A. Supplementary material

A.1. About Distortion Parameters

Our work targets two main types of in-the-wild situations. The first is the distortion that occurs in cameras for special purposes, such as fisheye and wide-angle cameras (*e.g.*, insta360), and the second is the distortion that occurs in low-cost cameras such as surveillance cameras. We defined the former and latter as "heavy" and "moderate" (equivalent to "light") distortion, respectively. Since there is no common benchmark and public dataset with such level of distortions, we randomly selected two sets of parameters (*i.e.*, k_1, k_2, k_3, p_1, p_2) that well reflect real-world situations at each distortion, and we synthesized videos and used them for evaluation. The d_1, d_2, d_3, and d_4 have distortion parameters of ($\mp$4.142, $\pm$4.956, $\mp$0.062, -0.488, -0.712), ($\mp$2.071, $\pm$2.478, $\mp$0.031, -0.010, -0.014), respectively. The original distortion present in H3.6M is (-0.207, 0.248, -0.003, -0.001, -0.001), which is almost identical to no distortion.

A.2. Bone-Length based ISO

Given a video clip with frame length of T, predicted 3D joints $\tilde{\mathbf{S}} = \{\tilde{\mathbf{s}}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times J\times 3}$ where $\tilde{\mathbf{s}}_t \in \mathbb{R}^{J\times 3}$ represents the predicted 3D joints at frame t can be obtained. Then, we can get the predicted bone-lengths (denoted as $\tilde{\mathbf{l}} = \{\tilde{l}_{t,j}\}_{t=1}^{T} \in \mathbb{R}^{T\times(J-1)}$ where $\tilde{l}_{t,j}$ denotes the predicted length of jth bone at frame t) from the predicted 3D joints by calculating the distance between adjacent joints. Finally, we can calculated the *bone-length symmetry* loss as follows:

$$\mathcal{L}_{\text{symmetry}} = \sum_{t=1}^{T} \sum_{(j_l,j_r)\in\mathcal{P}} \left| \tilde{l}_{t,j_l} - \tilde{l}_{t,j_r} \right|, \tag{8}$$

where $\mathcal{P}$ contains all the pair of bones that are symmetrical to the left and right (denoted as j_l and j_r, respectively). Also, the *bone-length consistency* loss is obtained by:

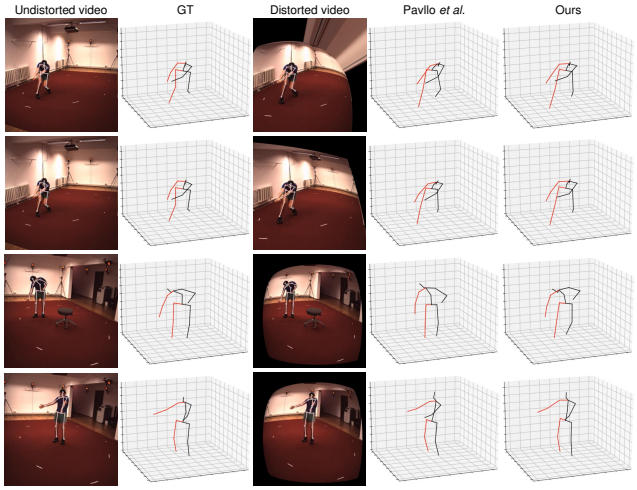
$$\mathcal{L}_{\text{consistency}} = \sum_{t=1}^{T-1} \sum_{j=1}^{J-1} \left| \tilde{l}_{t+1,j} - \tilde{l}_{t,j} \right|. \tag{9}$$

Thus, our final objective for the Inference Stage Optimization in *Scenario 2* is as follows:

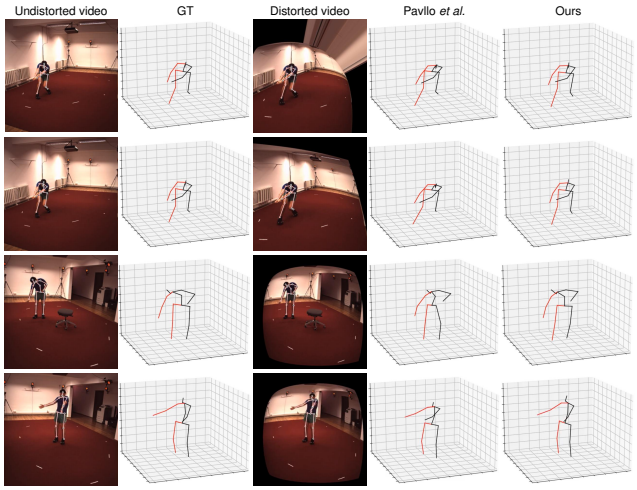
$$\mathcal{L}_{\text{ISO}} = \mathcal{L}_{\text{symmetry}} + \mathcal{L}_{\text{consistency}}. \tag{10}$$

A.3. Quantitative Results

In Table 2, we provided the average performance on each of the heavy distortion and moderate distortion. Table 6 shows the performance at each distortion (*i.e.*, d_1, d_2, d_3, and d_4). In addition, Table 7 shows the reconstruction accuracy (PCKh@0.5) for each action. The reported accuracy here are the average value for all kinds of distortions. We can notice that our method outperforms other methods regardless of the kinds of distortions and actions.



(a) Qualitative results under the *Scenario 1* setting.



(b) Qualitative results under the *Scenario 2* setting.

Figure 8: Qualitative results on heavily distorted videos of Human3.6M under the *Scenario 1* and *Scenario 2* setting.

A.4. Qualitative Results

Figure 8 shows qualitative results from videos with more diverse poses and distortions. We can notice that our method adapts better to the distorted environments than our base model [21], showing more similar results to the ground-truth 3D pose.

A.5. Performance in Undistorted Environments

Since our model is trained to be sensitive to all kinds of distortions, it performs well even in undistorted environments. Our method shows an MPJPE of 50.6mm in the test environment with no distortion. This is 2.1mm higher than the base model [21], but it is a reasonable trade-off because it has great advantages in other situations with distortions.

Method	*Scenario 1*			*Scenario 2*		
	MPJPE(↓)	P-MPJPE(↓)	PCKh@0.5(↑)	MPJPE(↓)	P-MPJPE(↓)	PCKh@0.5(↑)
Martinez *et al*. [17] ICCV'17	81.1 / 75.5	59.6 / 56.6	65.4 / 67.8	92.8 / 163.2	65.2 / 108.3	58.3 / 36.2
Zhao *et al*. [36] CVPR'19	90.7 / 81.8	66.0 / 62.3	61.3 / 65.1	104.6 / 134.7	76.0 / 94.9	52.6 / 37.3
Pavllo *et al*. [21] CVPR'19	83.2 / 76.6	61.4 / 57.3	65.5 / 69.1	94.4 / 133.8	65.6 / 79.2	57.5 / 38.2
Chen *et al*. [4] TCSVT'21	91.4 / 87.3	63.0 / 60.7	59.4 / 58.9	96.7 / 117.9	65.9 / 76.0	57.4 / 40.6
Liu *et al*. [16] CVPR'20	84.2 / 78.8	63.0 / 58.8	64.8 / 67.9	93.3 / 128.0	68.0 / 86.9	58.8 / 40.2
Ours	**64.1 / 59.8**	**48.0 / 44.7**	**77.3 / 79.5**	**69.1 / 63.1**	**49.7 / 45.9**	**74.7 / 77.8**

(a) Comparison of performance on (distortion d_1) / (distortion d_2).

Method	*Scenario 1*			*Scenario 2*		
	MPJPE(↓)	P-MPJPE(↓)	PCKh@0.5(↑)	MPJPE(↓)	P-MPJPE(↓)	PCKh@0.5(↑)
Martinez *et al*. [17] ICCV'17	63.8 / 62.3	49.1 / 48.2	75.9 / 77.0	75.0 / 61.6	51.7 / 46.5	69.2 / 79.0
Zhao *et al*. [36] CVPR'19	66.6 / 61.4	48.7 / 46.1	74.9 / 78.9	80.4 / 62.3	56.5 / 47.2	65.9 / 78.5
Pavllo *et al*. [21] CVPR'19	65.2 / 64.7	48.6 / 48.0	76.5 / 76.9	74.8 / 54.2	50.7 / 40.6	69.4 / 83.8
Chen *et al*. [4] TCSVT'21	64.6 / 60.7	47.9 / 44.7	76.3 / 79.3	77.1 / 53.1	52.9 / 39.6	69.3 / 85.2
Liu *et al*. [16] CVPR'20	69.2 / 68.3	51.3 / 50.6	74.4 / 74.9	72.3 / 55.6	50.7 / 42.3	70.4 / 83.2
Ours	**53.8 / 53.4**	**40.8 / 40.4**	**83.1 / 83.4**	**51.8 / 51.4**	**39.5 / 38.9**	**85.6 / 85.8**

(b) Comparison of performance on (distortion d_3) / (distortion d_4).

Table 6: Comparison with other state-of-the-art models on Human3.6M. The top two rows [17, 36] are based on a single-frame and others [21, 4, 16], including our method, are based on video with a frame length of 27. Best in bold, second-best underlined.

	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Martinez *et al*. [17] ICCV'17	80.5	79.9	70.2	74.3	69.7	58.3	74.7	79.5	67.6	56.4	72.2	73.5	65.4	78.8	72.5	71.6
Zhao *et al*. [36] CVPR'19	75.3	73.1	69.8	73.1	70.9	57.4	70.6	76.5	72.1	59.7	71.6	70.7	65.8	73.9	70.3	70.1
Pavllo *et al*. [21] CVPR'19	82.0	80.1	76.0	76.1	71.7	60.5	75.5	81.1	58.0	44.0	71.3	71.4	70.4	82.7	78.8	72.0
Chen *et al*. [4] TCSVT'21	73.9	74.9	66.9	70.7	67.0	59.9	70.5	72.6	66.7	56.2	69.8	70.0	64.1	73.6	70.1	68.5
Liu *et al*. [16] CVPR'20	80.6	78.3	72.5	74.1	70.5	59.7	74.9	79.9	52.9	41.0	69.7	71.3	70.5	82.6	79.0	70.5
Ours	**85.8**	**83.5**	**80.1**	**84.8**	**81.6**	**70.5**	**80.8**	**85.8**	**78.6**	**57.7**	**83.3**	**80.7**	**77.1**	**93.0**	**89.3**	**80.8**

(a) Reconstruction accuracy (PCKh@0.5) under the *Scenario 1* setting.

	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Martinez *et al*. [17] ICCV'17	59.2	61.0	62.0	57.1	62.2	48.9	58.3	67.6	69.6	65.5	63.1	57.9	60.4	59.9	57.2	60.7
Zhao *et al*. [36] CVPR'19	57.3	57.7	59.2	55.4	61.2	47.1	56.7	63.2	65.2	62.9	61.6	56.0	59.0	59.4	56.5	58.6
Pavllo *et al*. [21] CVPR'19	58.0	57.6	63.9	58.1	66.0	52.7	56.8	70.1	72.4	67.8	65.9	57.7	63.7	62.9	60.0	62.2
Chen *et al*. [4] TCSVT'21	57.0	57.9	66.1	59.6	66.7	55.4	56.3	70.5	72.7	70.0	66.8	58.8	63.6	64.3	61.3	63.1
Liu *et al*. [16] CVPR'20	62.2	62.2	63.7	58.9	65.6	51.7	58.9	70.7	70.1	66.8	65.9	59.9	64.0	64.6	61.9	63.1
Ours	**82.0**	**82.0**	**74.4**	**76.7**	**82.1**	**63.6**	**80.1**	**82.4**	**80.5**	**66.7**	**80.8**	**77.3**	**75.0**	**86.0**	**81.7**	**78.1**

(b) Reconstruction accuracy (PCKh@0.5) under the *Scenario 2* setting.

Table 7: Comparison with other state-of-the-art models on Human3.6M. The top two rows [17, 36] are based on a single-frame and others [21, 4, 16], including our method, are based on video with a frame length of 27. The reported performance is the average value for all kinds of distortions. Higher is better, best in bold, second-best underlined.