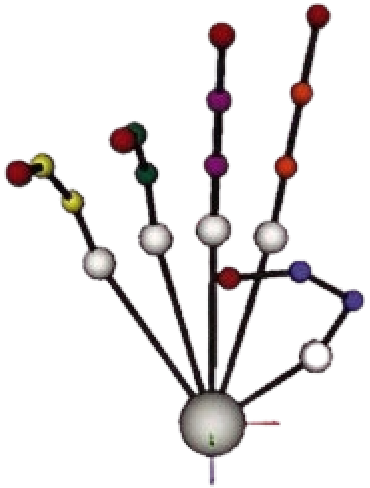
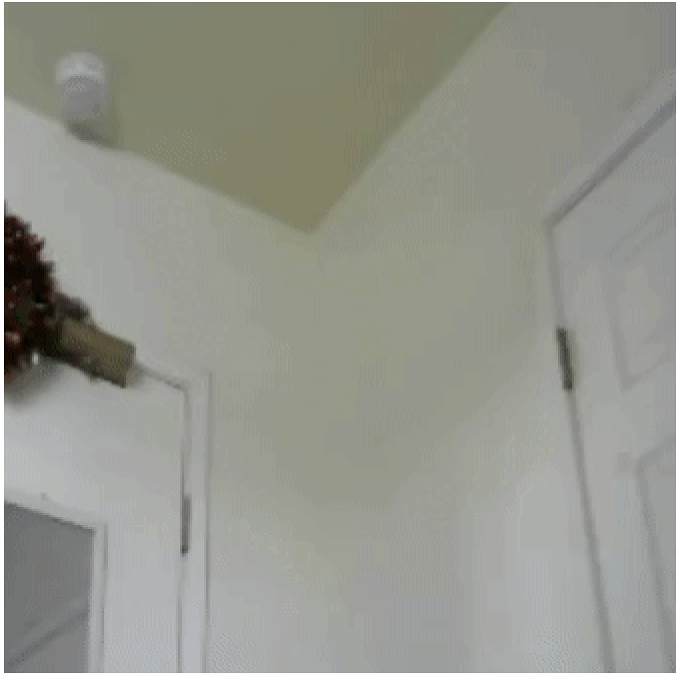
Comparison of Spatio-Temporal Hand Pose Denoising Models

Johnny Núñez Cano

Directors:
Dr. Sergio Escalera
MSc. German Barquero
MSc. Cristina Palmero



Hand Pose Estimation



https://github.com/NVIDIA-AI-IOT/trt_pose_hand

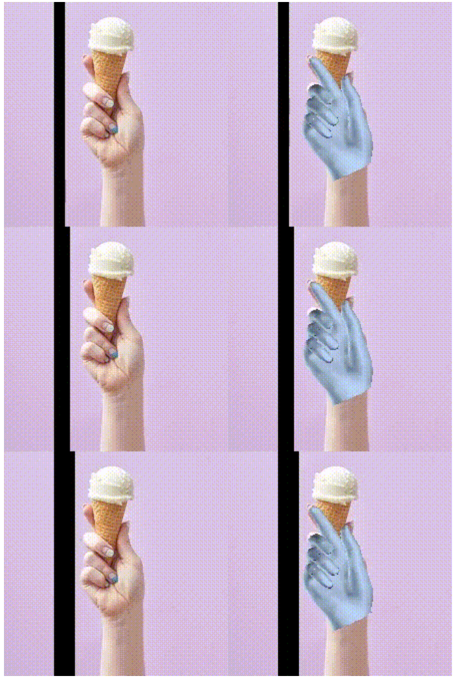
Estimation Algorithms

- FrankMocap

Rong, Yu, et al.. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In IEEE ICCV Workshops, 2021.

Estimation Algorithms

- FrankMocap
- Mesh Graphormer



Rong, Yu, et al.. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In IEEE ICCV Workshops, 2021.
Lin, Kevin and Wang, Lijuan and Liu, Zicheng.. Mesh Graphormer

Typical Errors

01 ...

Jitter

Small displacement.

02 ...

Inversion

Error on the same instance.

03 ...

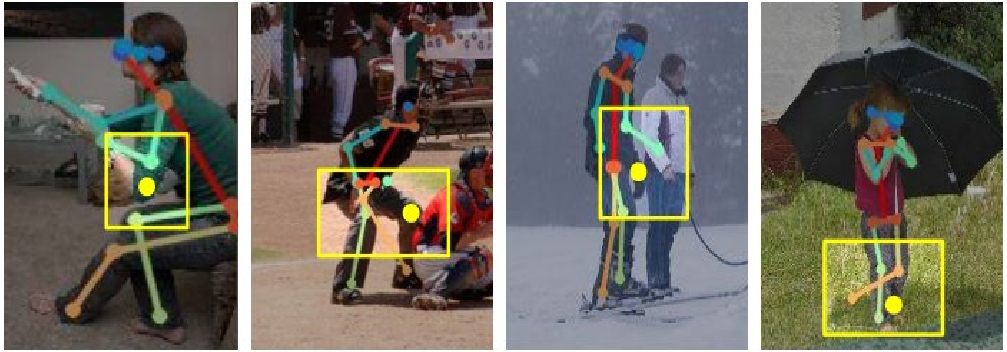
Swap

Error on different instances.

04 ...

Miss

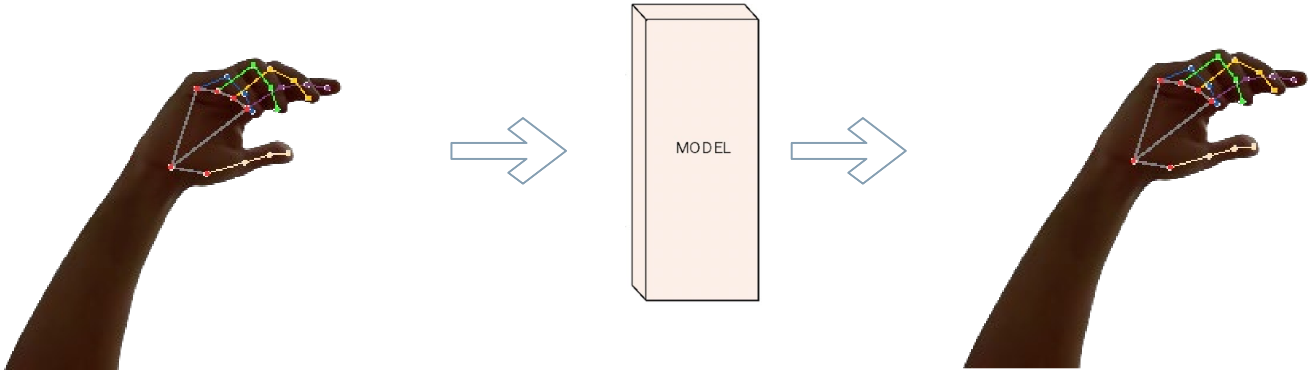
Not exists keypoint.



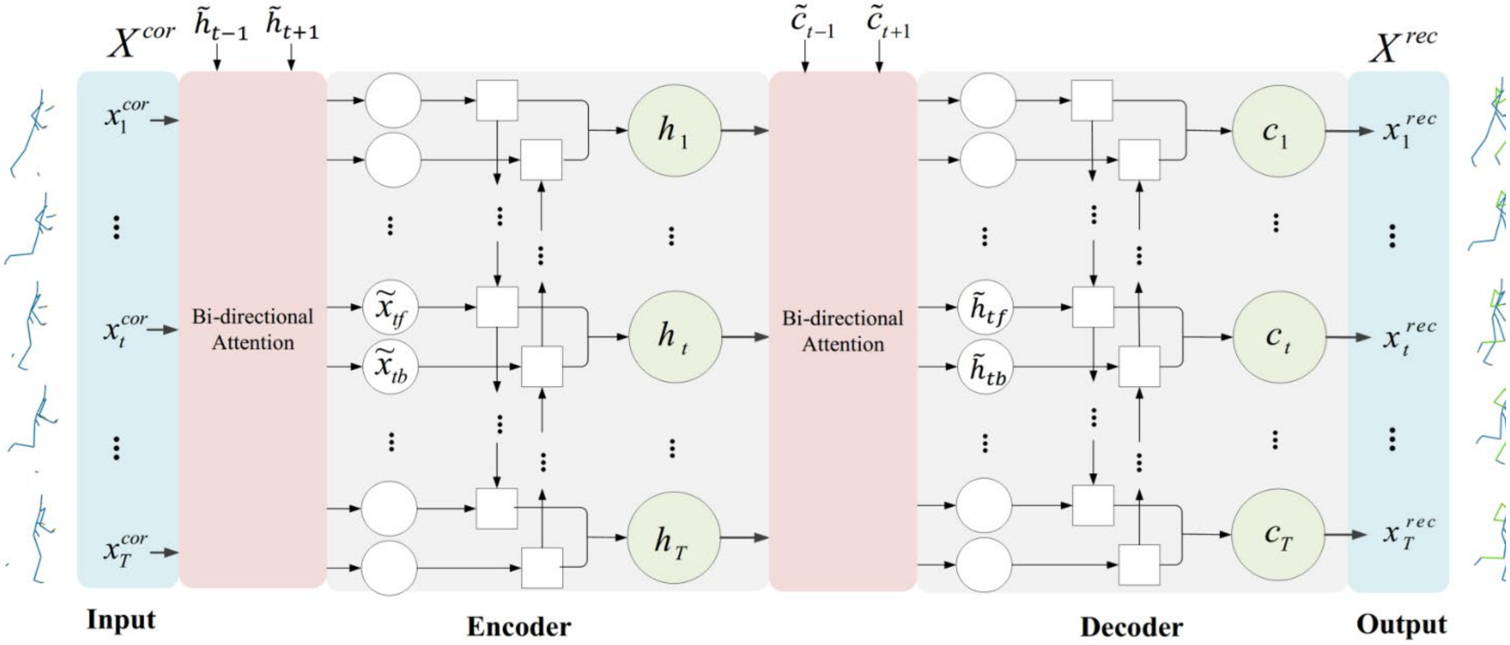
(a) Jitter (b) Inversion (c) Swap (d) Miss

Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee.. PoseFix: Model-Agnostic General Human Pose Refinement Network. CVPR, 2019

Our Goals

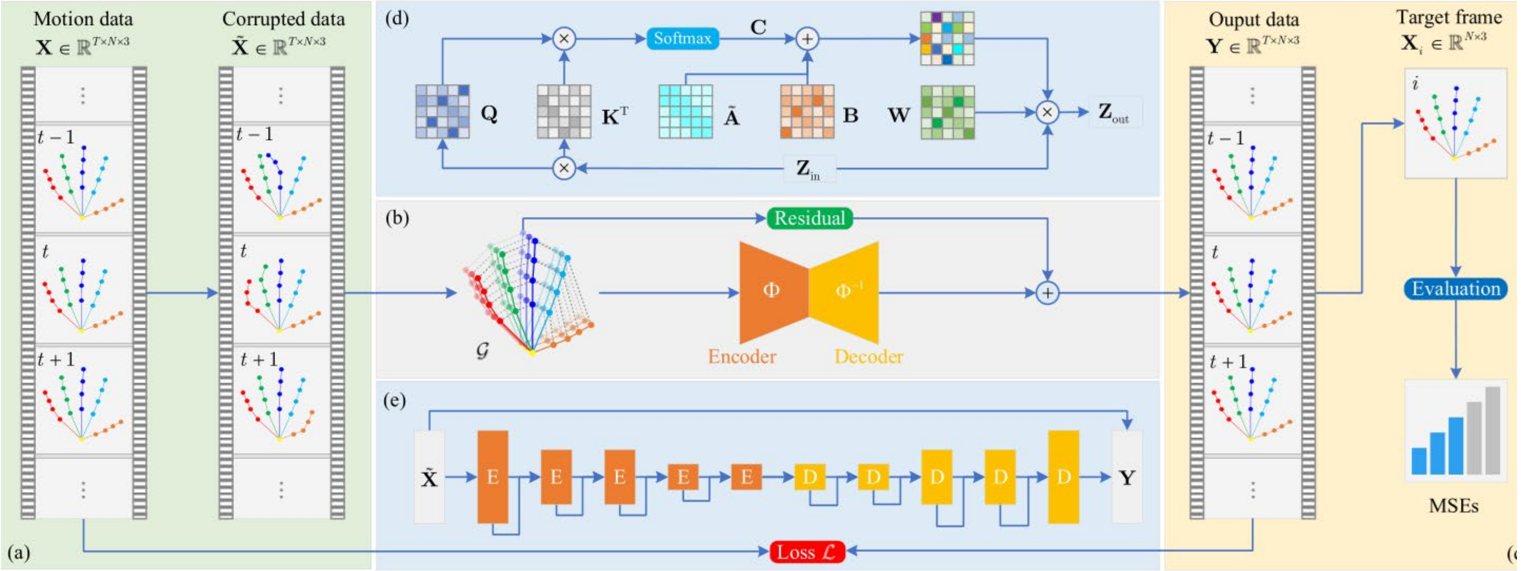


Related Work



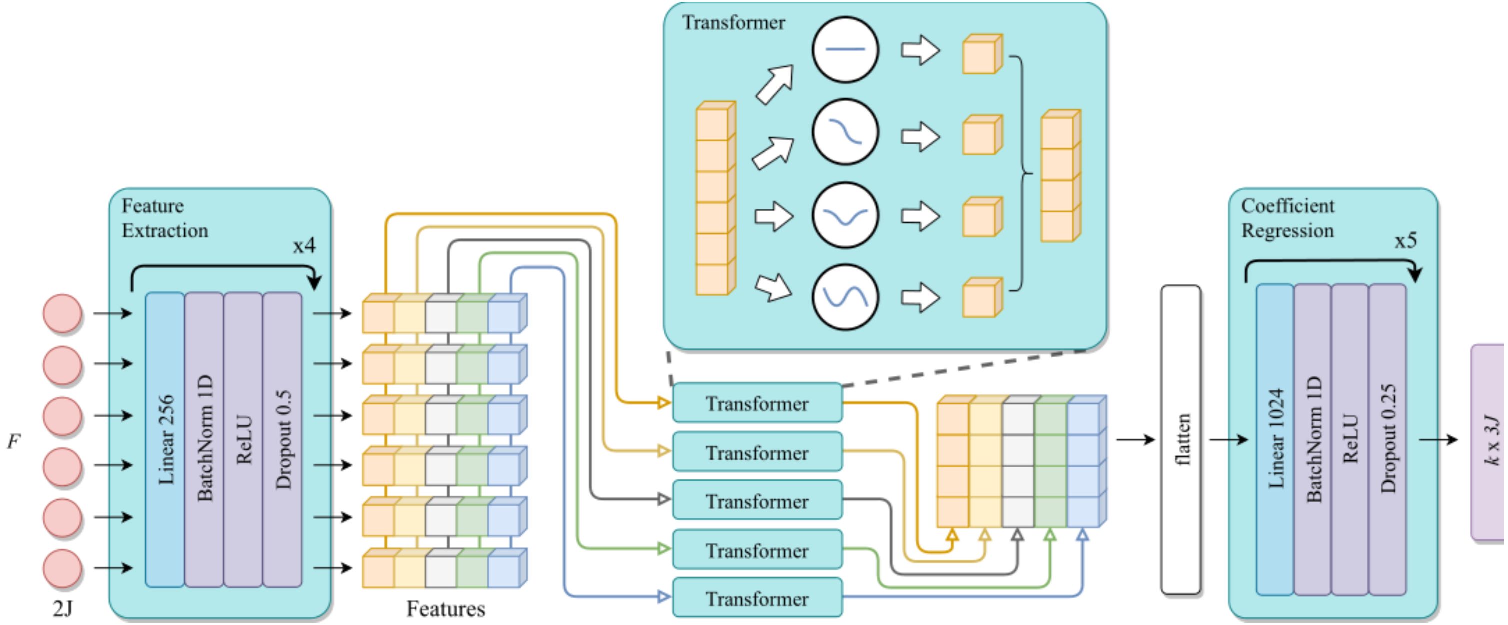
Qiongjie Cui, Huaijiang Sun, Yupeng Li, Yue Kong.. A Deep Bi-directional Attention Network for Human Motion Recovery. IJCAI, 2019

Related Work



Durham Research , Kanglei Zhou , Zhiyuan Cheng , Hubert P H Shum , Frederick W B Li , Liang , Xiaohui Liang. STGAE: Spatial-Temporal Graph Auto-Encoder for Hand Motion Denoising. ISMAR, 2021.

Related Work



Lin, Jiahao and Lee, Gim Hee.. Trajectory Space Factorization for Deep Video-Based 3D Human Pose Estimation. BMVC, 2019

UDIVA

Understanding
Dyadic
Interactions from
Video and
Audio signals

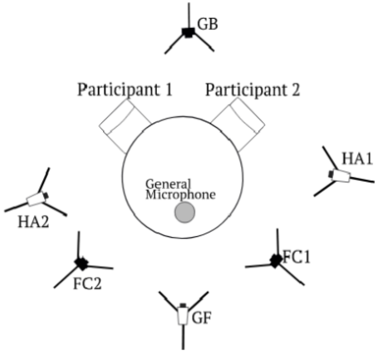
Palmero, Cristina, et al. "Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset." *WACV (Workshops)*. 2021.

UDIVA v0.5

- 4 tasks from 145 sessions: 116 for training (raw annotations), 18 for validation and 11 for testing (cleaned annotations).

Talk *Animals* *Lego* *Ghost*

UDIVA v0.5

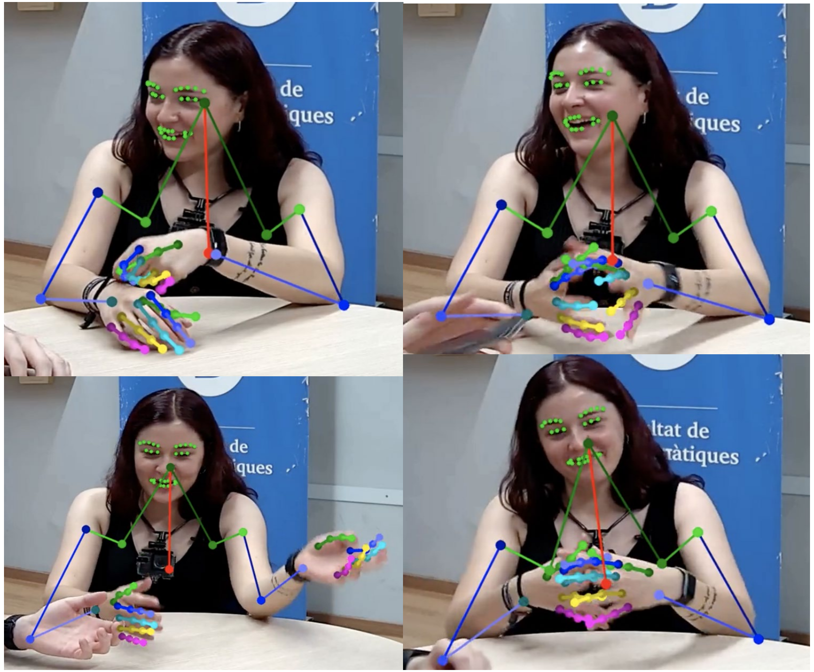

GB
Participant 1
Participant 2
HA1
General Microphone
HA2
FC1
FC2
GF


FC2


1902070
FC1

...

Landmarks extraction

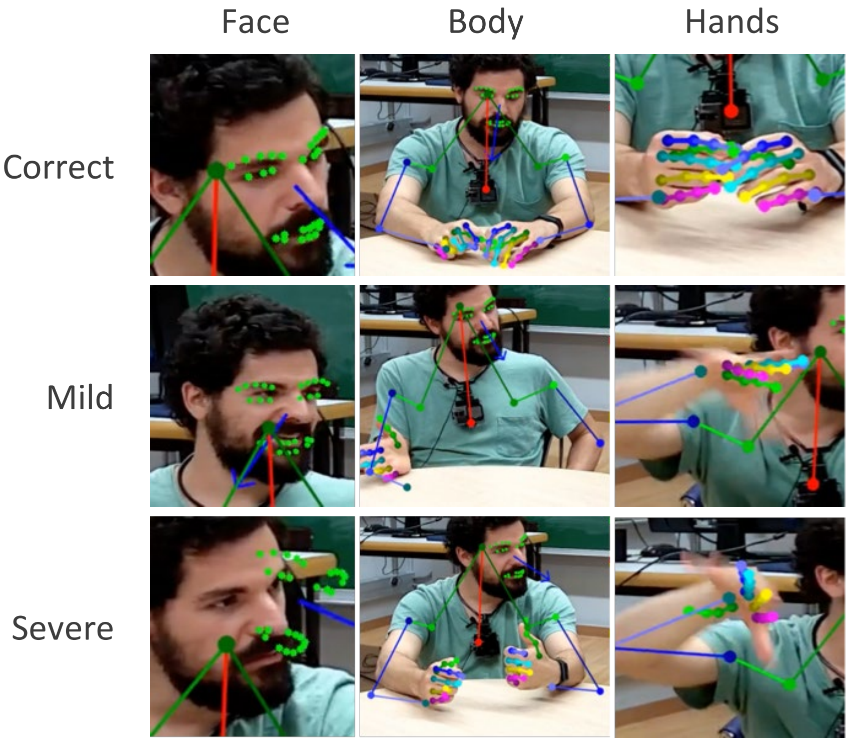


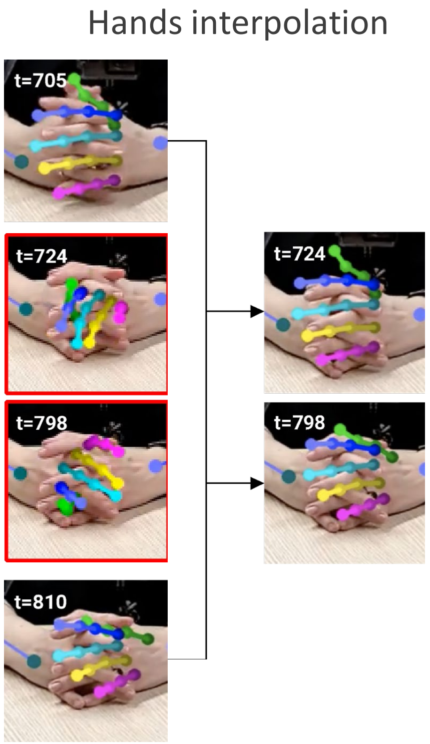
Hand Pose estimation Algorithm:
FrankMocap

Rong, Yu, et al.. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In IEEE ICCV Workshops, 2021.

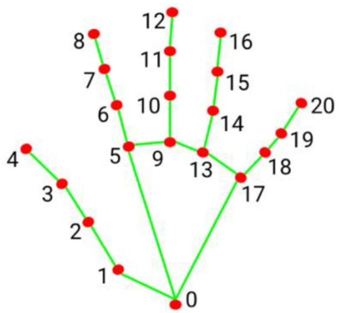
Visual inspection

- Validation (18 sessions) and test (11 sessions) sets underwent a visual inspection process.



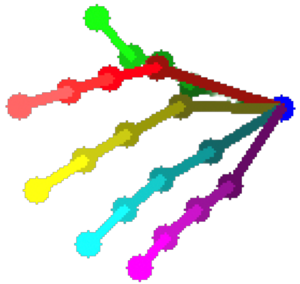


Hand Representation

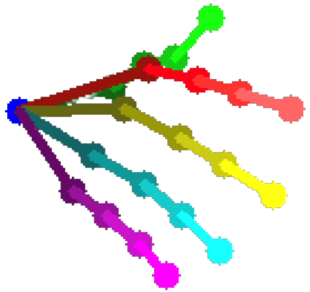


- 1x20 right hand landmarks
- 1x20 left hand landmarks (flip horizontally)
- Absolute coordinates were transformed to root-relative coordinates.

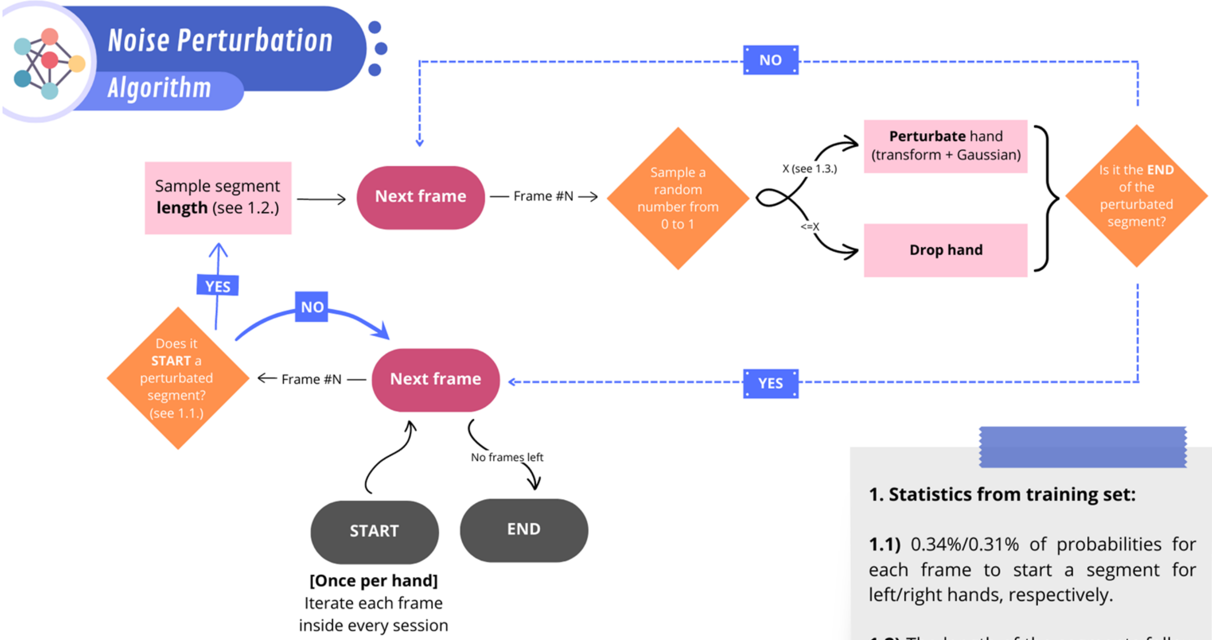
Left Hand



Left Hand Transformed



Perturbation Algorithm



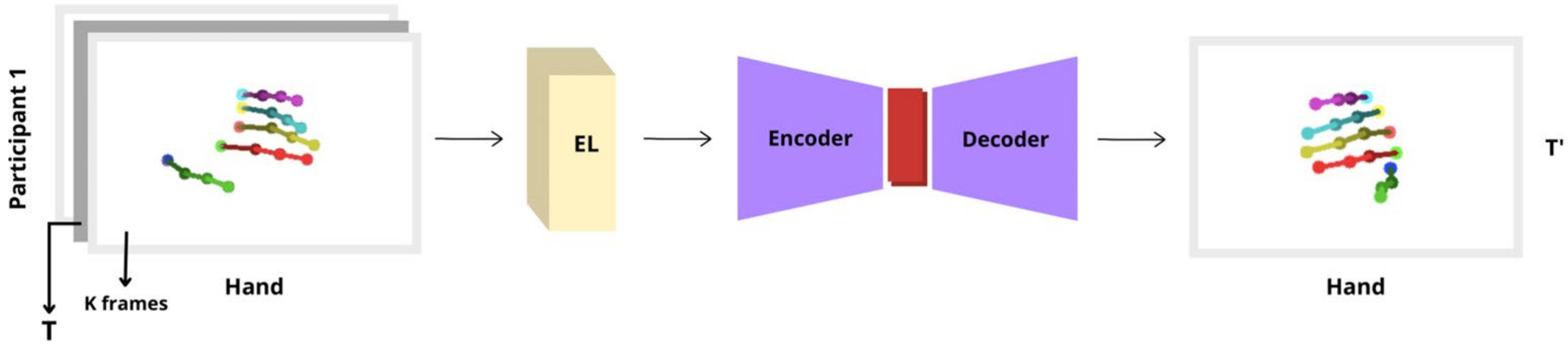
1. Statistics from training set:

1.1) 0.34%/0.31% of probabilities for each frame to start a segment for left/right hands, respectively.

1.2) The length of the segments follow a 'lognorm' distribution.

1.3) There are 14.4%/8.0% of left/right hands missing inside the segments (joints equal to 0).

Methodologies



- **Architectures**:
  - MLP ←→ MLP Frequential
  - Seq2Seq ←→ Seq2Seq-Bidirectional
  - Spatio-Temporal Transformer
- **Data**:
  - Participant 1
  - Two hands consecutively
  - Ground Truth: Cleaned Frames

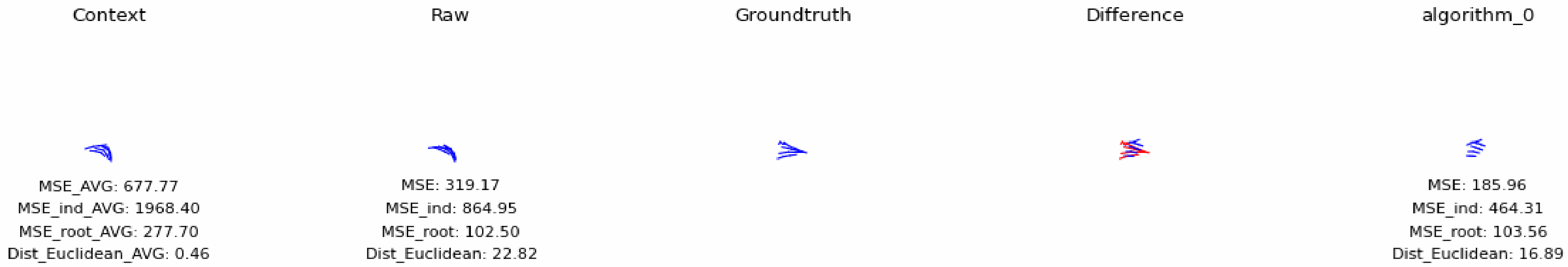
Visual Results

- MLP

Sample 21 - [H] Frame 0

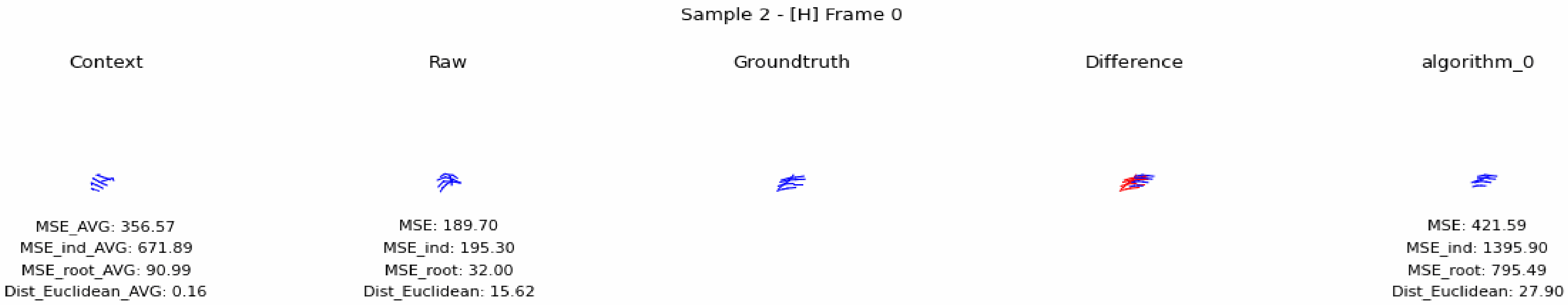
- MLP Frequential

Sample 31 - [H] Frame 0

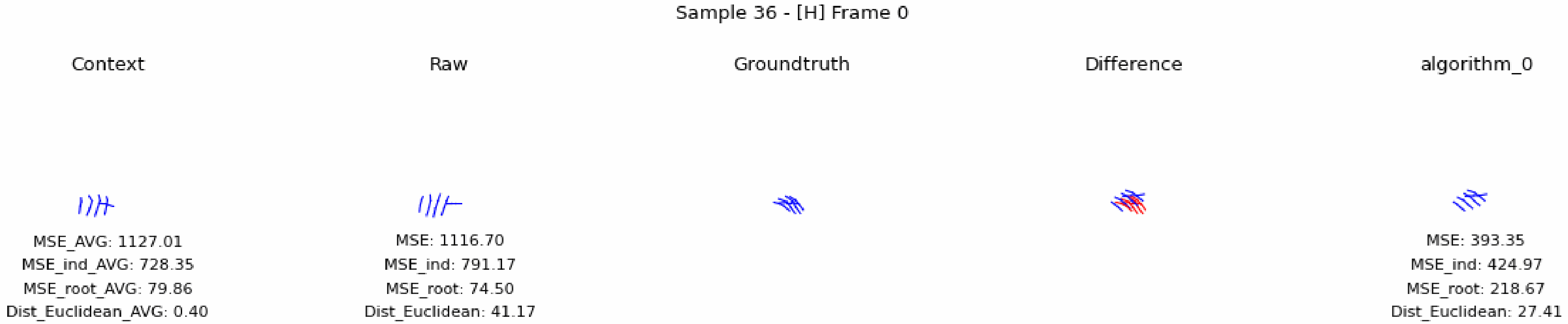


Visual Results

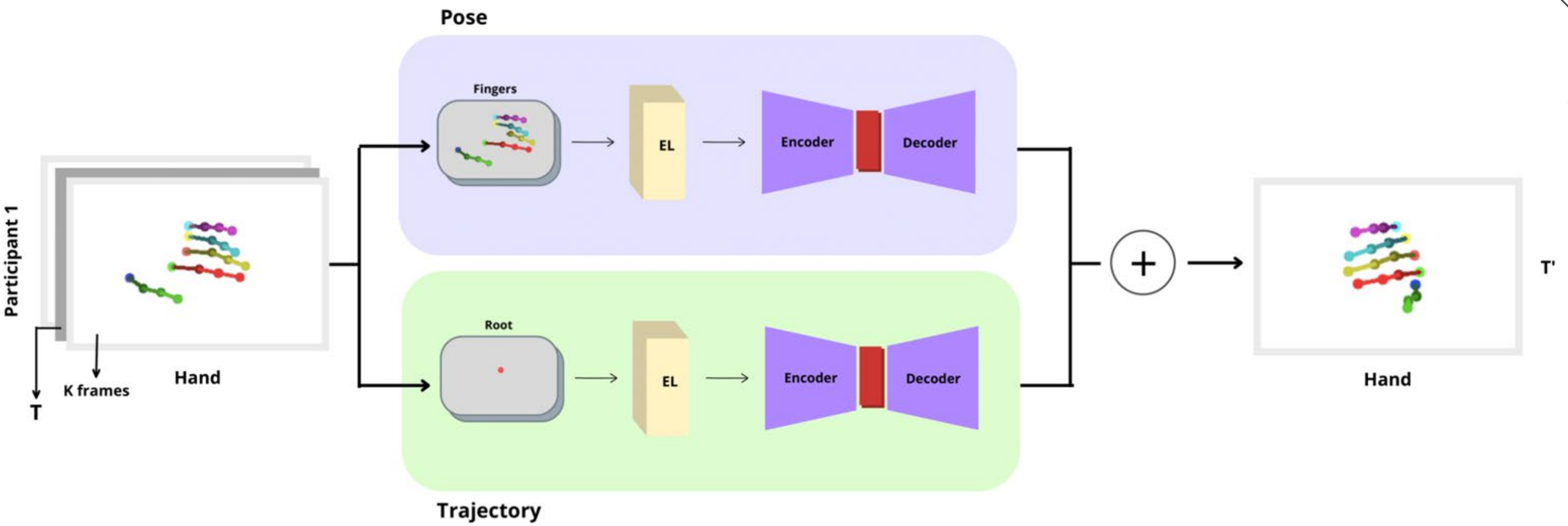
- RNN



- Transformer



Divide and Conquer



- **Pose Architectures**:
  - MLP
  - MLP Frequential
  - Transformer
- **Trajectory Architectures**:
  - Seq2Seq
    - LSTM

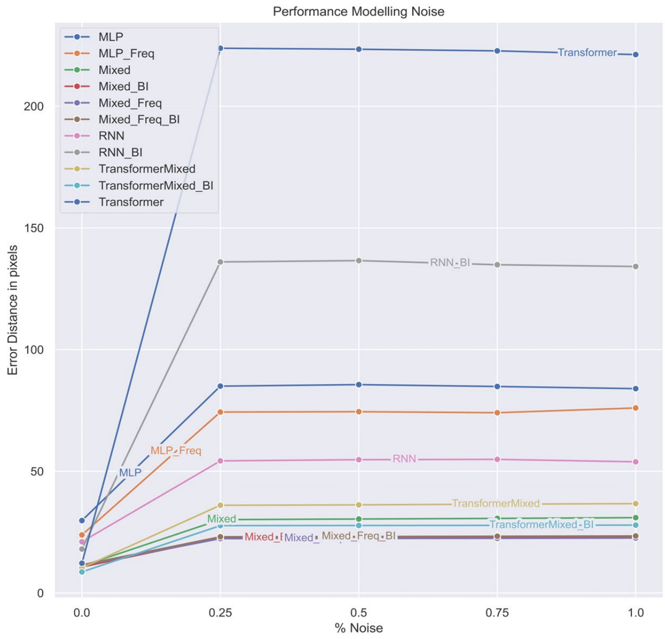
Ranking

- Baseline: 9.138526

Model	Prediction
TransformerMixed_BI_100	**8.649507**
TransformerMixed_100	10.361105
Mixed_BI_100	10.661939
Mixed_100	11.080834
Mixed_Freq_100	11.377095
Mixed_Freq_BI_100	11.522652
Transformer_100	12.230602
RNN_BI_100	18.057422
RNN_100	21.049572
MLP_Freq_100	23.856506
MLP_100	29.756484

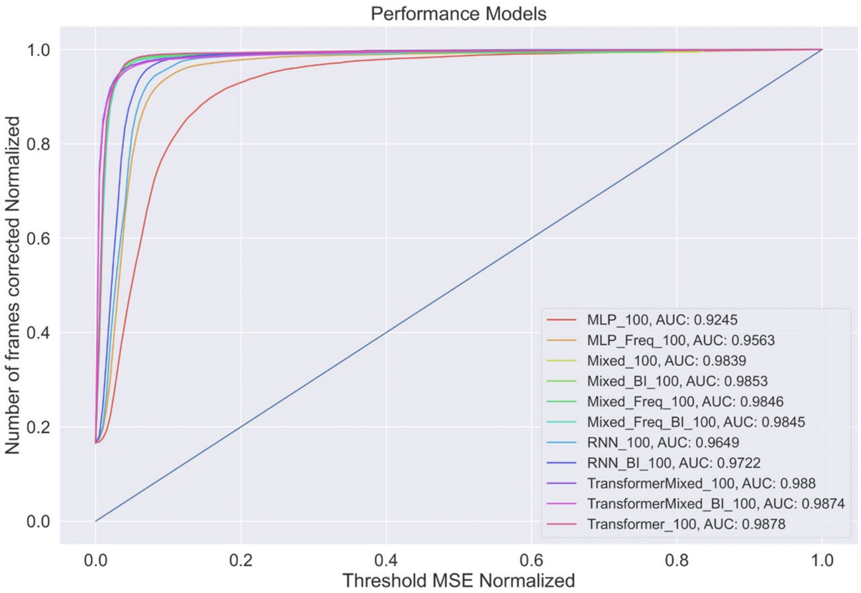
Mean Error of Euclidean Distance: Two hands

More Evaluation



Performance of the 100 observation length models respect to noise application.

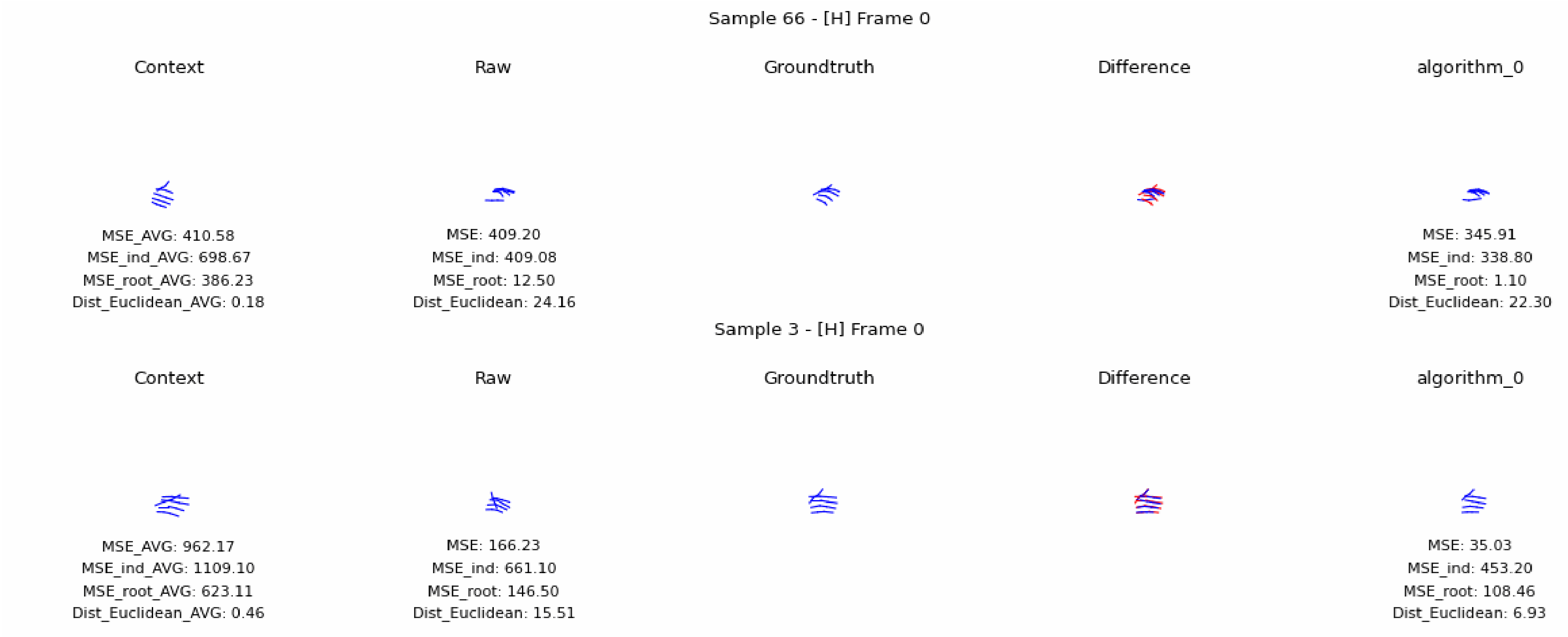
More Evaluation



Counting frames with MSE less than the threshold. Normalized values.

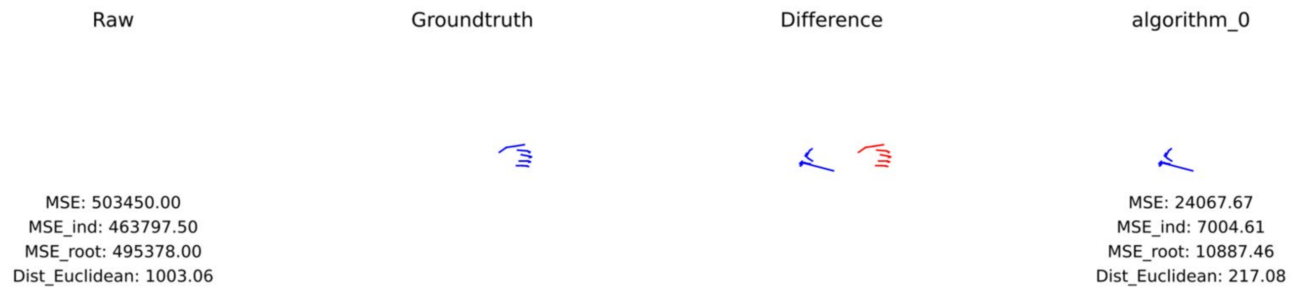
Best Algorithm Visual Result

- Mixed Transformed Bidirectional



First Case

Second Case



Third Case

Conclusions

- We carefully reviewed all related state-of-the-art literature on pose denoising.
- **The innovative Data Perturbation** algorithm allows to worsen the hand poses.
- The frequency space **improves** results in simple models over the original space, but not is representative.
- Mixed model **are more robust** applying more noise.
- **Pose and Trajectory disentanglement** is a good solution approach to find a robust denoising architecture algorithms.

Future Work

- How the observation window affects model performance.
- Not add noisy into cleaned frames.
- New recent Deep Learning Models.
- Use frequency domain only in trajectory.
- Instead of denoising a unique pose, denoising a sequence of poses.

Thanks!

Do you have any questions?