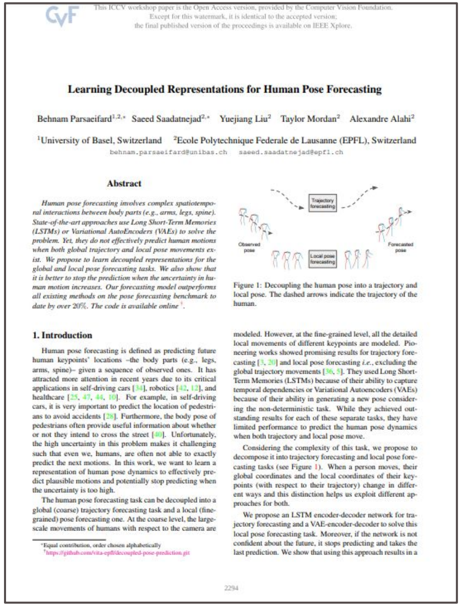
Graph neural networks, 3D-Convolutions, …

Model

Model Inspiration

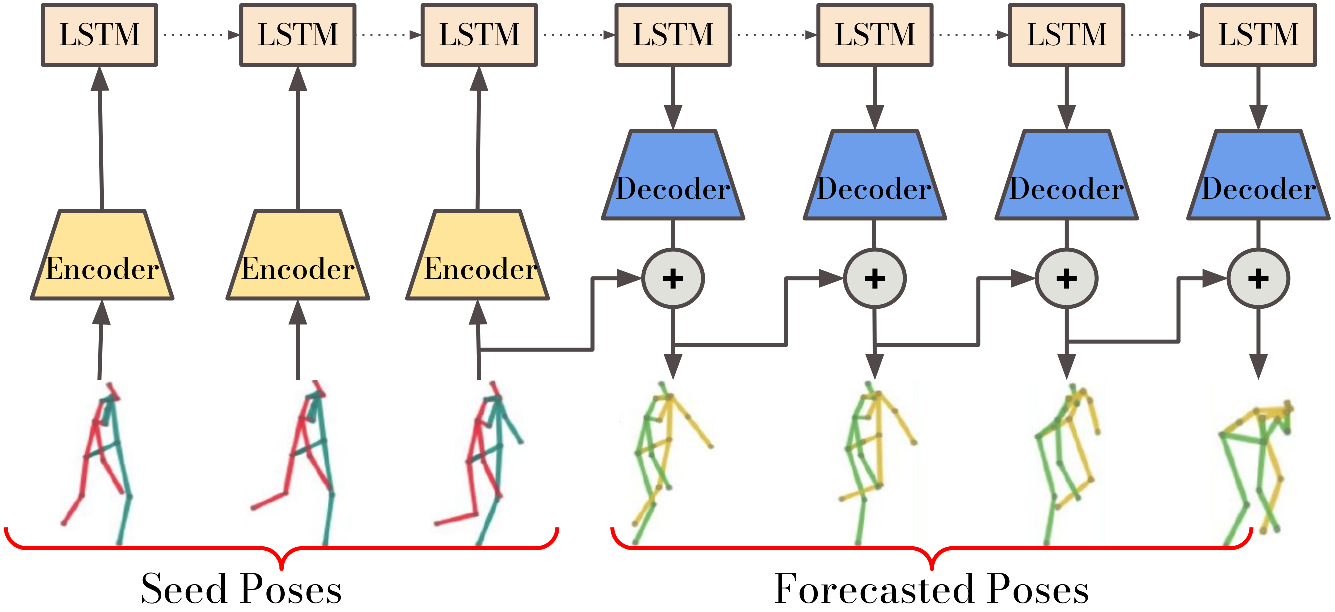






Proposed Model 1

- Skeleton-based pose prediction



Pose Representation

- Pose is parameterized as list of coordinates
  - One dimensional representation
  - Shape is 2N, where N is the number of joints

- Pose can be preprocessed:
  a. Normalize each coordinate by the maximum x & y values respectively
  b. Normalize each coordinate by the image size

$$\left[x1, y1, x2, y2, ..., xN, yN \right]$$

Model

- Encoder:
  - Maps input poses into higher-dimensional representation
  - One single fully connected layer might suffice

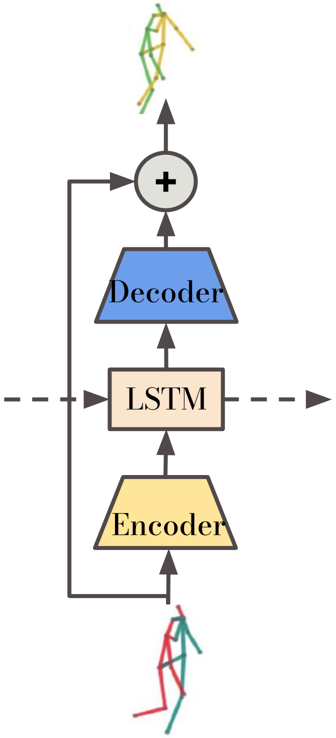
- Recurrent model:
  - Learns motion dynamics
  - One RNN (possibly with multiple cells) or Seq-to-Seq architecture
  - RNN can be either LSTM or GRU

- Decoder:
  - Maps output of predictor back to pose space
  - One single fully connected layer might suffice

Residual Connection

- Poses cannot change much between two consecutive time steps

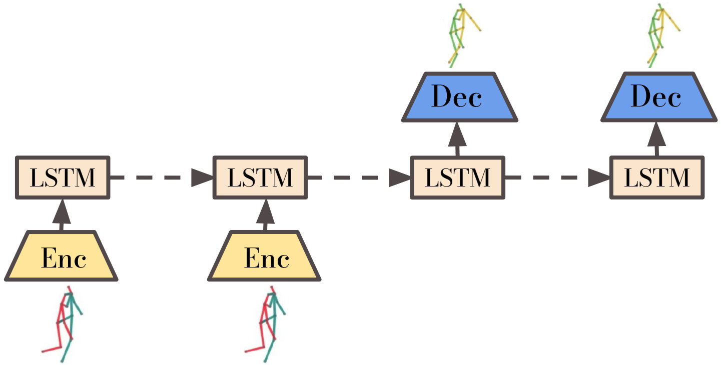
- Recurrent connection between time steps
  - Network must only model changes
  - Faster learning



Model Flow

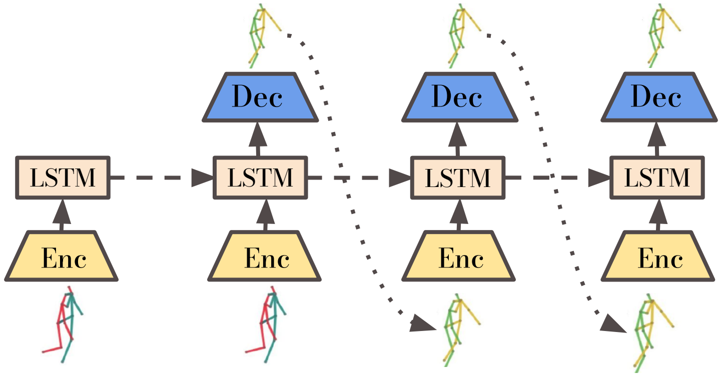
State space model:

- Autoregressive in feature space
- RNN carries pose and motion representations



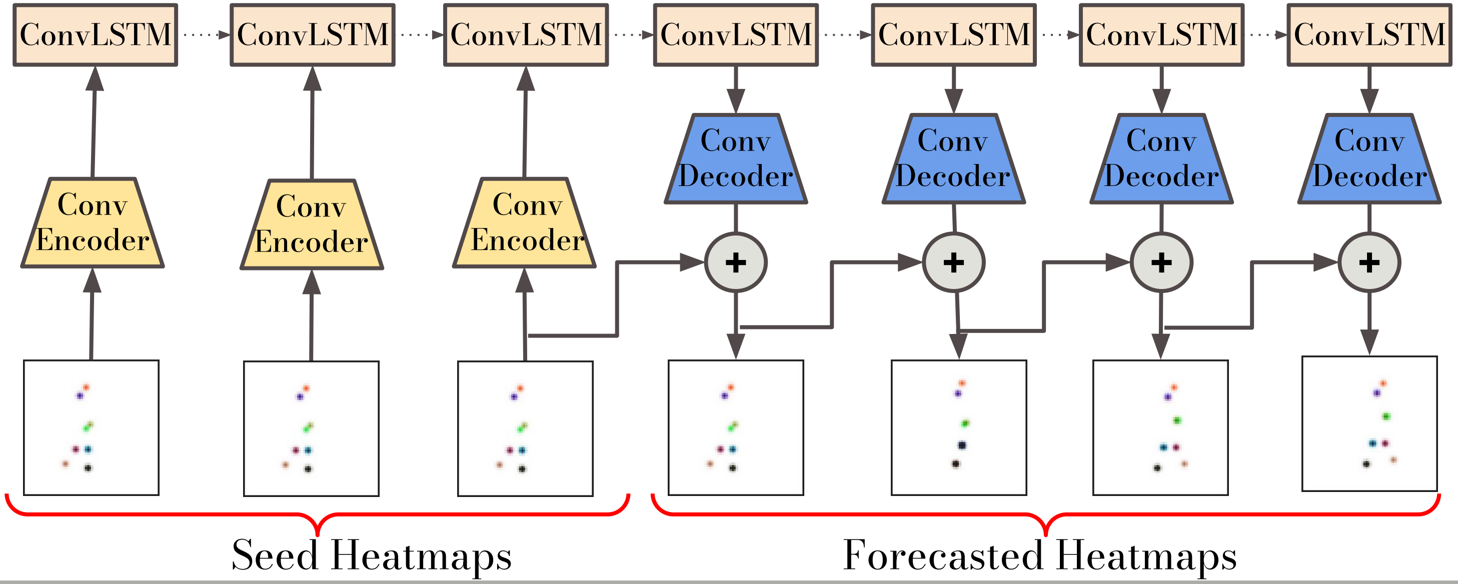
Autoregressive model:

- Predictions are re-encoded and used as inputs
- RNN must only model motion



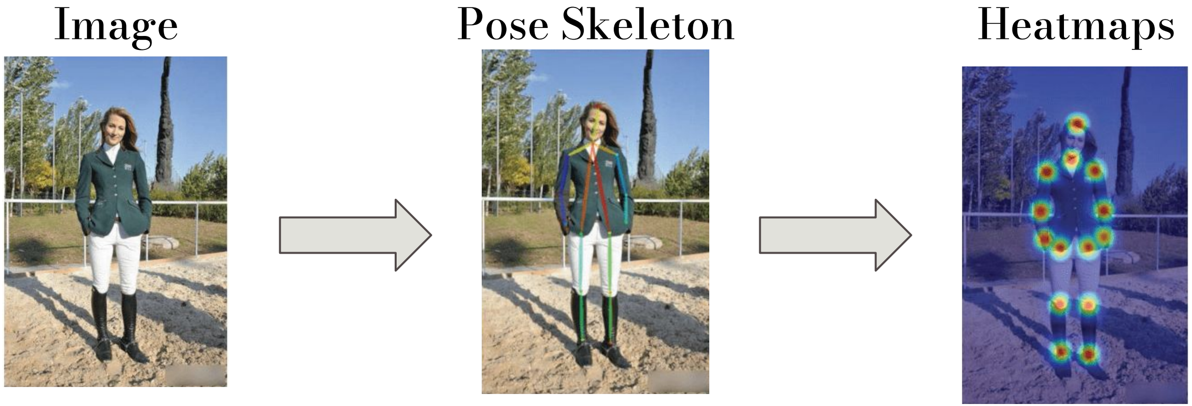
Proposed Model 2

- Heatmap-based pose prediction



Heatmap Representation

- Pose is parameterized as multi-channel heatmaps
  - One channel for each joint
  - Shape is (N, H, W), where N is the number of joints
  - Heatmaps are generated by fitting a Gaussian on the joint center



Model

- Encoder:
  - Maps input heatmaps into better representation for prediction
  - Convolutional encoder, e.g., VGG-ish or ResNet-like
- Recurrent model:
  - Learns motion dynamics
  - One ConvRNN (possibly with multiple cells) or Seq-to-Seq architecture
  - ConvRNN can be either ConvLSTM or ConvGRU
- Decoder:
  - Maps output of predictor back to pose space
  - Convolutional encoder, e.g., VGG-ish or ResNet-like
  - Mirrored version of encoder
- Model flow and residual connections are the same as for model 1

Datasets

Human 3.6M Dataset

- Dataset used as a benchmark for many tasks
  - Pose estimation and forecasting
  - Video prediction
- Stats and characteristics
  - 3.6 million 3D human poses and images
  - 11 professional actors (6 male, 5 female)
  - 17 scenarios (discussion, smoking, ...)
- Data is available in:
  - beast2:/home/cache/H36
  - Images + Annotations
  - Jupyter notebook with an example

Training & Evaluation

Training and Prediction

- Datasets:
  - Train & evaluate on Humans3.6M, both with poses and heatmaps
  - Use the official train-test splits
  - Use a downsampling factor of 8
    - [f1, f2, f3, …,f7, f8, f9, …, f65] → [f1, f9, f17, f25, f33, …, f65]

- Train and evaluate with sequence of size:
  - For pose vectors: (Batch_size, 20, 2 * N_kpts)
  - For heatmaps: (Batch_size, 20, N_kpts, 64, 64)
- Use 10 frames as seed frames, and predict the next 10 frames.

Training and Prediction

- Model:
  - Train & evaluate on pose-based and heatmap-based models
  - Choose your design choices: model flow, residual connections, …
  - Teacher forcing vs. no teacher forcing

- Criterion:
  - Use MSE or MAE as loss functions
  - (Optional) Add an addition perceptual loss (SSIM, LPIPS, …) or adversarial loss

Evaluation

- Measure performance only on the 10 predicted frames
- Evaluate using the following metrics:
  - MSE
  - MAE
  - PDJ
  - PCK
  - MPJPE
- Qualitative evaluation by observing predicted frames

Project Goals and Deliverables

Passing Requirements

1. Implement both models, pipelines and utils
2. Train your models to achieve best possible results on Humans3.6M
   - You must implement and train the described models
   - Make changes and train further models to achieve better results
3. Create overview notebook
4. Write project report

Deliverables

- Complete codebase
    - Clean and structured
    - Not just a notebook!
- Trained model checkpoint and (tensorboard, WandB, …) logs
- Overview notebook (.ipynb & .html) showing main functionalities:
    - Load data
    - Load pretrained model
    - Display some results
- Project report

Grading

- Results and Experiments **55%-60%:**
  - Performing several experiments and obtaining good results
  - Additional experiments: ablation study, changes in the model, …
- Codebase & Overview Notebook **20%:**
  - Implement all functionalities
  - Modularity and structure
- Report **20%-25%**

Project Report

- Document your work in the project report
- Try to be brief, but readable and informative
- Include figures and tables
- Use *BibTex* for the references
- I expect 6-12 pages, but highly depends on number and size of imgs/tables
- Use the following template
    - https://www.overleaf.com/read/tmnvhrsdmjrp

Additional Experiment Ideas

- Try your own ideas!
- Investigate data preprocessing
- Tweak the model
  - Change modules (num. layers, num. kernels, ...)
  - Investigate different predictors (ConvLSTM, Seq-to-Seq)
- Investigate different training strategies or transfer learning:
  - Use additional loss functions
  - Use adversarial supervision
- Make changes to the model
  - Stochastic model: https://arxiv.org/abs/1802.07687
  - Transformer based: https://arxiv.org/abs/2111.12073

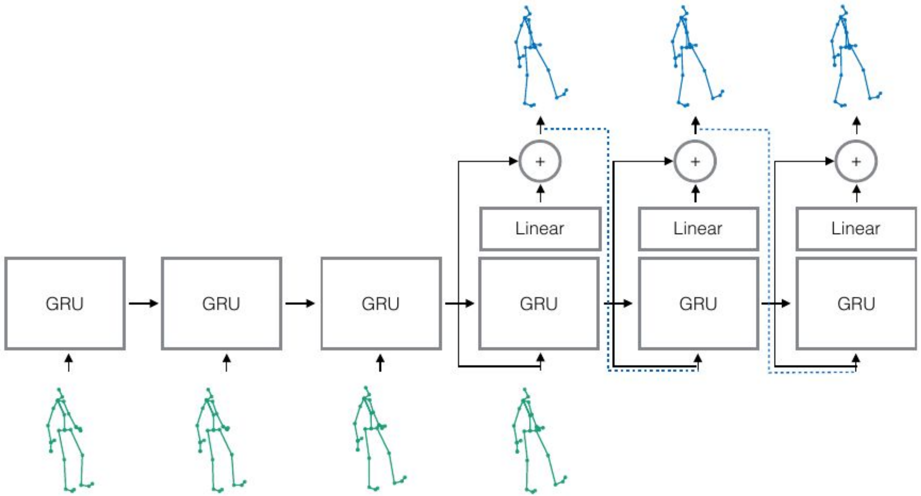
Important Dates

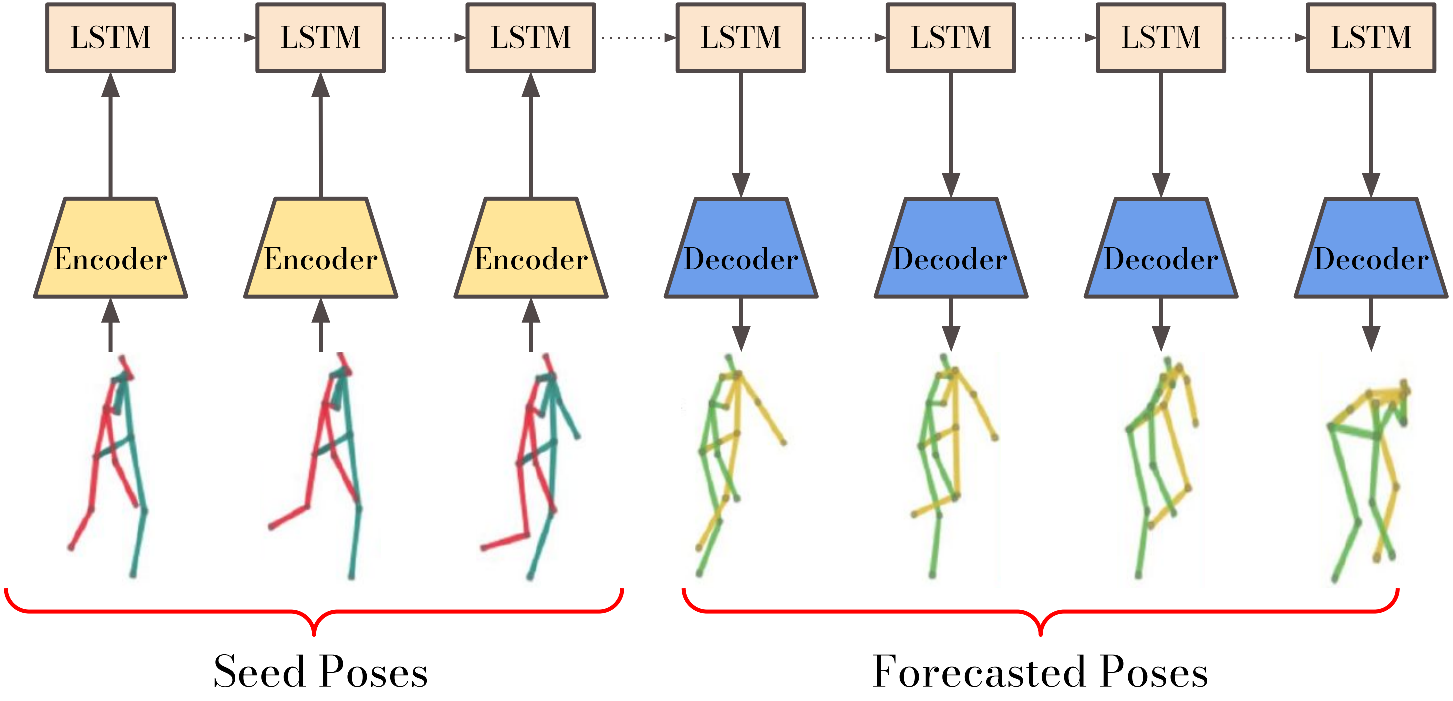
- **05.07**: Starting date
- **29.08-09.09**: Revision session
- **15.09**: Draft submission due
- **30.09**: Final submission:

Questions?

References

1. Parsaeifard, Behnam, et al. "Learning decoupled representations for human pose forecasting." IEEE/CVF International Conference on Computer Vision (ICCV). 2021.
2. Tanke, Julian, Chintan Zaveri, and Juergen Gall. "Intention-based Long-Term Human Motion Anticipation." 2021 International Conference on 3D Vision (3DV). IEEE, 2021.
3. Karapetyan, Villar-Corrales et al. "Video Prediction at Multiple Scales with Hierarchical Recurrent Networks." arXiv preprint arXiv:2203.09303 (2022).
4. Martinez, Julieta, Michael J. Black, and Javier Romero. "On human motion prediction using recurrent neural networks." IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
5. Catalin Ionescu, 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 (PAMI). 2014


GRU
GRU
GRU
GRU
GRU
GRU
Linear
Linear
Linear
+
+
+


LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
Encoder
Encoder
Encoder
Decoder
Decoder
Decoder
Decoder
Seed Poses
Forecasted Poses