SIGGRAPH 2021

ANINDITA GHOSH[1,3*], NOSHABA CHEEMA[1,2,3], CENNET OGUZ[1],
CHRISTIAN THEOBALT[2,3], PHILIPP SLUSALLEK[1,3]

TEXT-BASED MOTION SYNTHESIS WITH A HIERARCHICAL TWO-STREAM RNN

[1]DFKI, [2]MAX-PLANCK INSTITUTE OF INFORMATICS, [3]SAARLAND INFORMATICS CAMPUS

*anindita.ghosh@dfki.de

PROBLEM

Mapping natural language text descriptions to 3D pose sequences for human motions, where the input texts may describe single actions with sequential information e.g., "a person walks four steps forward" or multiple superimposed actions e.g., "a person walks forward for 2 steps, while spinning their arms".

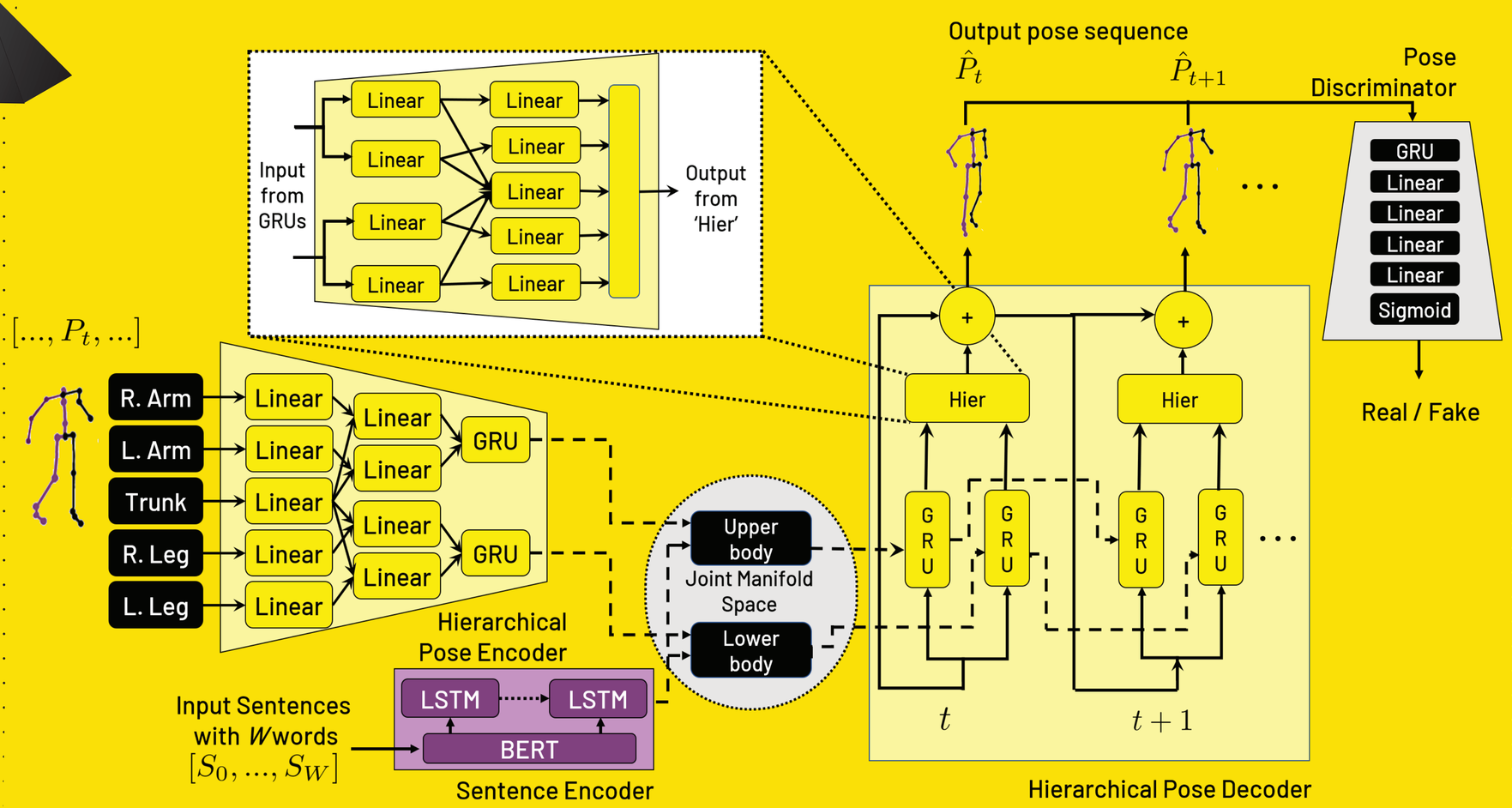
RELATED WORK

Existing text-to-motion mapping methods can either generate motions that stay in one place [1], or generates simple actions on global trajectories , e.g., walking [2,3]. However, these methods fail to translate long-range dependencies and correlations in complex sentences and do not generalize well to complex actions involving synchronized limb movements, e.g. dancing.

In contrast, we propose an RNN based hierarchical two-stream model to explore a finer joint-level mapping between language and 3D pose sequences. Our model can generate animated 3D pose sequences depicting multiple sequential or superimposed actions provided in long, compositional sentences.

OUR APPROACH

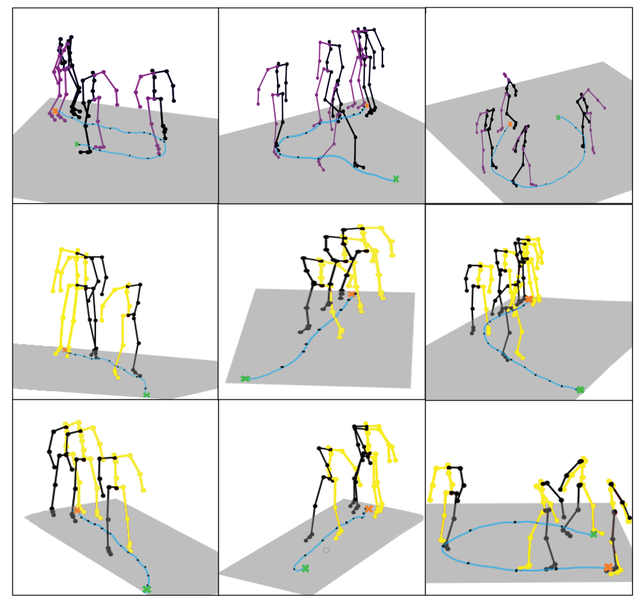
We introduce a hierarchical joint embedding space to learn embeddings of pose and language simultaneously. We separate our intermediate pose embeddings hierarchically to limb embeddings such that our model learns features from the different components of the body. We have a two-stream sequential network to separately learn the upper and the lower body movements and focus on the end joints of the body. We introduce contextualized BERT embeddings [4] with handpicked word feature embeddings to improve text understanding. Lastly, we add a pose discriminator with an adversarial loss to further improve the plausibility of the synthesized motions.



RESULTS

▼ Our method (white) shows more than 50% improvement on both the mean Average Positional Error (APE) and the Average Variance Error (AVE) of joint positions over the state-of-the-art methods of JL2P [1] (purple) and Lin et al. [2] (black).

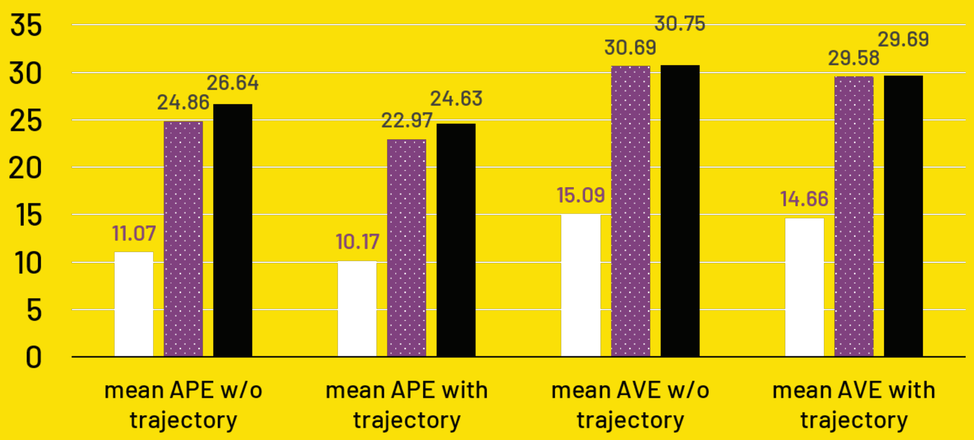
▶ Our method accurately synthesizes a trajectory that matches the semantics of a given sentence compared to the benchmark methods as shown in the Figures.



"A human walks forward two steps, pivots 180 degrees, and walks two steps back to where they started."

"A person walks two steps forwards, rotates to their left 180 degrees into the opposite direction and continues walking for two steps then stops."

"A human walks in a counterclockwise circle, completing one round in 8 steps"



REFERENCES

1. Plappert, Matthias, et al. "Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks." Robotics and Autonomous Systems 2018.
2. Lin, Angela S., et al. "Generating Animated Videos of Human Activities from Natural Language Descriptions." Visually Grounded Interaction and Language Workshop, NeurIPS 2018.
3. Ahuja, Chaitanya., et al. "Language2pose: Natural language grounded pose forecasting." 2019 International Conference on 3D Vision, IEEE, 2019.
4. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

This research is funded by the BMBF grants XAINES (01|W20005) and IMPRESS (01|S20076), EU Horizon 2020 grant Carousel+ (101017779) and an IMPRS-CS Fellowship. Computational resources provided by the BMWi grants 01MK20004D and 01MD19001B.