

Long-term Human Motion Prediction with Scene Context

Zhe Cao¹, Hang Gao¹, Karttikeya Mangalam¹, Qi-Zhi Cai²,
Minh Vo³, and Jitendra Malik¹

UC Berkeley
Nanjing University
Facebook Reality Lab

ECCV 2020 (Oral)

By Mahdiar

Problem Long-term Human Motion Prediction

Given the scene image and the persons past pose and location history in 2D, predict his future poses and locations in 3D



View-Point

- Human movement is goal-directed
- It is influenced by the spatial layout of the objects in the scene
 - Heading towards the window
 - Finding a path through the space avoiding collisions
 - It is crucial to perceive the environment
 - If not navigate a new room with lights off



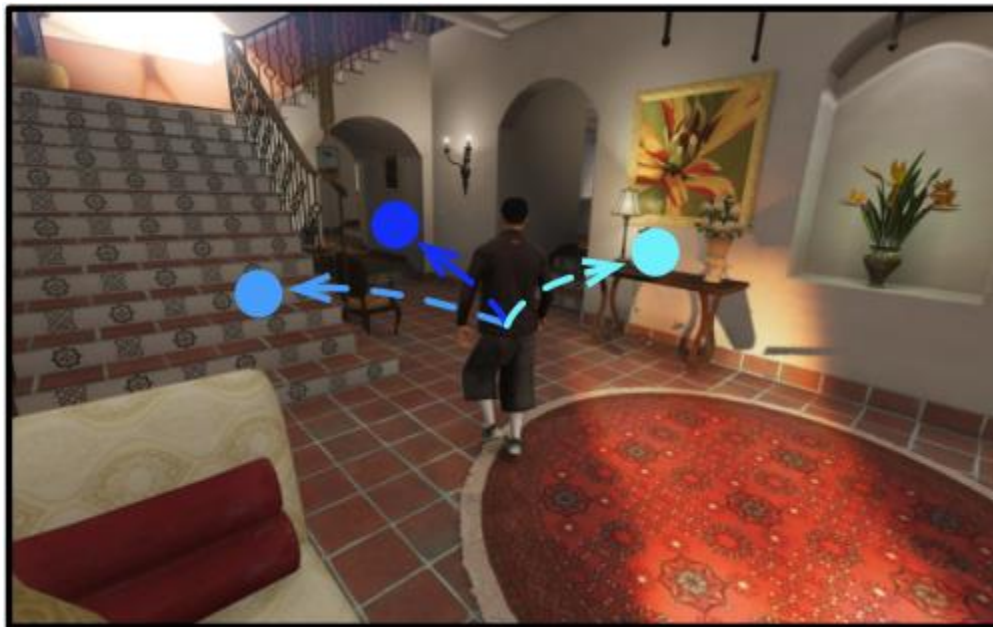
Intuitive Solution

- When we decide to do sth
 - First defining the goal



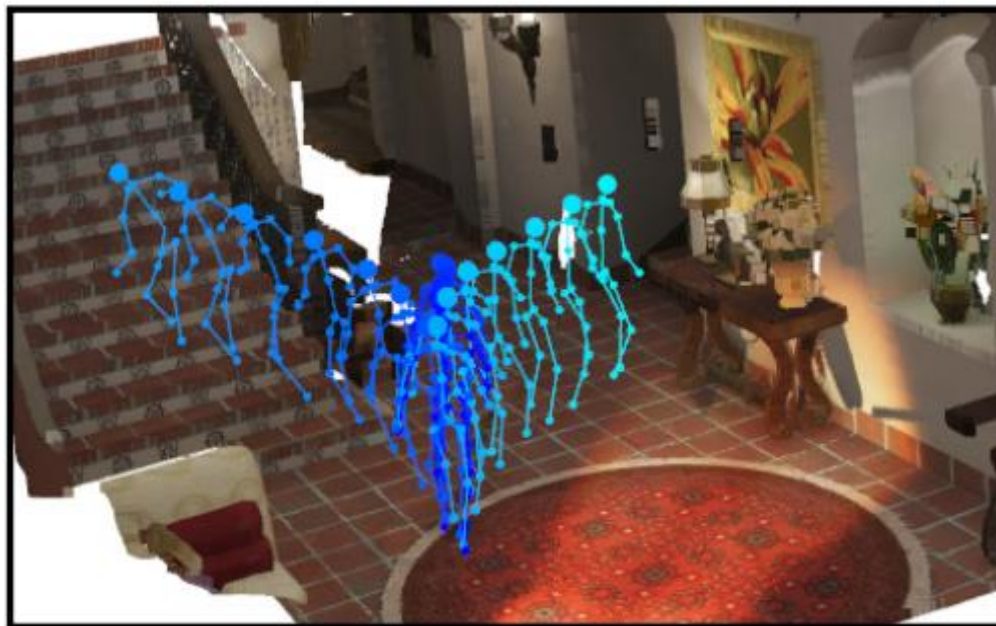
Intuitive Solution

- When we decide to do sth
 - Then planning a plausible path



Intuitive Solution

- When we decide to do sth
 - Finally taking the action



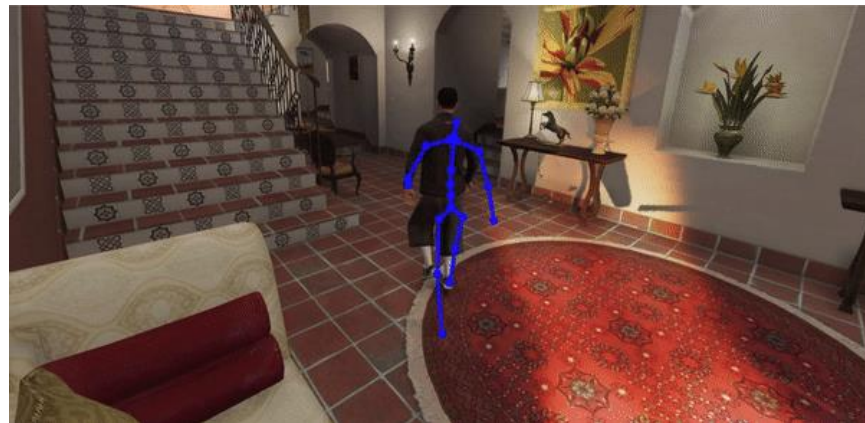
The Core Idea

1. To understand long term behavior, we must reason in terms of goals
2. Plausibility of the path using the scene context

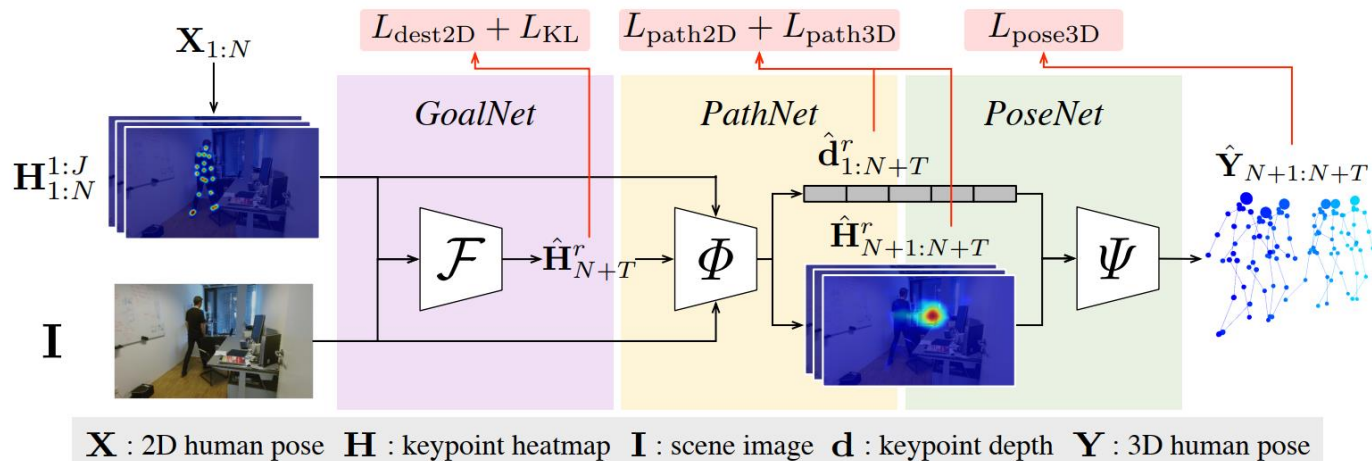
Problem Statement

Given 2D pose history and 2d scene context

Predict the next T-step 3D human poses



Overall Pipeline



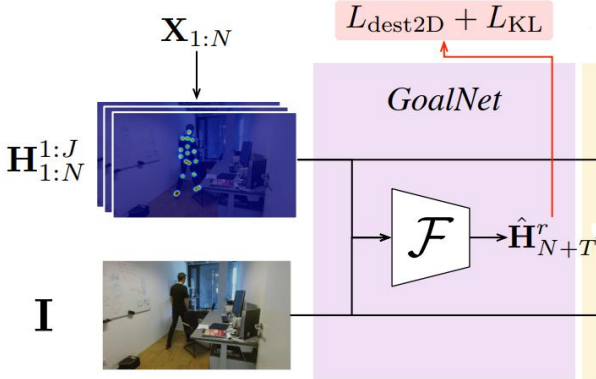
- Inputs
 - A N-step 2D human pose history $X_{1:N}$
 - 2D image of the scene I (the Nth frame scene context)
- Outputs
 - Next T future poses in 3D $Y_{N+1:N+T}$

GoalNet

Allow the model to express uncertainty of human motion

By learning a distribution of possible motion destinations, instead of a single hypothesis.

Using conditional VAE to learn the distribution (latent space)



Index	Input	Data	Operator	Output shape
(1)	-	scene image	-	$3 \times 256 \times 448$
(2)	-	stacked heatmaps	-	$(N \times J) \times 64 \times 112$
(3)	(1)		$7 \times 7, \text{ stride } 2$	$64 \times 128 \times 224$
(4)	(3)		MaxPool, stride 2	$64 \times 64 \times 112$
(5)	(4)		$3 \times 3, \text{ stride } 1$	$64 \times 64 \times 112$
(6)	(2)		$3 \times 3, \text{ stride } 1$	$64 \times 64 \times 112$
(7)	(5), (6)		$3 \times 3, \text{ stride } 2$ $3 \times 3, \text{ stride } 1$	$128 \times 32 \times 56$
(8)	(7)		$3 \times 3, \text{ stride } 2$ $3 \times 3, \text{ stride } 1$	$256 \times 16 \times 28$
(9)	(8)		$3 \times 3, \text{ stride } 2$ $3 \times 3, \text{ stride } 1$	$512 \times 8 \times 14$
(10)	(9)	encoder feat.	GlobalAvgPool	$512 \times 1 \times 1$
(11)	(10)	μ	Linear	$Z \times 1 \times 1$
(12)	(10)	σ	Linear	$Z \times 1 \times 1$
(13)	(11), (12)	\mathbf{z}	Sample from $\mathcal{N}(\mu, \sigma)$	$Z \times 8 \times 14$
(14)	(13)		$3 \times 3, \text{ stride } 1$	$512 \times 8 \times 14$
(15)	(14)		$3 \times 3, \text{ stride } 1$	$512 \times 8 \times 14$
(16)	(15)		Upsample $2 \times$	$512 \times 16 \times 28$
(17)	(16)		$3 \times 3, \text{ stride } 1$	$256 \times 16 \times 28$
(18)	(17)		Upsample $2 \times$	$256 \times 32 \times 56$
(19)	(18)		$3 \times 3, \text{ stride } 1$	$128 \times 32 \times 56$
(20)	(19)		Upsample $2 \times$	$128 \times 64 \times 112$
(21)	(20)	decoder feat.	$3 \times 3, \text{ stride } 1$ $3 \times 3, \text{ stride } 1$	$64 \times 64 \times 112$
(22)	(21)	goal heatmap pred.	$1 \times 1, \text{ stride } 1$	$1 \times 64 \times 112$

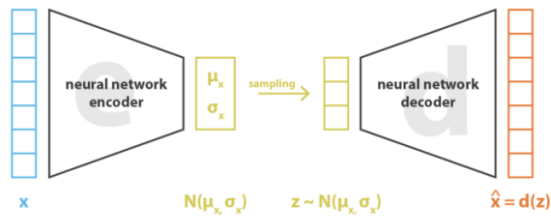
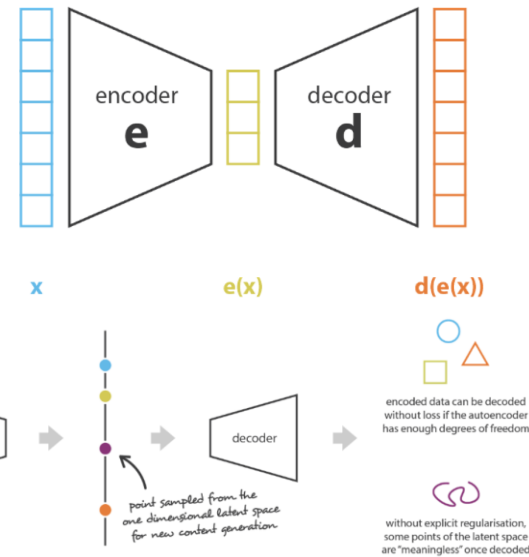
Quick Review of VAE

- Simple auto-encoder
 - Goal: compress the data into a smaller representation
 - Loss: reconstruction loss

$$Loss = ||x - d(e(x))||^2$$

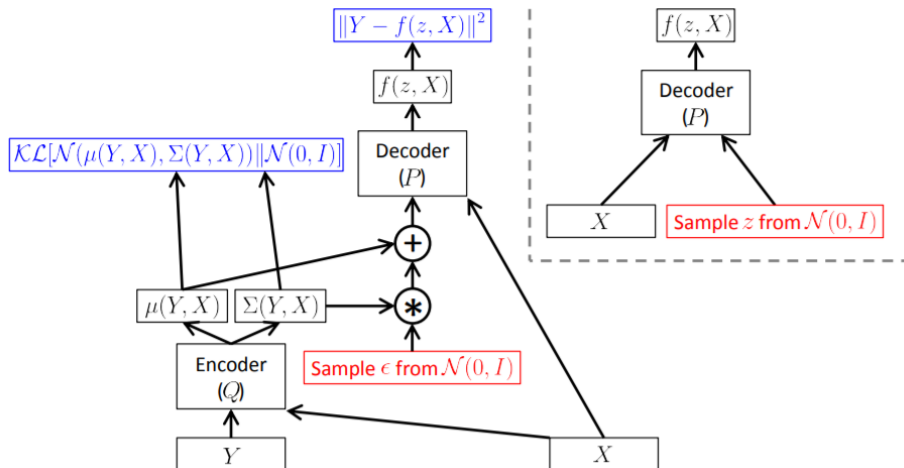
- Can not be used for generative process
- Variational auto-encoder
 - Adding a regularization term in the loss function

$$Loss = ||x - \hat{x}||^2 + KL[N(\mu_x, \sigma_x), N(0, I)]$$



Quick Review of VAE

- Conditional Variational Auto-encoder
 - No control on the data generation process on VAE
 - From where should be pick the sample (e.g. MNIST generation)?



GoalNet

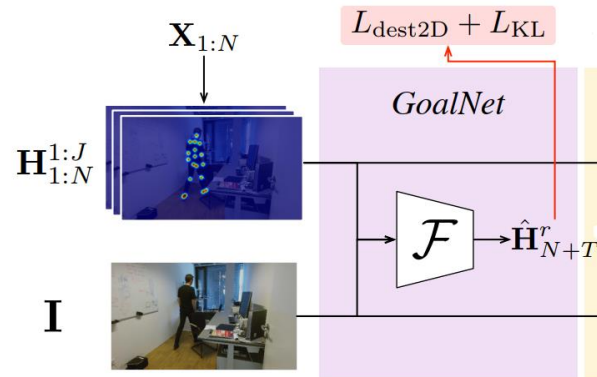
- Learns a distribution of possible 2D destinations
- Conditioned on the 2D pose history and the scene

$$z \sim Q(z | H_{1:N}^{1:J}, I) \equiv \mathcal{N}(\mu, \sigma)$$
$$\mu, \sigma = F_{enc}(H_{1:N}^{1:J}, I)$$

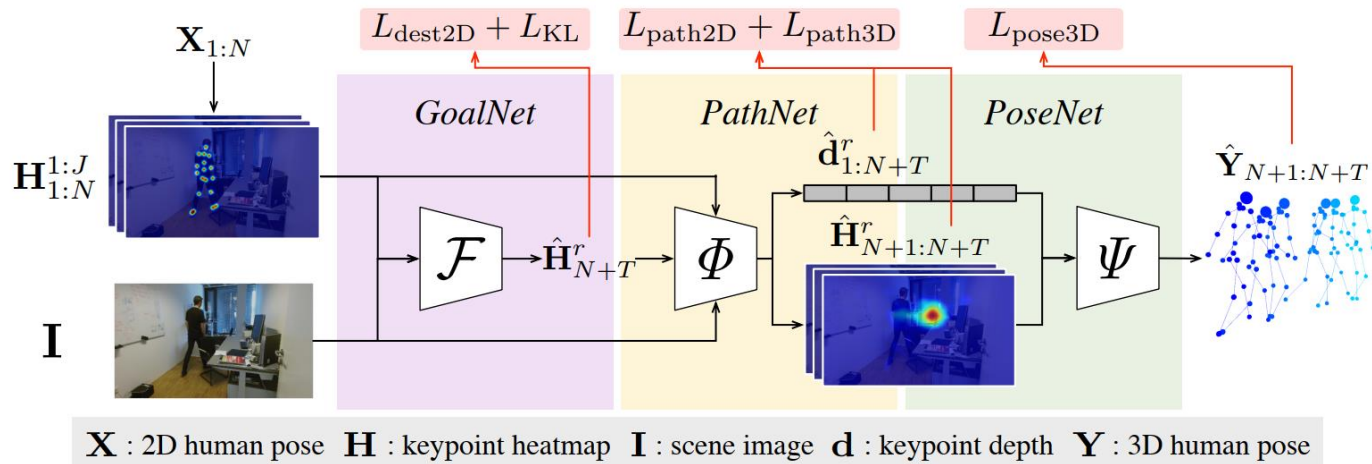
- Decoding
- The whole process

$$\hat{H}_{N+T}^r = F_{dec}(z, I)$$
$$\hat{H}_{N+T}^r = F(H_{1:N}^{1:J}, I)$$
$$L_{dest2D} = \left\| X_{N+T}^r - \widehat{X}_{N+T}^r \right\|$$
$$L_{KL} = KL[Q(z | H_{1:N}^{1:J}, I) || \mathcal{N}(0, 1)]$$

- During test: samples a set of latent variables $\{z\}$ from $\mathcal{N}(0, 1)$ and map them to multiple plausible 2D destinations



Overall Pipeline



PathNet

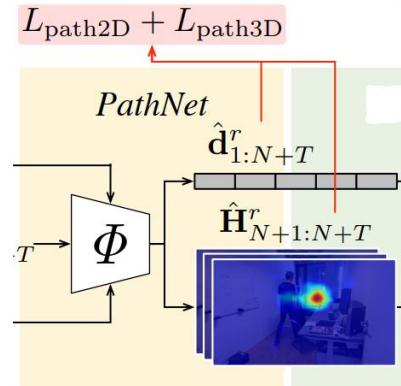
- The destination determines where to move
- The scene context determines how to move
- Input
 - Pose history, scene image, 2d destination
- Backbone
 - Hourglass54
- Output

- Global 3D path $(\hat{H}_{N+1:N+T}^r, \hat{d}_{1:N+T}^r)$
- The 3D path $\hat{Y}_{1:N+T}^r$ would be obtained using
 - Depth $(\hat{d}_{1:N+T}^r)$, 2D path $\hat{X}_{1:N+T}^r$, Camera intrinsic (K)

$$L_{path2D} = \|X_{N+1:N+T}^r - \hat{X}_{N+1:N+T}^r\|$$

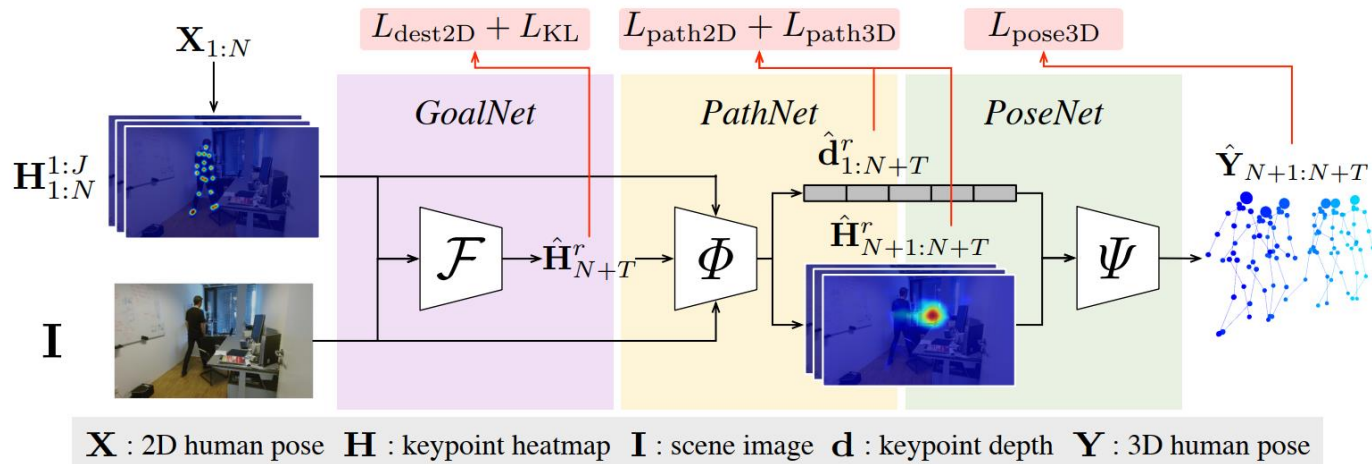
$$L_{path3D} = \left\| Y_{1:N+T}^r - \hat{Y}_{1:N+T}^r \right\| + \left\| \hat{Y}_{1:N+T-1}^r - \hat{Y}_{2:N+T}^r \right\|$$

- During training: the ground-truth destination to train
- During testing: predictions from the GoalNet



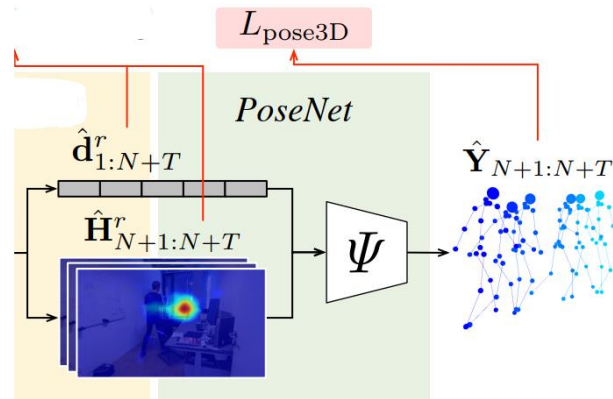
Index	Input	Data	Operator	Output shape
(1)	-	scene image	-	$3 \times 256 \times 448$
(2)	-	stacked heatmaps	-	$(N \times J) \times 256 \times 448$
(3)	-	goal heatmap	-	$1 \times 256 \times 448$
(4)	-	initial depth.	-	$N \times 1 \times 1$
(5)	-	2D pose sequence	-	$N \times J \times 2$
(6)	(1), (2), (3)		7×7 , stride 2	$128 \times 128 \times 224$
(7)	(6)		$\begin{bmatrix} 3 \times 3, \text{stride } 2 \\ 3 \times 3, \text{stride } 1 \end{bmatrix}$	$256 \times 64 \times 112$
(8)	(7)	backbone feat ₁ .	HourglassStack	$256 \times 64 \times 112$
(9)	(8)	backbone feat ₂ .	HourglassStack	$256 \times 64 \times 112$
(10)	(9)	backbone feat ₃ .	HourglassStack	$256 \times 64 \times 112$
(11)	(8) or (9) or (10)		$\begin{bmatrix} 3 \times 3, \text{stride } 1 \\ 3 \times 3, \text{stride } 1 \end{bmatrix}$	$256 \times 64 \times 112$
(12)	(11)	heatmap pred.	1×1 , stride 1	$T \times 64 \times 112$
(13)	(8) or (9) or (10)		$\begin{bmatrix} 3 \times 3, \text{stride } 2 \\ 3 \times 3, \text{stride } 1 \\ 3 \times 3, \text{stride } 2 \\ 3 \times 3, \text{stride } 1 \end{bmatrix}$	$384 \times 32 \times 56$
(14)	(13)		GlobalAvgPool	$512 \times 16 \times 28$
(15)	(14)		Linear	$512 \times 1 \times 1$
(16)	(4), (5)		Linear	$256 \times 1 \times 1$
(17)	(15), (16)		Linear	$256 \times 1 \times 1$
(18)	(17)	depth pred.	Linear	$(N + T) \times 1 \times 1$

Overall Pipeline



PoseNet

- Instead of predicting the pose from scratch
 - First obtain a noisy 3D poses $\bar{Y}_{1:N}$
 - Given $X_{1:N}$ into 3D using the human
 - Torso depth $d_{1:N}^r$
 - Camera intrinsic K



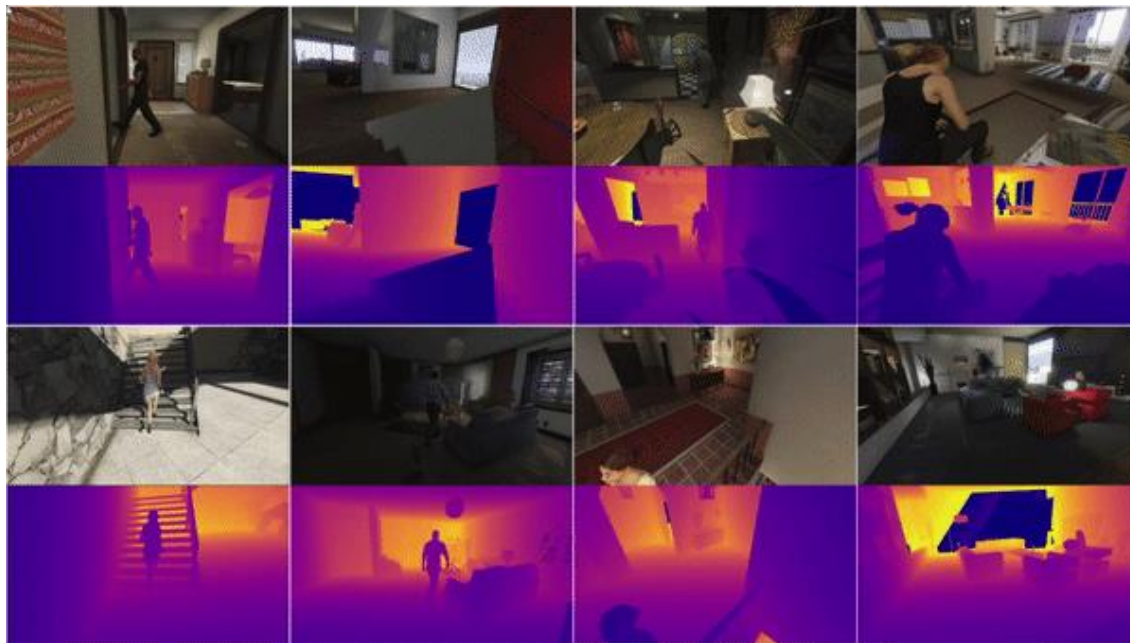
- Next replicate the present 3D pose \bar{Y}_N to the future 3D path location for the initial estimation of future 3D poses $\bar{Y}_{N+1:N+T}$
- Finally revising the initialization using transformer network

$$\hat{Y}_{N+1:N+T} = \psi(\bar{Y}_{1:N+T})$$
$$L_{pose3D} = ||Y_{N+1:N+T} - \hat{Y}_{N+1:N+T}||$$

- During training, ground-truth 3D path is used for estimating 3D pose $Y_{1:N+T}^r$
- During testing, the predicted 3D path from PathNet $\hat{Y}_{1:N+T}^r$

GTA-IM Dataset

- Existing datasets have relatively noisy 3D human pose annotations and limited long-range human motion
- focus on the task of human pose estimation or parts segmentation with limited interactable objects



GTA-IM Dataset

- Using GTA game engine
- controlling characters, cameras, and action tasks
- Randomizing the goal destination inside the 3D scene, the specific task to do, the walking style; controlling the lighting condition, camera
- In total, one million RGBD frames of 1920×1080 resolution with the ground-truth 3D human pose (98 joints), human segmentation, and camera pose



Datasets Usage

- GTA-IM
 - 8 scenes for training and 2 scenes for evaluation (each scene has several floors)
- PROX
 - Captured using the Kinect-One
 - 12 different 3D scenes and RGB sequences of 20 subjects moving in and interacting with the scenes
 - 52 sequences for training and 8 for testing

Results

GTA-IM

- No prior work that predicts 3D human pose with global movement using 2D pose sequence as input

Metric: Mean Per Joint Position Error (MPJPE)	Time step (second)	3D path error (mm)				3D pose error (mm)				
		0.5	1	1.5	2	0.5	1	1.5	2	All ↓
	TR [54]	277	352	457	603	291	374	489	641	406
	TR [54] + VP [43]	157	240	358	494	174	267	388	526	211
	VP [43] + LTD [63]	124	194	276	367	121	180	249	330	193
Transformer network sequence-to-sequence modeling	Ours (deterministic)	104	163	219	297	91	158	237	328	173
	Ours (samples=4)	114	162	227	310	94	161	236	323	173
performing 3D prediction directly from 2D	Ours (samples=10)	110	154	213	289	90	154	224	306	165
	Ours w/ xyz. output	122	179	252	336	101	177	262	359	191
	Ours w/o image	128	177	242	320	99	179	271	367	196
	Ours w/ masked image	120	168	235	314	96	170	265	358	189
	Ours w/ RGBD input	100	138	193	262	93	160	235	322	172
Treating the entire problem as a single-stage sequence to sequence task	Ours w/ GT destination	104	125	146	170	85	133	178	234	137


Results

GTA-IM

- No prior work that predicts 3D human pose with global movement using 2D pose sequence as input

Time step (second)	3D path error (mm)				3D pose error (mm)				
	0.5	1	1.5	2	0.5	1	1.5	2	All ↓
TR [54]	277	352	457	603	291	374	489	641	406
TR [54] + VP [43]	157	240	358	494	174	267	388	526	211
VP [43] + LTD [63]	124	194	276	367	121	180	249	330	193
Ours (deterministic)	104	163	219	297	91	158	237	328	173
Ours (samples=4)	114	162	227	310	94	161	236	323	173
Ours (samples=10)	110	154	213	289	90	154	224	306	165
Ours w/ xyz. output	122	179	252	336	101	177	262	359	191
Ours w/o image	128	177	242	320	99	179	271	367	196
Ours w/ masked image	120	168	235	314	96	170	265	358	189
Ours w/ RGBD input	100	138	193	262	93	160	235	322	172
Ours w/ GT destination	104	125	146	170	85	133	178	234	137

First predicting future 2D pose using TR
Then lifting to 3D



Results

GTA-IM

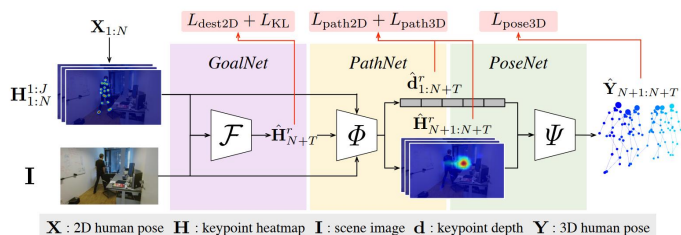
- No prior work that predicts 3D human pose with global movement using 2D pose sequence as input

Time step (second)	3D path error (mm)				3D pose error (mm)				
	0.5	1	1.5	2	0.5	1	1.5	2	All ↓
TR [54]	277	352	457	603	291	374	489	641	406
TR [54] + VP [43]	157	240	358	494	174	267	388	526	211
VP [43] + LTD [63]	124	194	276	367	121	180	249	330	193
Ours (deterministic)	104	163	219	297	91	158	237	328	173
Ours (samples=4)	114	162	227	310	94	161	236	323	173
Ours (samples=10)	110	154	213	289	90	154	224	306	165
Ours w/ xyz. output	122	179	252	336	101	177	262	359	191
Ours w/o image	128	177	242	320	99	179	271	367	196
Ours w/ masked image	120	168	235	314	96	170	265	358	189
Ours w/ RGBD input	100	138	193	262	93	160	235	322	172
Ours w/ GT destination	104	125	146	170	85	133	178	234	137

Combining 2D-to-3D human pose estimation method and 3D human pose prediction method

Results GTA-IM

- No prior work that predicts 3D human pose with global movement using 2D pose sequence as input

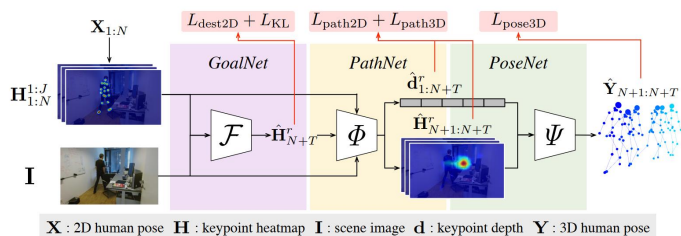


Removing GoalNet
directly using PathNet to produce
deterministic 3D path predictions

Time step (second)	3D path error (mm)				3D pose error (mm)				All ↓
	0.5	1	1.5	2	0.5	1	1.5	2	
TR [54]	277	352	457	603	291	374	489	641	406
TR [54] + VP [43]	157	240	358	494	174	267	388	526	211
VP [43] + LTD [63]	124	194	276	367	121	180	249	330	193
Ours (deterministic)	104	163	219	297	91	158	237	328	173
Ours (samples=4)	114	162	227	310	94	161	236	323	173
Ours (samples=10)	110	154	213	289	90	154	224	306	165
Ours w/ xyz. output	122	179	252	336	101	177	262	359	191
Ours w/o image	128	177	242	320	99	179	271	367	196
Ours w/ masked image	120	168	235	314	96	170	265	358	189
Ours w/ RGBD input	100	138	193	262	93	160	235	322	172
Ours w/ GT destination	104	125	146	170	85	133	178	234	137

Results GTA-IM

- No prior work that predicts 3D human pose with global movement using 2D pose sequence as input



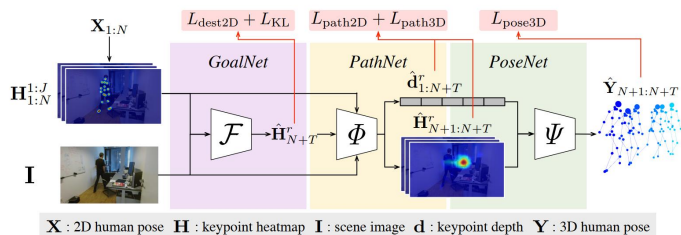
Stochastic mode

Enabling sampling multiple 3D paths during inference selecting the predictions among all samples that best matches ground truth to report the error

Time step (second)	3D path error (mm)				3D pose error (mm)				All ↓
	0.5	1	1.5	2	0.5	1	1.5	2	
TR [54]	277	352	457	603	291	374	489	641	406
TR [54] + VP [43]	157	240	358	494	174	267	388	526	211
VP [43] + LTD [63]	124	194	276	367	121	180	249	330	193
Ours (deterministic)	104	163	219	297	91	158	237	328	173
Ours (samples=4)	114	162	227	310	94	161	236	323	173
Ours (samples=10)	110	154	213	289	90	154	224	306	165
Ours w/ xyz. output	122	179	252	336	101	177	262	359	191
Ours w/o image	128	177	242	320	99	179	271	367	196
Ours w/ masked image	120	168	235	314	96	170	265	358	189
Ours w/ RGBD input	100	138	193	262	93	160	235	322	172
Ours w/ GT destination	104	125	146	170	85	133	178	234	137

Results GTA-IM

- No prior work that predicts 3D human pose with global movement using 2D pose sequence as input

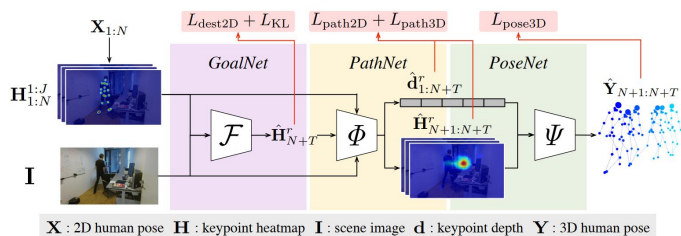


directly regressing 3D coordinates ←
 The 3D path as the depth + 2D heatmap center has better strong correlation to the image appearance

Time step (second)	3D path error (mm)				3D pose error (mm)				All ↓
	0.5	1	1.5	2	0.5	1	1.5	2	
TR [54]	277	352	457	603	291	374	489	641	406
TR [54] + VP [43]	157	240	358	494	174	267	388	526	211
VP [43] + LTD [63]	124	194	276	367	121	180	249	330	193
Ours (deterministic)	104	163	219	297	91	158	237	328	173
Ours (samples=4)	114	162	227	310	94	161	236	323	173
Ours (samples=10)	110	154	213	289	90	154	224	306	165
Ours w/ xyz. output	122	179	252	336	101	177	262	359	191
Ours w/o image	128	177	242	320	99	179	271	367	196
Ours w/ masked image	120	168	235	314	96	170	265	358	189
Ours w/ RGBD input	100	138	193	262	93	160	235	322	172
Ours w/ GT destination	104	125	146	170	85	133	178	234	137

Results GTA-IM

- No prior work that predicts 3D human pose with global movement using 2D pose sequence as input



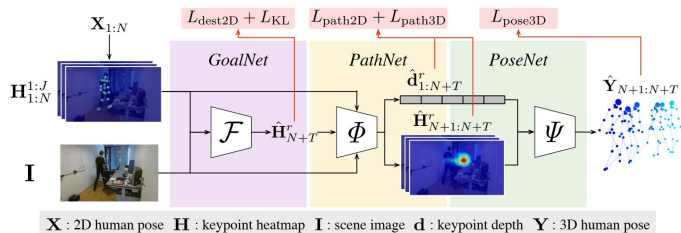
replacing pixels outside human crop by ImageNet mean pixel values



Time step (second)	3D path error (mm)				3D pose error (mm)				
	0.5	1	1.5	2	0.5	1	1.5	2	All ↓
TR [54]	277	352	457	603	291	374	489	641	406
TR [54] + VP [43]	157	240	358	494	174	267	388	526	211
VP [43] + LTD [63]	124	194	276	367	121	180	249	330	193
Ours (deterministic)	104	163	219	297	91	158	237	328	173
Ours (samples=4)	114	162	227	310	94	161	236	323	173
Ours (samples=10)	110	154	213	289	90	154	224	306	165
Ours w/ xyz. output	122	179	252	336	101	177	262	359	191
Ours w/o image	128	177	242	320	99	179	271	367	196
Ours w/ masked image	120	168	235	314	96	170	265	358	189
Ours w/ RGBD input	100	138	193	262	93	160	235	322	172
Ours w/ GT destination	104	125	146	170	85	133	178	234	137

Results GTA-IM

- No prior work that predicts 3D human pose with global movement using 2D pose sequence as input



using ground-truth 2D destinations instead of predicted ones

The most difficult part → finding the goal



Time step (second)	3D path error (mm)				3D pose error (mm)				
	0.5	1	1.5	2	0.5	1	1.5	2	All ↓
TR [54]	277	352	457	603	291	374	489	641	406
TR [54] + VP [43]	157	240	358	494	174	267	388	526	211
VP [43] + LTD [63]	124	194	276	367	121	180	249	330	193
Ours (deterministic)	104	163	219	297	91	158	237	328	173
Ours (samples=4)	114	162	227	310	94	161	236	323	173
Ours (samples=10)	110	154	213	289	90	154	224	306	165
Ours w/ xyz. output	122	179	252	336	101	177	262	359	191
Ours w/o image	128	177	242	320	99	179	271	367	196
Ours w/ masked image	120	168	235	314	96	170	265	358	189
Ours w/ RGBD input	100	138	193	262	93	160	235	322	172
Ours w/ GT destination	104	125	146	170	85	133	178	234	137

Results

PROX

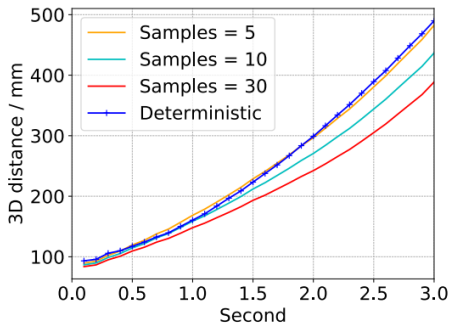
Time step (second)	3D path error (mm)				3D pose error (mm)				
	0.5	1	1.5	2	0.5	1	1.5	2	All ↓
TR [54]	487	583	682	783	512	603	698	801	615
TR [54] + VP [43]	262	358	461	548	297	398	502	590	326
VP [43] + LTD [63]	194	263	332	394	216	274	335	394	282
Ours w/o GTA-IM pretrain	192	258	336	419	192	273	352	426	280
Ours (deterministic)	189	245	317	389	190	264	335	406	270
Ours (samples=3)	192	245	311	398	187	258	328	397	264
Ours (samples=6)	185	229	285	368	184	249	312	377	254
Ours (samples=10)	181	222	273	354	182	244	304	367	249
Ours w/ gt destination	193	223	234	237	195	235	276	321	237

Uncertainty of future motion in the real dataset is larger.
Stochastic predictions have more advantage

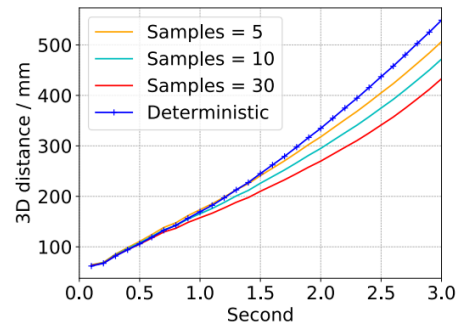
Results

GTA-IM

stochastic model can
achieve better results with a small
number of samples especially in
the long-term prediction

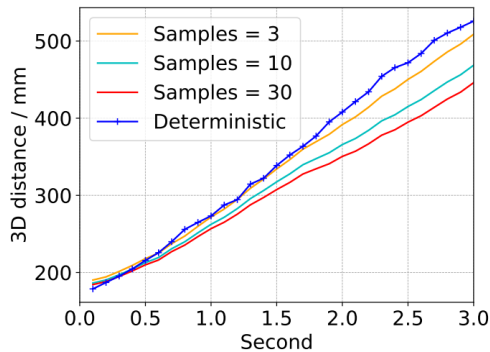


(a) predicted 3D paths

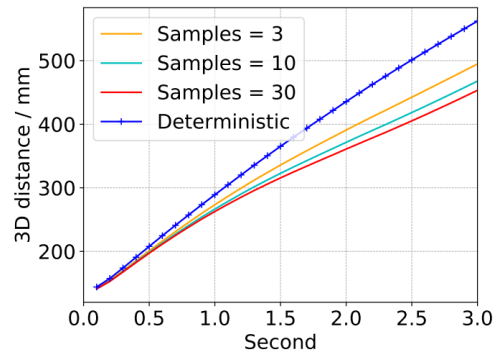


(b) predicted 3D poses

PROX



(a) predicted 3D paths



(b) predicted 3D poses

Results

Qualitative Results

3D Path Prediction on PROX



History of human center position

Predicted 3D path

Results

Qualitative Results

3D Path Prediction on GTA-IM



History of human center position

Predicted 3D path

Results

Qualitative Results

3D Pose Prediction

Example 1



Input

Thank you for your attention
