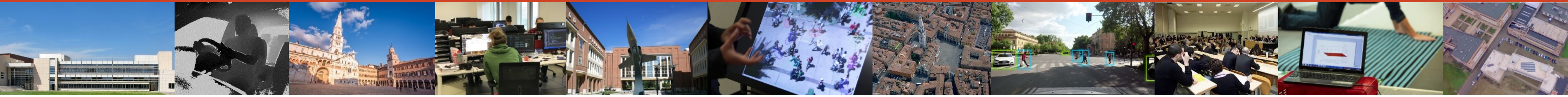
SnapPose3D: Diffusion-Based Single-Frame 2D-to-3D Lifting of Human Poses



Alessandro Simoni[1], Riccardo Catalini[1], Davide Di Nucci[1,2], Guido Borghi[1],
Davide Davoli[2], Lorenzo Garattoni[2], Gianpiero Francesca[2], Yuki Kawana[3], **Roberto Vezzani**[1]

[1]University of Modena and Reggio Emilia, [2]Toyota Motor Europe, [3] Woven by Toyota

SnapPose3D:

Diffusion-Based

Single-Frame

2D-to-3D Lifting

of Human Poses

Problem definition

I'm sorry, this paper is not about Medical Deep Learning nor Breast Imaging

SnapPose3D:

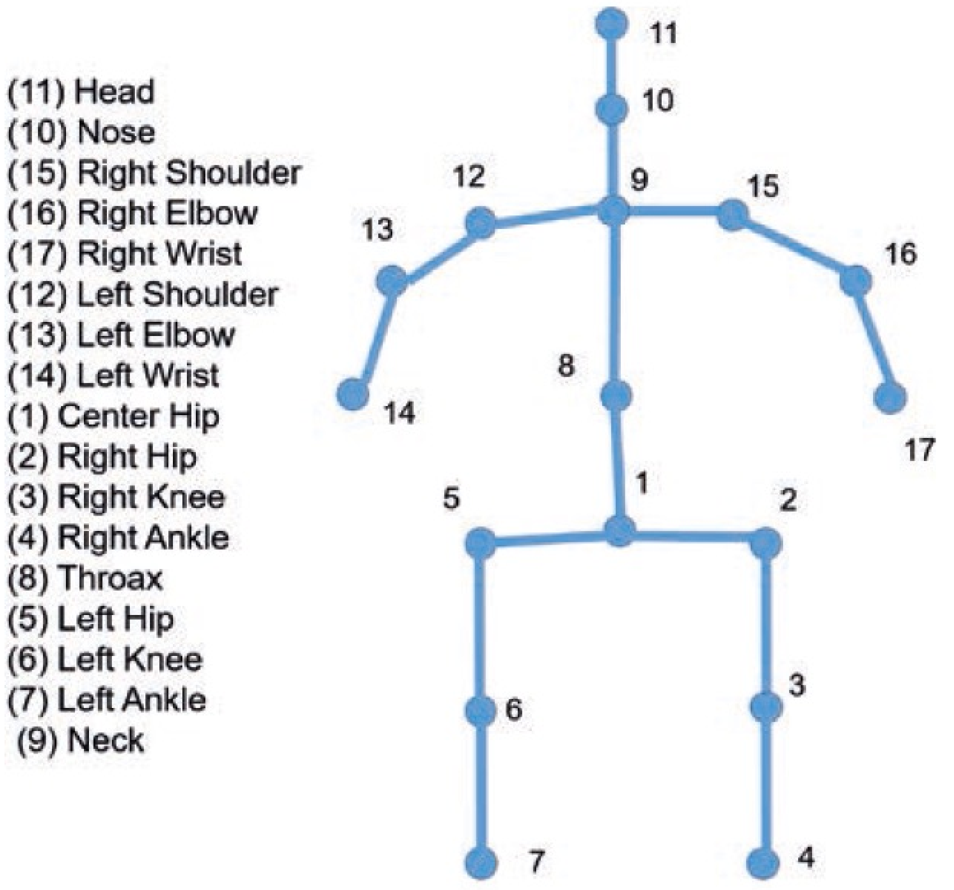
Diffusion-Based

Single-Frame

2D-to-3D Lifting

of Human Poses

Problem definition



SnapPose3D:

Diffusion-Based

Single-Frame

2D-to-3D Lifting

of Human Poses

Problem definition

- Given an image with a person and a 2D pose detector
- Given the coordinates of the joints in the 2D image space
- **Lift them to find the corresponding 3D coordinates (with respect to the camera)**
- During lifting, we allow changes of the X and Y coordinates as well
- A bit different than a backprojection

SnapPose3D:

Diffusion-Based

Single-Frame

2D-to-3D Lifting

of Human Poses

Problem definition

Only one frame as input

- ✓ No motion data
- ✓ No previous frames
- ✓ No tracking
- ✓ No calibration data

SnapPose3D:

Diffusion-Based

Single-Frame

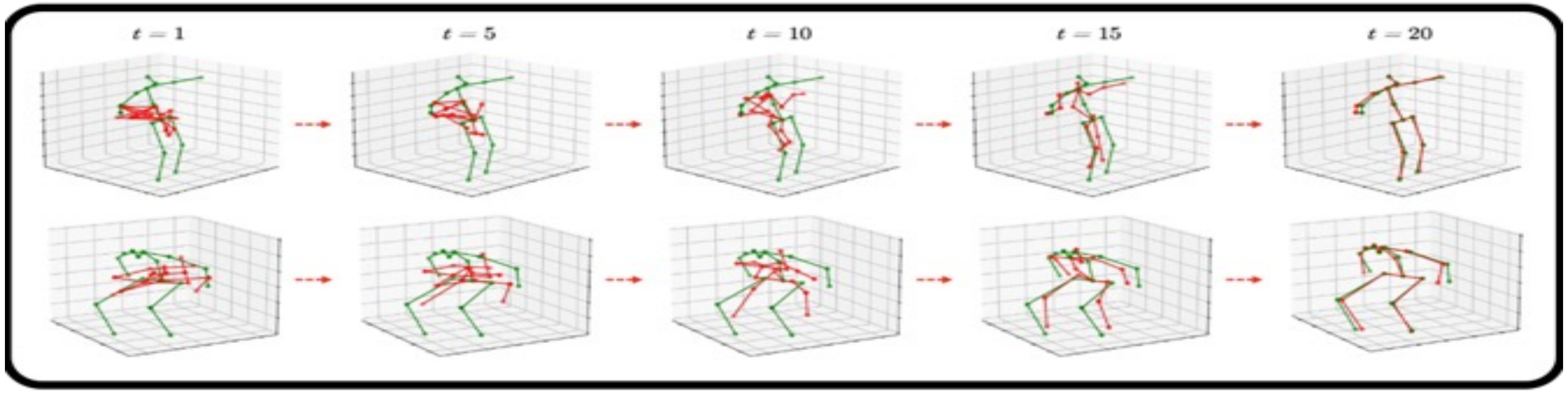
2D-to-3D Lifting

of Human Poses

Problem definition

Randomly generate a 3d posture
(random means implausible and inhuman)

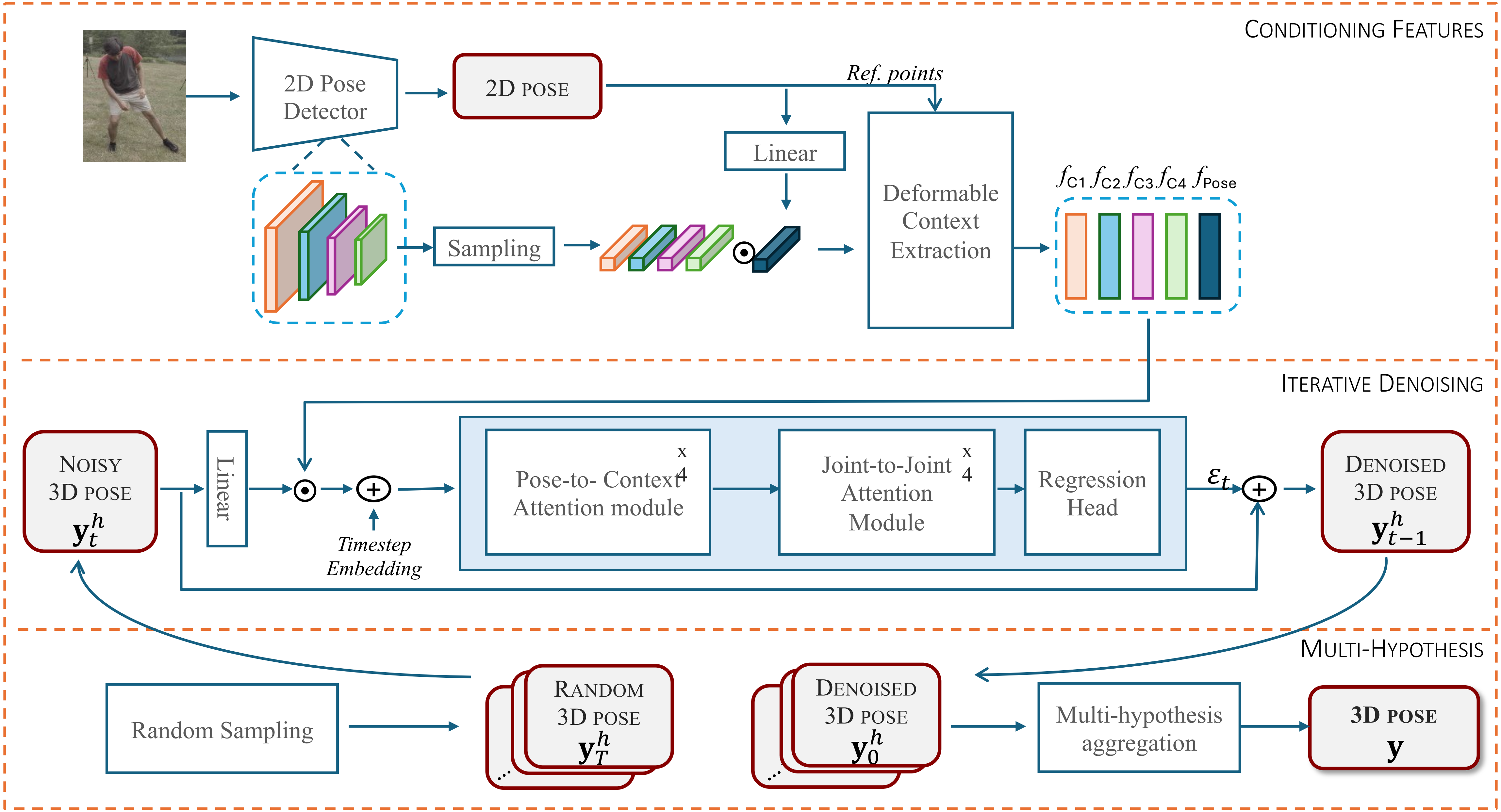
Iteratively refine the posture using
information (conditioning) from the current frame



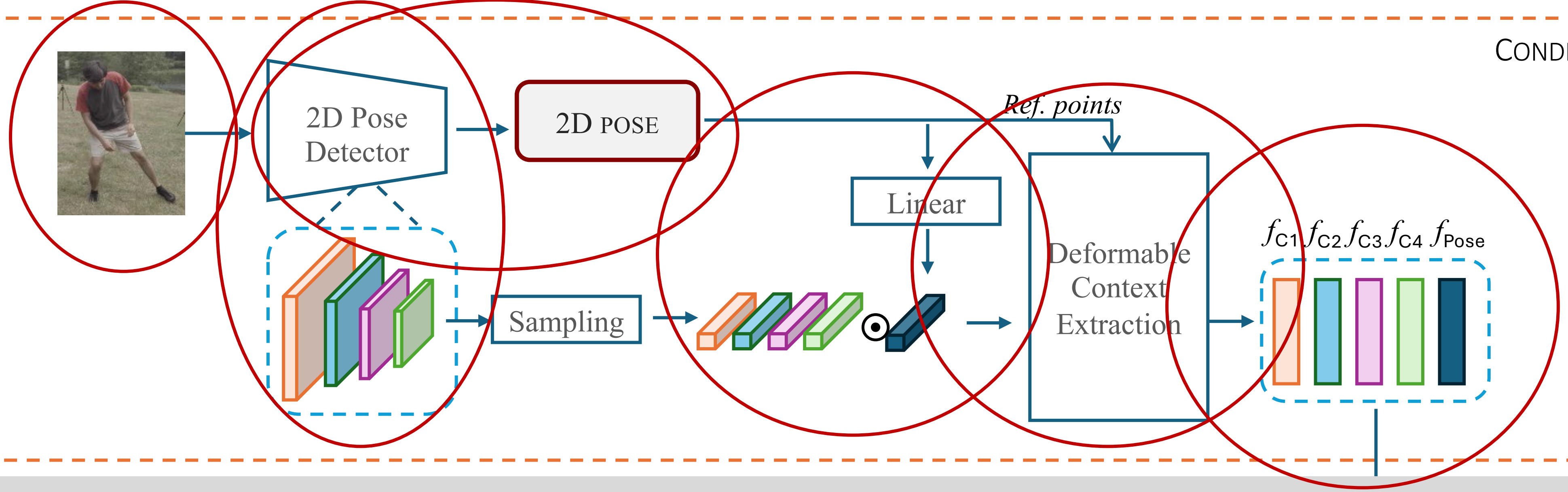
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

TOYOTA
TOYOTA MOTOR EUROPE

THE ALGORITHM

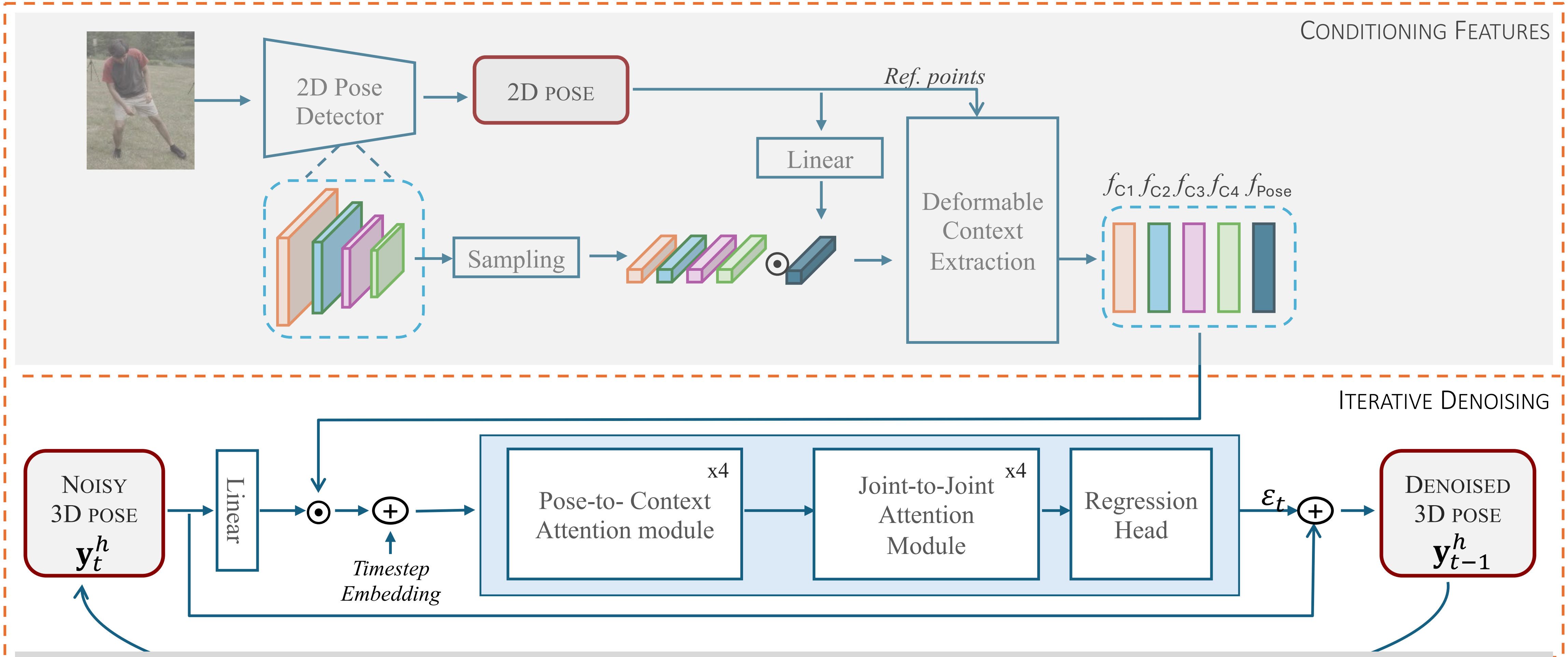


OVERALL SCHEMA



- We start with a crop around a detected person
- We use HRNet32 to obtain the 2D Pose
- We extract context features directly from the 2D pose estimator intermediate layers
- Context features are concatenated with a linear projection of the pose itself
- The features are collected around the 2D keypoints with a deformable module
- Final tensor size: *(L+1) x J x d* (*L* layers + 1 pose = 5) x *J* joints x *d* embedding size

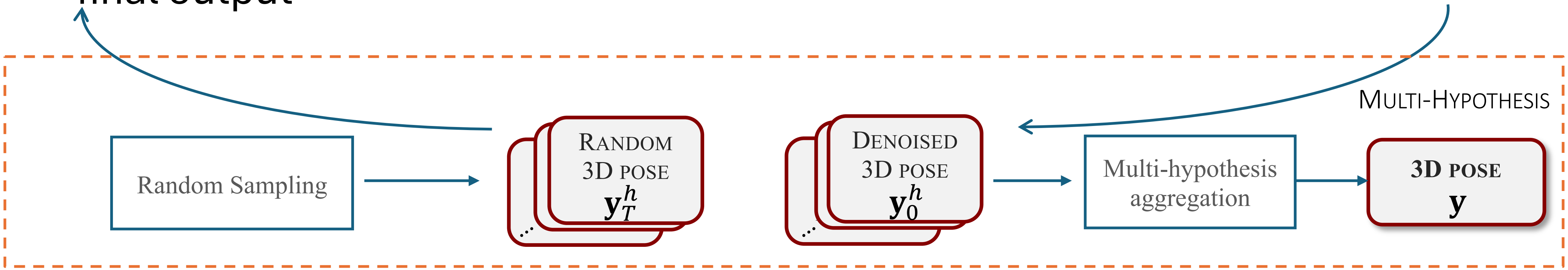
OVERALL SCHEMA



- The tensor is used as conditioning in a denoising module
- A linear projection to a d-size vector of the 3D pose is concatenated to the tensor
- Two attention blocks, one among the 6 channels, and one among the J joints
- Finally, a regression head to convert back the embedding to a 3D pose

Multiple Hypothesis Aggregation

- During inference, SnapPose3D can be used to **generate multiple hypotheses** by sampling different initial 3D poses (from a Gaussian distribution), allowing an arbitrary number (H) of predictions.
- The distribution of the output hypotheses is not Gaussian -> average and median operators provide different results
- We studied the best way to aggregate the hypotheses to generate only one final output



Aggregation / Selection algorithms

not general enough

SELECTION

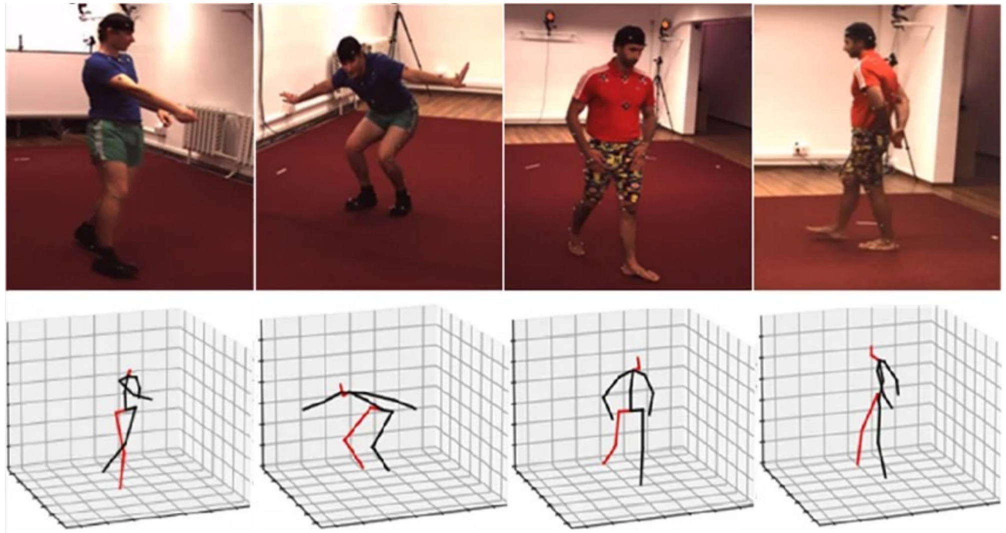
- ~~If the intrinsic camera parameters are given, the re-projection error of the 3D poses when compared with the initial 2Dposes could be exploited as a selection schema~~
- **R (random selection):** a single hypothesis is randomly selected as the final pose estimation among the available H ones. Using H=1 solves the task
- **B (best/oracle selection):** selection of the one closest to the ground truth ("Supervision from an Oracle")
  - it requires the knowledge of the ground truth → not useful in real and online cases

AGGREGATION

- **A (averaging)**: is the standard aggregation approach, though it is sensitive to outliers.
- **M (Median operator)**, which selects the central element after independently sorting the 3 × J vectors (each containing the H values)

Datasets

- **Human3.6M**
  - 11 subjects, 15 daily activities.
  - Train on subjects (S1, S5, S6, S7, S8), test on (S9, S11).
- **H36MA**
  - To underline the advantage of having a multi-hypothesis system
  - **H36MA**[52] is a subset of Human3.6M.
  - H36MA contains 6.4% of the test set with highly ambiguous cases.
- **MPI-INF-3DHP**
  - More than 1.3 million frames
  - 8 actors performing 8 activities
  - official train/test split



Training details

- We train SnapPose3D with a batch size of 128 for 50 epochs on both datasets.
- We use Adam optimizer with a starting learning rate of 6e−4 using a linear decay with factor 0.993.
- We apply **random horizontal flipping as data augmentation** for both images and skeletons.
- We implement SnapPose3D using the PyTorch framework and the Diffusers library for the diffusion-based technique.

Metrics

- Mean Per Joint Position Error (MPJPE)
  - average Euclidean distance between estimated joint 3D coordinates and their ground truth in millimeters.
- P-MPJPE
  - aligned version of the MPJPE
  - MPJPE after rigid alignment of the prediction with the ground truth via Procrustes Analysis.
- PCK
  - Percentage of Correct Keypoints within a **Th=150mm** range
  - Area Under Curve (AUC) metrics for a range of PCK thresholds.

Competitors

- We tested SnapPose3D against several single-frame methods.
- Competitors have been divided into three groups:
  - Methods providing a single hypothesis as output
  - Methods directly comparable with SnapPose, generating multiple hypotheses and aggregating them without any additional knowledge
  - Methods providing multiple outputs but exploiting the ground-truth to select the best result (oracle)

MPJPE on Human3.6M

Protocol 1 (**MPJPE**) (mm)	Dir	Dis	Eat	Gre	Phn	Pht	Pos	Pur	Sit	SitD	Smk	Wait	WlkD	Walk	WlkT	**Avg**
Baseline methods																
Pavlakos [33]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Martinez [28]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Zhao [57]	48.2	60.8	51.8	64.0	64.6	53.6	51.1	67.4	88.7	57.7	73.2	65.6	48.9	64.8	51.9	60.8
Sun [46]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1
Yang [55]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Hossain [15]	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Liu [26]	46.3	52.2	47.3	50.7	55.5	67.1	49.2	**46.0**	60.4	71.1	51.5	50.1	54.5	40.3	43.7	52.4
Xu [53]	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	**59.7**	71.5	51.4	48.6	53.9	39.9	44.1	51.9
Zhao [59]	45.2	50.8	48.0	50.0	54.9	65.0	48.2	47.1	60.2	70.0	51.6	48.7	54.1	**39.7**	43.1	51.8
Zhao [58]	**37.7**	**42.0**	**39.0**	**42.7**	**43.5**	**37.6**	**41.1**	51.4	61.7	**43.4**	**49.4**	**43.0**	**33.9**	45.2	**35.0**	**43.4**
Diffusion-based methods	**Aggregation of multiple hypotheses**															
(H=10,A) Diffupose [6]	43.4	50.7	45.4	50.2	49.6	53.4	48.6	**45.0**	56.9	70.7	47.8	48.2	51.3	43.1	43.4	49.4
(H=5,A) Diffpose [11]	42.8	49.1	45.2	48.7	52.1	63.5	46.3	45.2	58.6	66.3	50.4	47.6	52.0	**37.6**	40.2	49.7
(H=20, A) *SnapPose3D*	**36.4**	41.8	**39.6**	42.1	42.7	**36.7**	**43.0**	53.5	62.0	**42.1**	51.3	40.6	**32.0**	45.3	35.3	43.2
(H=20, M) *SnapPose3D*	43.2	**41.3**	41.4	**39.0**	**40.5**	44.8	43.9	47.5	**46.2**	50.5	**43.2**	**40.0**	41.5	40.5	**34.5**	**42.8**
Diffusion-based methods	**One hypothesis selected**															
(H=200, B) DiffPose [14]	38.1	43.1	**35.3**	43.1	46.6	48.2	39.0	**37.6**	51.9	59.3	41.7	47.6	45.4	37.4	36.0	43.3
(H=20, R) *SnapPose3D*	38.7	44.4	41.6	44.6	45.0	38.8	45.6	55.6	65.3	44.4	54.1	43.1	34.3	48.2	37.7	45.6
(H=20, B) *SnapPose3D*	**37.5**	**35.5**	35.7	**33.4**	**34.9**	**38.6**	**37.8**	41.1	**40.0**	**43.7**	**37.4**	**34.5**	**35.8**	**34.8**	**29.2**	**36.9**

MPJPE on Human3.6M

Protocol 1 (MPJPE) (mm)	Dir	Dis	Eat	Gre	Phn	Pht	Pos	Pur	Sit	SitD	Smk	Wait	WlkD	Walk	WlkT	Avg
Baseline methods																
Pavlakos [33]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Martinez [28]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Zhao [57]	48.2	60.8	51.8	64.0	64.6	53.6	51.1	67.4	88.7	57.7	73.2	65.6	48.9	64.8	51.9	60.8
Sun [46]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1
Yang [55]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Hossain [15]	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Liu [26]	46.3	52.2	47.3	50.7	55.5	67.1	49.2	**46.0**	60.4	71.1	51.5	50.1	54.5	40.3	43.7	52.4
Xu [53]	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	**59.7**	71.5	51.4	48.6	53.9	39.9	44.1	51.9
Zhao [59]	45.2	50.8	48.0	50.0	54.9	65.0	48.2	47.1	60.2	70.0	51.6	48.7	54.1	**39.7**	43.1	51.8
Zhao [58]	**37.7**	**42.0**	**39.0**	**42.7**	**43.5**	**37.6**	**41.1**	51.4	61.7	**43.4**	**49.4**	**43.0**	**33.9**	45.2	**35.0**	**43.4**
Diffusion-based methods	Aggregation of multiple hypotheses															
(H=10,A) Diffupose [6]	43.4	50.7	45.4	50.2	49.6	53.4	48.6	**45.0**	56.9	70.7	47.8	48.2	51.3	43.1	43.4	49.4
(H=5,A) Diffpose [11]	42.8	49.1	45.2	48.7	52.1	63.5	46.3	45.2	58.6	66.3	50.4	47.6	52.0	**37.6**	40.2	49.7
(H=20, A) *SnapPose3D*	**36.4**	41.8	**39.6**	42.1	42.7	**36.7**	**43.0**	53.5	62.0	**42.1**	51.3	40.6	**32.0**	45.3	35.3	43.2
(H=20, M) *SnapPose3D*	43.2	**41.3**	41.4	**39.0**	**40.5**	44.8	43.9	47.5	**46.2**	50.5	**43.2**	**40.0**	41.5	40.5	**34.5**	**42.8**
Diffusion-based methods	One hypothesis selected															
(H=200, B) DiffPose [14]	38.1	43.1	**35.3**	43.1	46.6	48.2	39.0	**37.6**	51.9	59.3	41.7	47.6	45.4	37.4	36.0	43.3
(H=20, R) *SnapPose3D*	38.7	44.4	41.6	44.6	45.0	38.8	45.6	55.6	65.3	44.4	54.1	43.1	34.3	48.2	37.7	45.6
(H=20, B) *SnapPose3D*	**37.5**	**35.5**	35.7	**33.4**	**34.9**	**38.6**	**37.8**	41.1	**40.0**	**43.7**	**37.4**	**34.5**	**35.8**	**34.8**	**29.2**	**36.9**

P-MPJPE on Human3.6M

Protocol 2 (**P-MPJPE**) (mm)	Dir	Dis	Eat	Gre	Phn	Pht	Pos	Pur	Sit	SitD	Smk	Wait	WlkD	Walk	WlkT	**Avg**
Baseline methods																
Sun [46]	42.1	44.3	45.0	45.4	51.5	53.0	43.2	41.3	59.3	73.3	51.0	44.0	48.0	38.3	44.8	48.3
Martinez [28]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Hossain [15]	36.9	37.9	42.8	40.3	46.8	46.7	37.7	36.5	48.9	52.6	45.6	39.6	43.5	35.2	38.5	42.0
Pavlakos [33]	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Liu [26]	35.9	40.0	38.0	41.5	42.5	51.4	37.8	36.0	48.6	56.6	41.8	38.3	42.7	**31.7**	36.2	41.2
Yang [55]	**26.9**	**30.9**	36.3	39.9	43.9	47.4	**28.8**	**29.4**	**36.9**	58.4	41.5	**30.5**	29.5	42.5	32.2	37.7
Zhao [58]	32.3	34.6	**33.1**	**35.1**	**35.2**	**30.1**	32.1	42.6	48.8	**36.0**	**39.1**	33.1	**27.5**	37.1	**29.3**	**35.4**
Diffusion-based methods	**Aggregation of multiple hypotheses**															
(H=10,A) Diffupose [6]	35.9	40.3	36.7	41.4	39.8	43.4	37.1	35.5	46.2	59.7	39.9	38.0	41.9	32.9	34.2	39.9
(H=5, A) Diffpose [11]	33.9	38.2	36.0	39.2	40.2	46.5	35.8	**34.8**	48.0	52.5	41.2	36.5	40.9	**30.3**	33.8	39.2
(H=20, A) *SnapPose3D*	**31.5**	34.7	**32.7**	35.1	34.2	**29.9**	**33.1**	42.8	47.5	**34.4**	38.6	31.6	**26.5**	36.7	29.7	34.8
(H=20, M) *SnapPose3D*	35.7	**34.3**	34.0	**32.5**	**32.9**	36.7	35.6	38.2	**37.1**	38.4	**34.3**	**31.5**	32.8	32.4	**29.0**	**34.5**
Diffusion-based methods	**One hypothesis selected**															
(H=200, B) DiffPose [14]	**28.1**	31.5	**28.0**	30.8	33.6	35.3	**28.5**	**27.6**	40.8	44.6	31.8	32.1	32.6	28.1	26.8	32.0
(H=20, R) *SnapPose3D*	34.4	37.7	35.1	38.1	36.9	32.4	36.3	45.5	51.7	37.2	42.0	34.6	29.2	40.1	32.6	37.8
(H=20, B) *SnapPose3D*	30.8	**29.2**	29.0	**27.8**	**28.2**	**31.6**	30.6	33.0	**32.1**	**33.3**	**29.6**	**27.2**	**28.3**	**27.9**	**24.5**	**29.7**

P-MPJPE on Human3.6M

Protocol 2 (**P-MPJPE**) (mm)	Dir	Dis	Eat	Gre	Phn	Pht	Pos	Pur	Sit	SitD	Smk	Wait	WlkD	Walk	WlkT	**Avg**
Baseline methods																
Sun [46]	42.1	44.3	45.0	45.4	51.5	53.0	43.2	41.3	59.3	73.3	51.0	44.0	48.0	38.3	44.8	48.3
Martinez [28]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Hossain [15]	36.9	37.9	42.8	40.3	46.8	46.7	37.7	36.5	48.9	52.6	45.6	39.6	43.5	35.2	38.5	42.0
Pavlakos [33]	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Liu [26]	35.9	40.0	38.0	41.5	42.5	51.4	37.8	36.0	48.6	56.6	41.8	38.3	42.7	**31.7**	36.2	41.2
Yang [55]	**26.9**	**30.9**	36.3	39.9	43.9	47.4	**28.8**	**29.4**	**36.9**	58.4	41.5	**30.5**	29.5	42.5	32.2	37.7
Zhao [58]	32.3	34.6	**33.1**	**35.1**	**35.2**	**30.1**	32.1	42.6	48.8	**36.0**	**39.1**	33.1	**27.5**	37.1	**29.3**	**35.4**
Diffusion-based methods	Aggregation of multiple hypotheses															
(H=10,A) Diffupose [6]	35.9	40.3	36.7	41.4	39.8	43.4	37.1	35.5	46.2	59.7	39.9	38.0	41.9	32.9	34.2	39.9
(H=5, A) Diffpose [11]	33.9	38.2	36.0	39.2	40.2	46.5	35.8	**34.8**	48.0	52.5	41.2	36.5	40.9	**30.3**	33.8	39.2
(H=20, A) *SnapPose3D*	**31.5**	34.7	**32.7**	35.1	34.2	**29.9**	**33.1**	42.8	47.5	**34.4**	38.6	31.6	**26.5**	36.7	29.7	34.8
(H=20, M) *SnapPose3D*	35.7	**34.3**	34.0	**32.5**	**32.9**	36.7	35.6	38.2	**37.1**	38.4	**34.3**	**31.5**	32.8	32.4	**29.0**	**34.5**
Diffusion-based methods	One hypothesis selected															
(H=200, B) DiffPose [14]	**28.1**	31.5	**28.0**	30.8	33.6	35.3	**28.5**	**27.6**	40.8	44.6	31.8	32.1	32.6	28.1	26.8	32.0
(H=20, R) *SnapPose3D*	34.4	37.7	35.1	38.1	36.9	32.4	36.3	45.5	51.7	37.2	42.0	34.6	29.2	40.1	32.6	37.8
(H=20, B) *SnapPose3D*	30.8	**29.2**	29.0	**27.8**	**28.2**	**31.6**	30.6	33.0	**32.1**	**33.3**	**29.6**	**27.2**	**28.3**	**27.9**	**24.5**	**29.7**

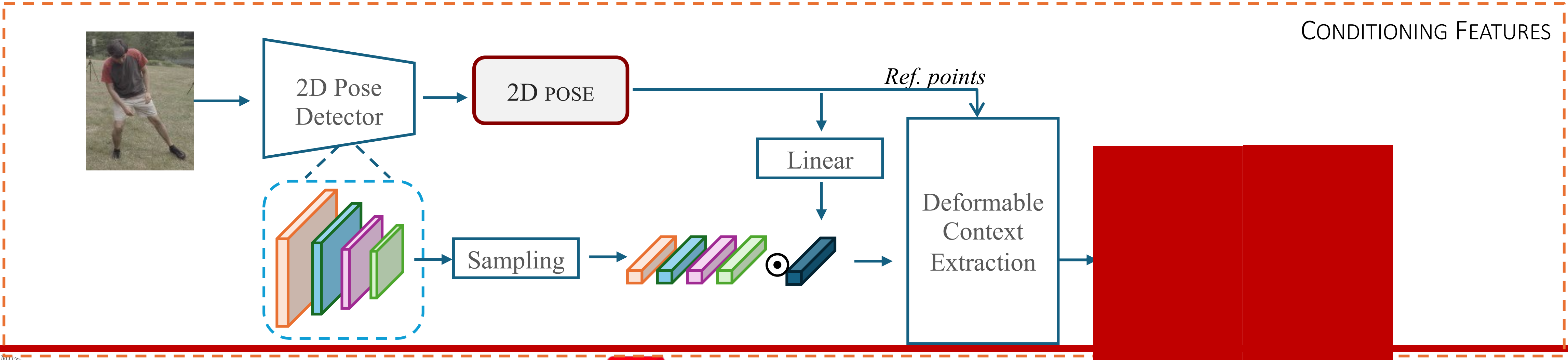
Results on MPI-INF-3DHP

Method	Pavvlo [34]	Zheng [60]	Wang [51]	Li [24]	Zheng [60]	Zhang [56]	Zhao [58]	*SnapPose3D* (H=20, R)	*SnapPose3D* (H=20, M)
PCK ↑	86.0	88.6	86.9	93.8	95.4	94.4	96.8	96.1	**97.2**
AUC ↑	51.9	56.4	62.1	63.3	63.2	66.5	70.7	70.2	**71.4**
MPJPE ↓	84.0	77.1	68.1	58.0	57.7	54.9	44.7	45.5	**42.2**

Ablation: role of the Context and the 2D Pose

Pose	Context	Multiple Hypothesis Aggregation A (avg)	M (median)	B (Pose-level)	B (Joint-level)
✓		59.9	59.8	56.8	49.9
	✓	45.3	45.4	42.3	37.6
✓	✓	**43.2**	**42.8**	**36.9**	**28.6**

Masking part of the conditioning
Metric: MPJPE
Dataset: Human3.6M



Confidence Score

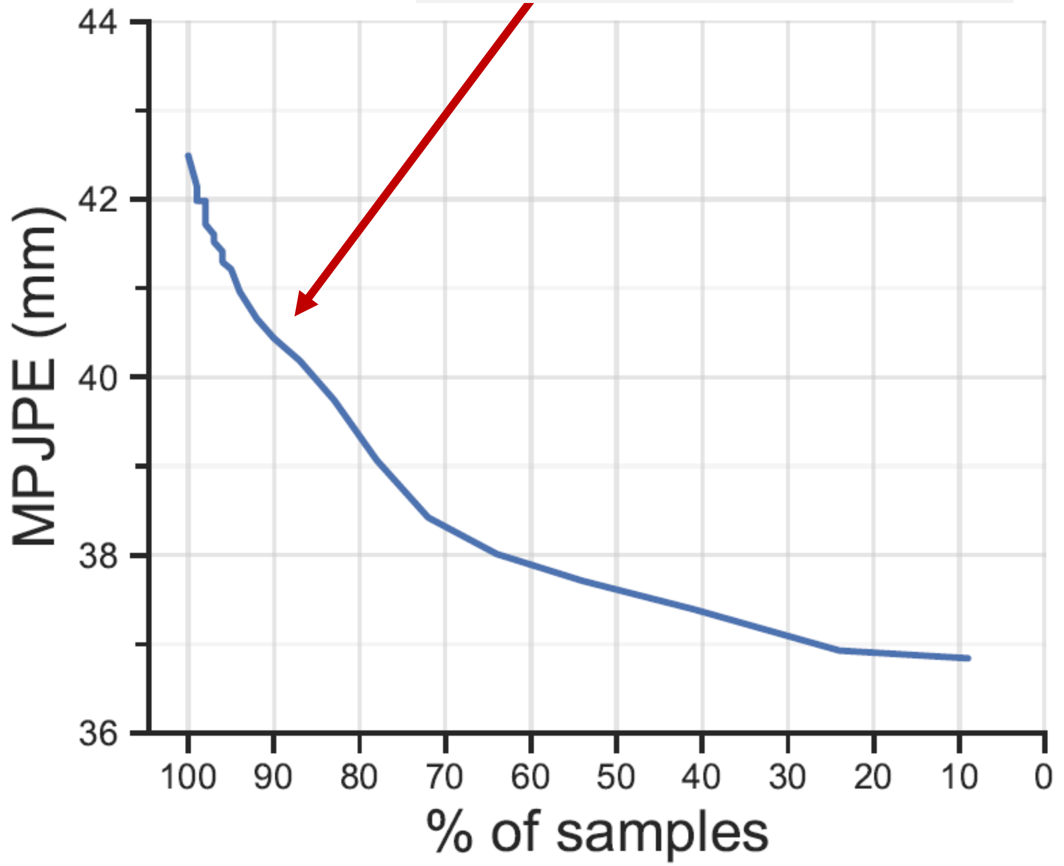
- The generation of multiple hypotheses allows:
  - improvement of performance
  - the definition of a confidence score.

- Variance (between the H different solutions) as Confidence score of the final prediction

- Graph: ordering the results by confidence and keeping a % of samples

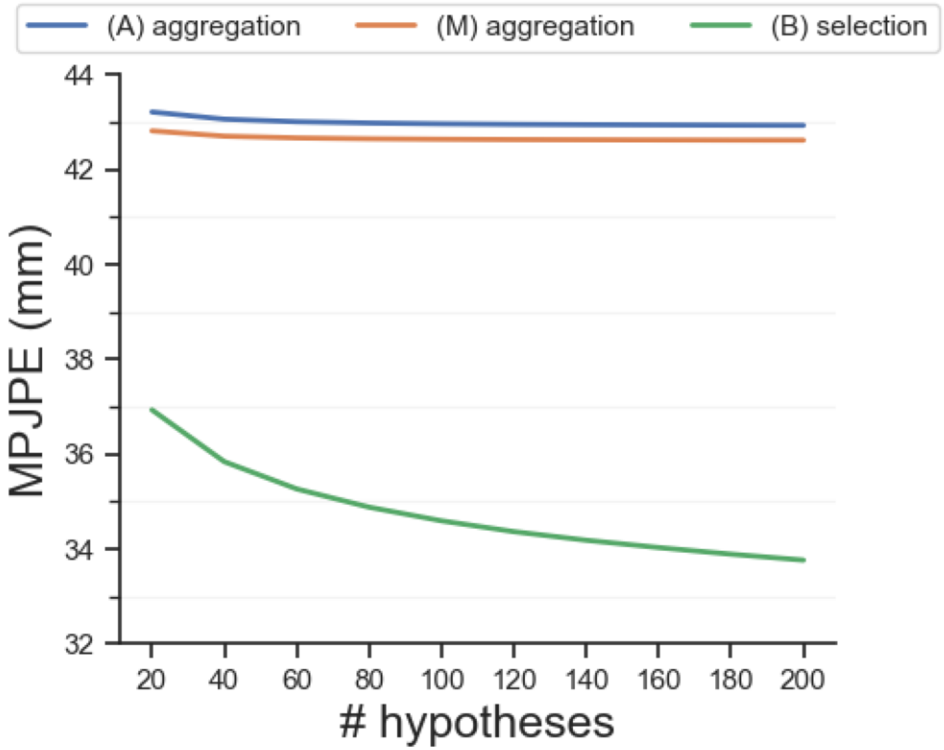
- The confidence metric does not require the ground truth

Reduction of 2mm of the error by removing the 10% of most unreliable outputs



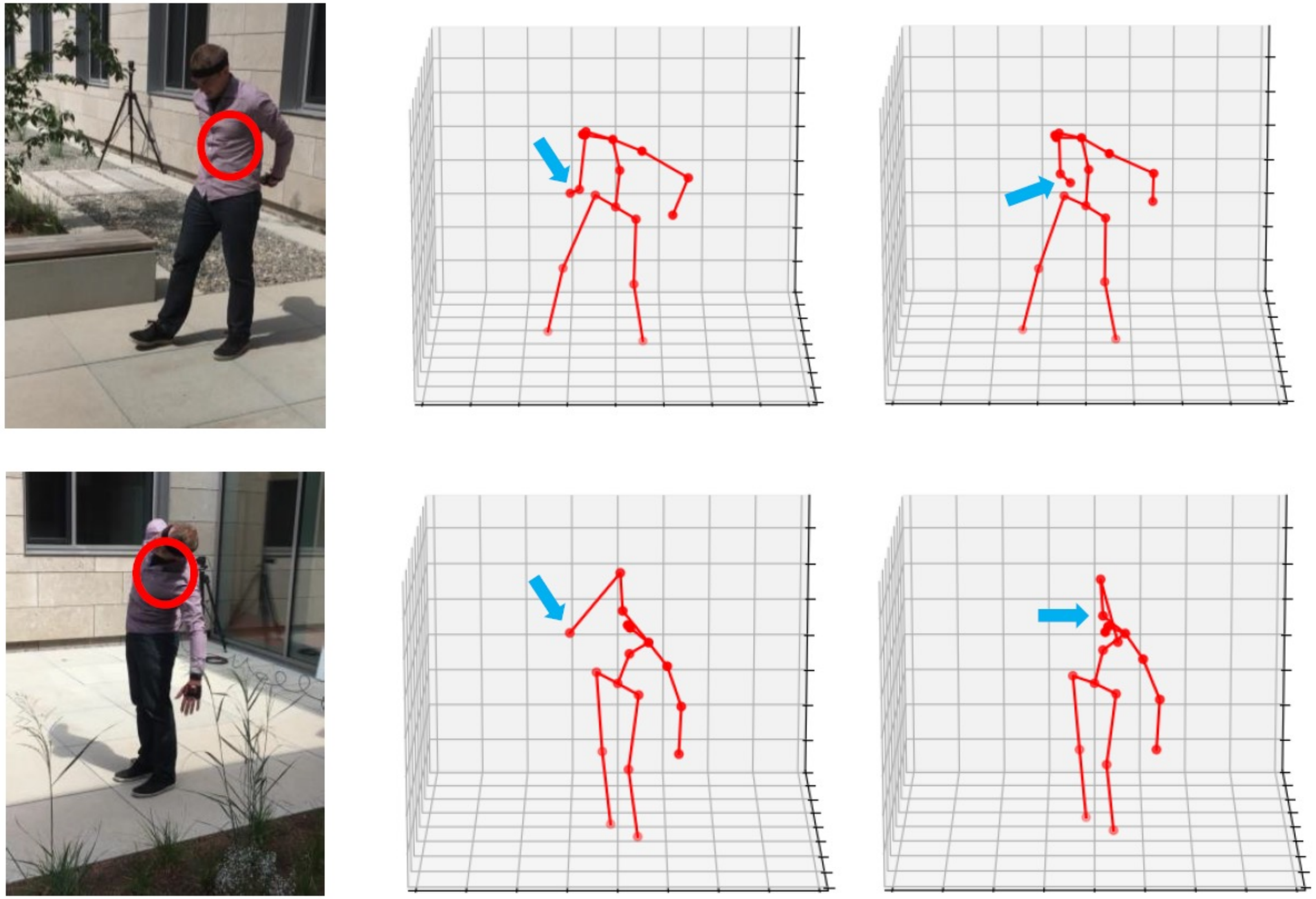
Number of generated hypotheses

- MPJPE error (Protocol 1), Human3.6M
- Error does not decrease a lot after H=40 using aggregation
- Use higher values of H only if you use an oracle selection
- H=20 is a good trade off between performance and computational cost



In-the-wild videos - 3DPW

Some qualitative results

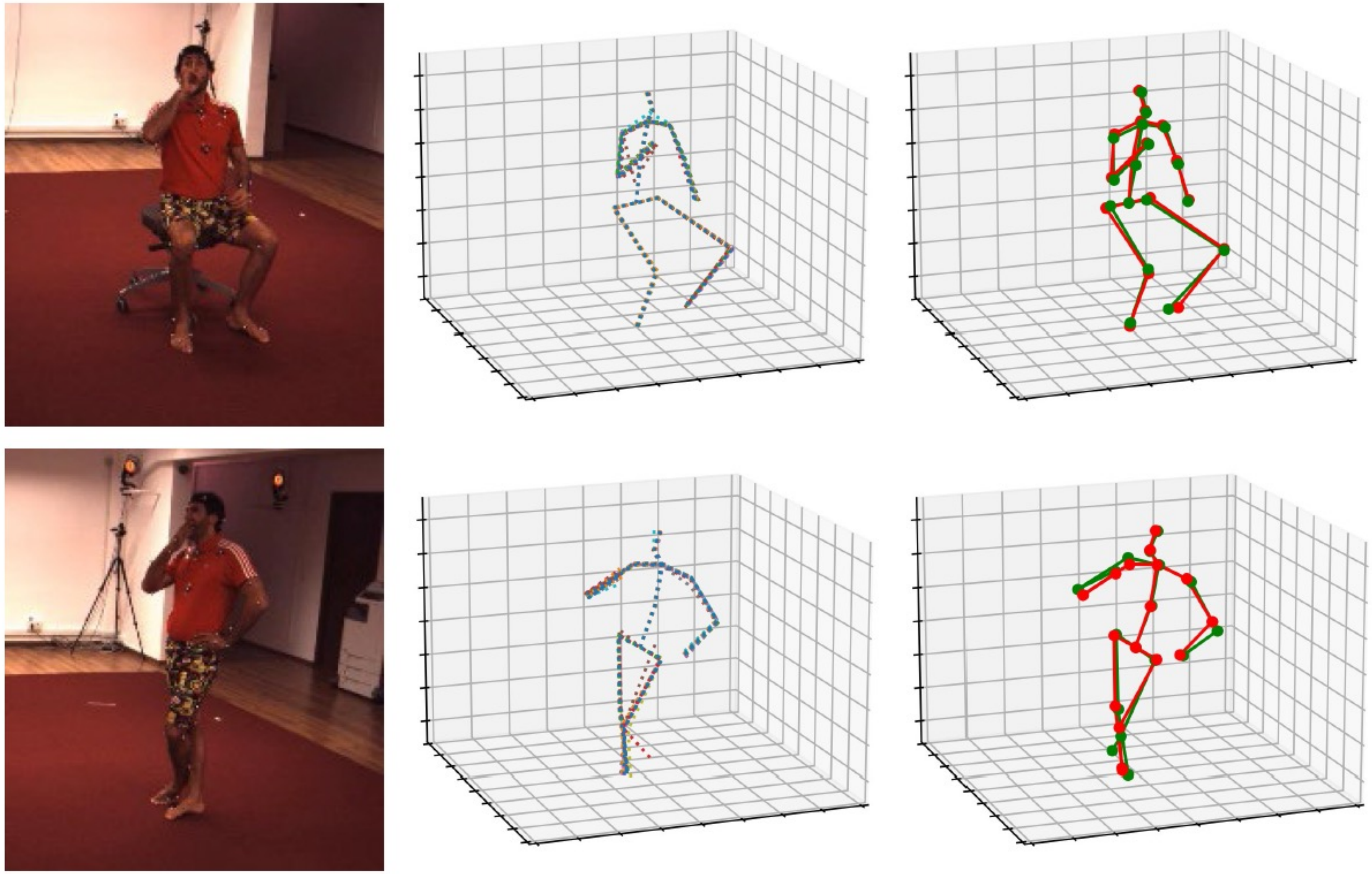


Zhao et al. [4]

SnapPose3D

HUMAN 3.6M

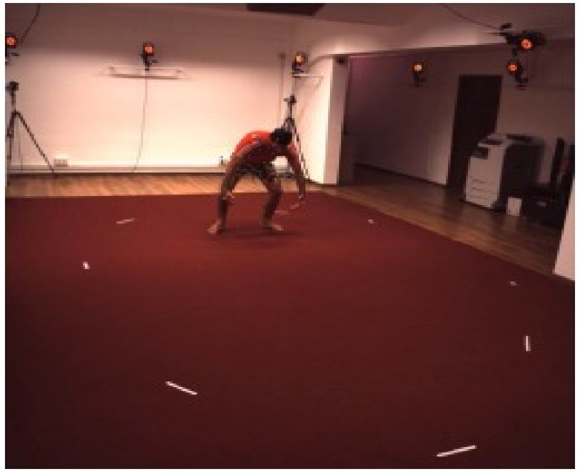
Some qualitative results

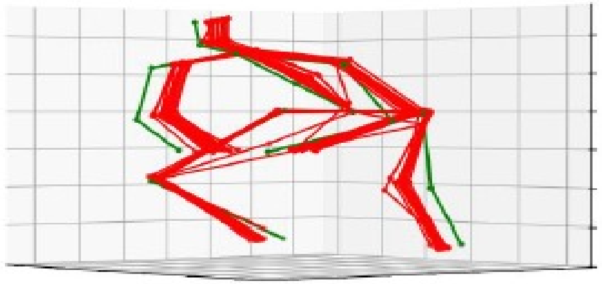


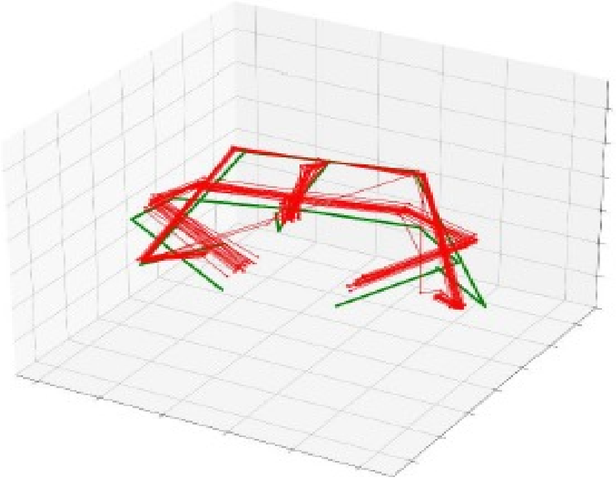
20 hypotheses SnapPose3D and GT

H36MA

Some qualitative results







Conclusions

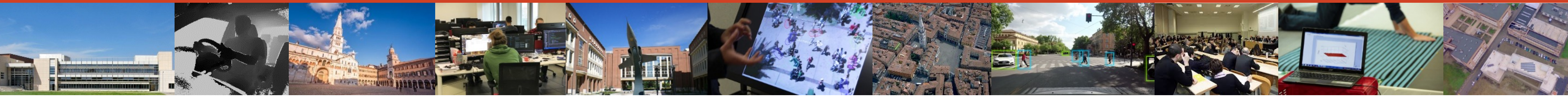
- A 2D to 3D Pose lifting has been proposed
- The method works on a single image input
- The denoising diffusion model can generate multiple plausible hypothese conditioned on the context and the 2d pose
- Aggregating the set of generated solutions improves the results and allows to compute a reliabily (confidence) score

- Connected work: SnapPose3D has been used as core block in a simultaneous 3D pose & 3D gaze estimation (**GazeD**)

ICPR 2026
International Conference on Pattern Recognition LYON

Thank you
for your attention

SnapPose3D: Diffusion-Based Single-Frame 2D-to-3D Lifting of Human Poses



Alessandro Simoni[1], Riccardo Catalini[1], Davide Di Nucci[1,2], Guido Borghi[1],
Davide Davoli[2], Lorenzo Garattoni[2], Gianpiero Francesca[2], Yuki Kawana[3], **Roberto Vezzani**[1]

[1]University of Modena and Reggio Emilia, [2]Toyota Motor Europe, [3] Woven by Toyota

UNIVERSITAS STUDIORUM MUTINENSIS ET REGIENSIS 1175

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

TOYOTA
TOYOTA MOTOR EUROPE