# A generic diffusion-based approach for 3D human pose prediction in the wild

Saeed Saadatnejad, Ali Rasekh, Mohammadreza Mofayezi, Yasamin Medghalchi, Sara Rajabzadeh, Taylor Mordan, Alexandre Alahi

## Overview

**Task**: Predicting a sequence of future 3D poses of a person given a sequence of past observed ones

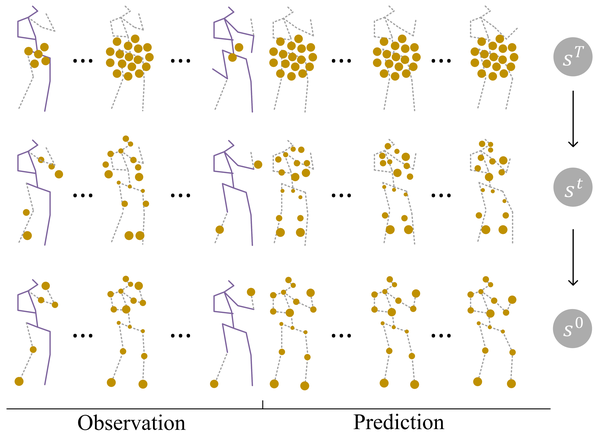
**Challenge**: Predict accurately in noisy observation (partial occlusion, whole frame missing or inaccurate observations)

**Approach**: All missing elements are treated as noise and denoised with our conditional diffusion model by simultaneously:
1) predicting poses for the future frame
2) repairing the noisy observations

## Related work

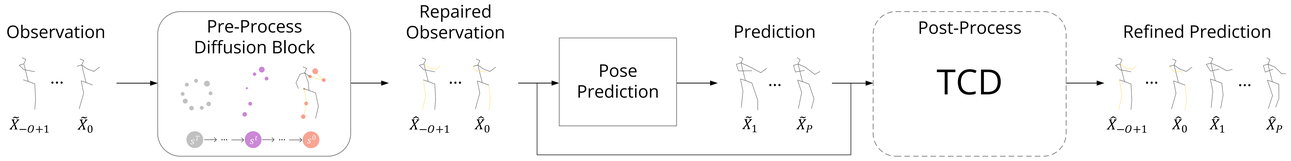
- Deterministic models [1,2]: not diverse
- Stochastic models [3,4,5]: not accurate
- Noisy observations in the real-world [6]
  - not perfect because of not modeling the noise

## Approach



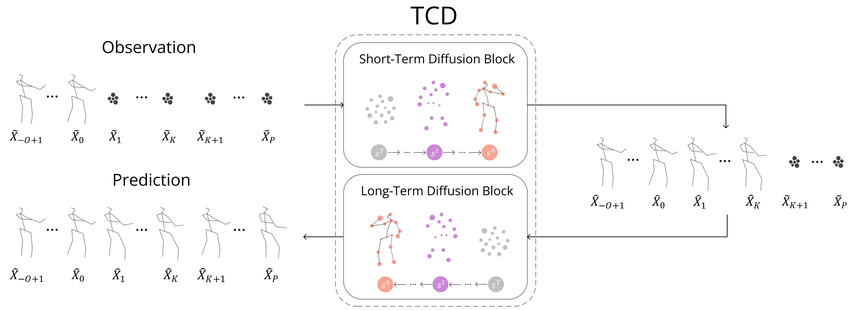

partial occlusion (first column), missing whole frame (second column), or inaccurate observations (third column)

## Generic framework




Improve any existing predictor in a black box manner in two steps:
1) pre-processing to repair the observations and
2) postprocessing to refine the predicted poses.

## Method




## Results

| Model | Human3.6M [26] ADE ↓ | FDE ↓ | MMADE ↓ | MMFDE ↓ | HumanEva-I [53] ADE ↓ | FDE ↓ |
|---|---|---|---|---|---|---|
| Pose-Knows [58] | 461 | 560 | 522 | 569 | 269 | 296 |
| MT-VAE [61] | 457 | 595 | 716 | 883 | 345 | 403 |
| HP-GAN [6] | 858 | 867 | 847 | 858 | 772 | 749 |
| BoM [7] | 448 | 533 | 514 | 544 | 271 | 279 |
| GMVAE [19] | 461 | 555 | 524 | 566 | 305 | 345 |
| DeLiGAN [22] | 483 | 534 | 520 | 545 | 306 | 322 |
| DSF [62] | 493 | 592 | 550 | 599 | 273 | 290 |
| DLow [63] | 425 | 518 | 495 | 531 | 251 | 268 |
| Motron [52] | 375 | 488 | – | – | – | – |
| Multi-Objective [35] | 414 | 516 | – | – | 228 | 236 |
| GSPS [40] | 389 | 496 | 476 | 525 | 233 | 244 |
| STARS [60] | 358 | 445 | 442 | 471 | 217 | 241 |
| TCD (ours) | **356** | **396** | **463** | **445** | **199** | **215** |

| | Model | 80ms | 320ms | 560ms | 720ms | 880ms | 1000ms |
|---|---|---|---|---|---|---|---|
| Noisy Observation | Zero-Vel | 84.9 | 138.2 | 169.9 | 184.2 | 193.7 | 198.2 |
| | HRI [39] | 65.2 | 104.5 | 130.0 | 141.6 | 151.1 | 157.1 |
| | PGBIG [36] | 67.0 | 107.1 | 132.1 | 143.5 | 152.9 | 158.8 |
| | TCD (ours) | **11.2** | **51.3** | **75.4** | **85.4** | **95.4** | **104.5** |
| Repaired Observation | Pre(ours) + Zero-Vel | 24.1 | 76.3 | 107.6 | 121.7 | 131.7 | 136.7 |
| | Pre(ours) + HRI [39] | 11.4 | 48.6 | 78.3 | 92.7 | 105.0 | 112.8 |
| | Pre(ours) + PGBIG [36] | 11.1 | **47.9** | 77.2 | 91.7 | 103.5 | 110.8 |
| | Pre(ours) + TCD (ours) | **10.8** | 49.9 | **74.4** | **84.9** | **95.1** | **104.2** |
| Perfect Observation | LTD-50-25 [41] | 12.2 | 50.7 | 79.6 | 93.6 | 105.2 | 112.4 |
| | HRI [39] | 10.4 | 47.1 | 77.3 | 91.8 | 104.1 | 112.1 |
| | PGBIG [36] | 10.3 | **46.6** | 76.3 | 90.9 | 102.6 | 110.0 |
| | TCD (ours) | **9.9** | 48.8 | **73.7** | **84.0** | **94.3** | **103.3** |
| Post-processed Prediction | HRI [39] + TCD (ours) | 10.3 | 47.3 | 72.9 | 83.8 | 94.0 | 102.9 |
| | PGBIG [36] + TCD (ours) | 10.2 | 46.1 | 72.4 | 83.6 | 93.9 | 102.8 |


## Source code

github.com/vita-epfl/DePOSit

DePOSit

## References

[1] Mao et al, History repeats itself: Human motion prediction via motion attention, ECCV'20
[2] Ma et al, Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction, CVPR'22
[3] Salzman et al, Motron: Multimodal Probabilistic Human Motion Forecasting, CVPR'22
[4] Mao et al, Generating Smooth Pose Sequences for Diverse Human Motion Prediction, ICCV'21
[5] Ma et al, Multi-Objective Diverse Human Motion Prediction With Knowledge Distillation, CVPR'22
[6] Cui et al, Towards Accurate 3D Human Motion Prediction From Incomplete Observations, CVPR'21