Camera Distortion-aware 3D Human Pose Estimation in Video with Optimization-based Meta-Learning

Hanbyel Cho, Yoonshin Cho, Jaemyung Yu, Junmo Kim

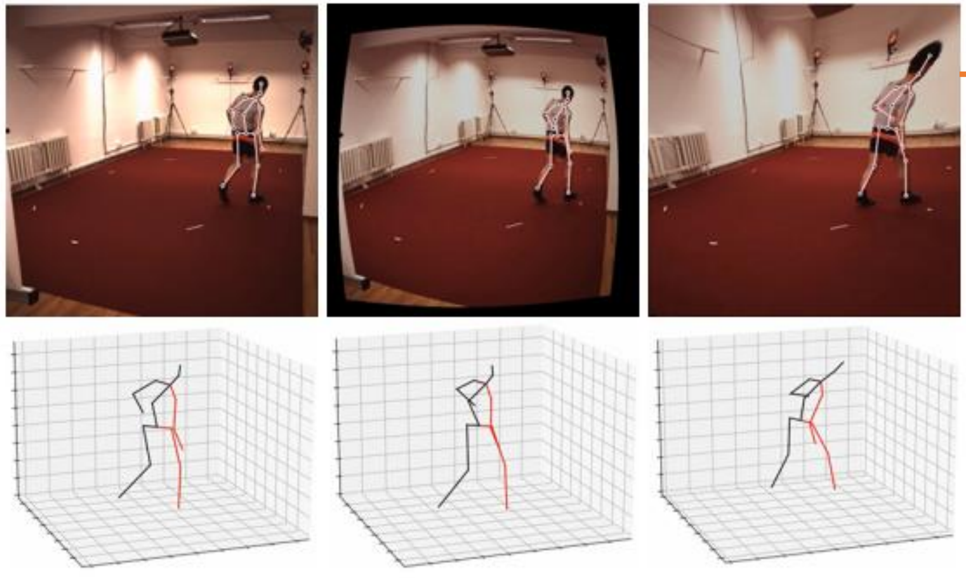
(ICCV 2021)

03.30.2022

전북대학교 학부생 김세희

Introduction

- Existing 3D human pose estimation algorithms trained on distortion-free datasets suffer performance drop when applied to new scenarios with a specific camera distortion.
- Propose a simple yet effective model for 3D human pose estimation in video that can quickly adapt to any distortion environment by utilizing MAML, a representative optimization-based meta-learning algorithm.



(a) Undistorted (b) Distortion 1 (c) Distortion 2

Human3.6M

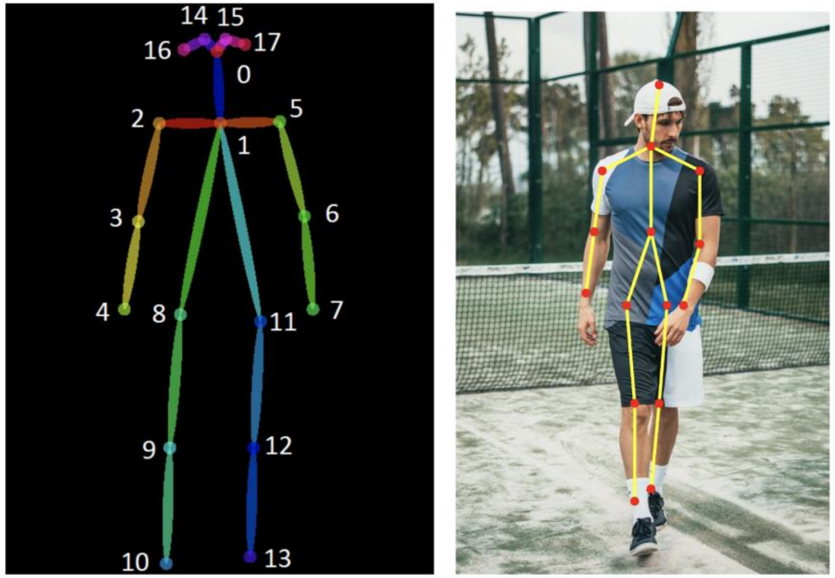
Condition	MPJPE(↓)	P-MPJPE(↓)	PCKh@0.5(↑)
Undistorted	48.5	37.1	87.1
Distortion 1	94.4(+45.9)	65.6(+28.5)	57.7(-29.4)
Distortion 2	133.8(+85.3)	79.2(+42.1)	38.2(-48.9)

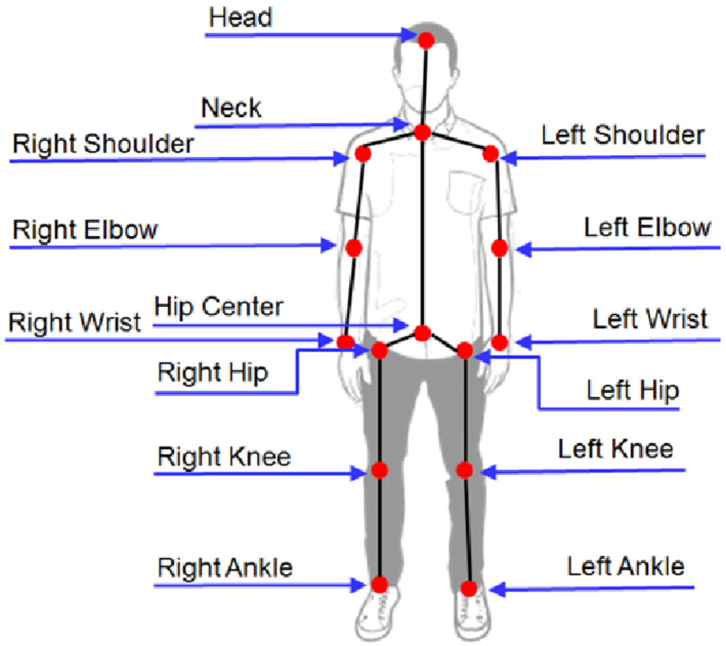
Preliminary

Background

- What is Human Pose Estimation?

  : a way of identifying and classifying the joints in the human body.



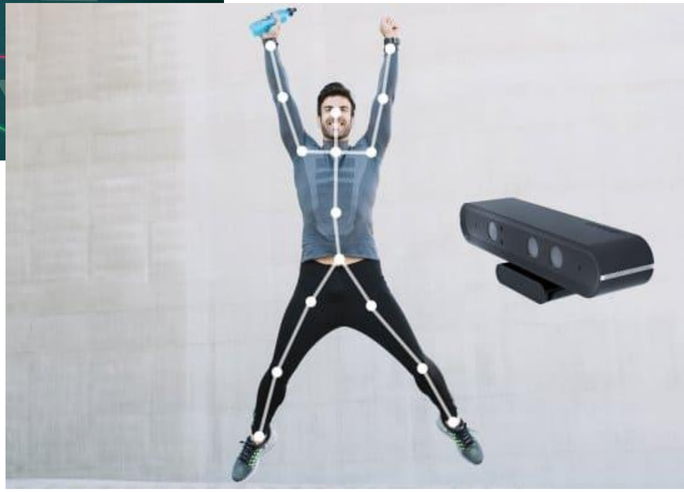


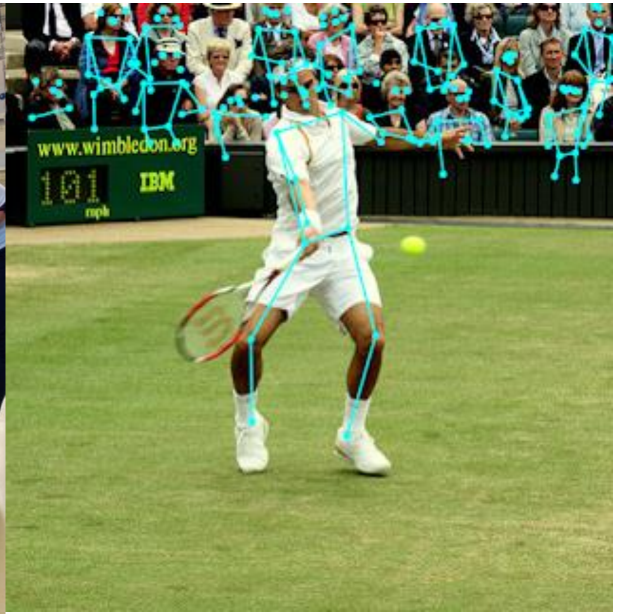
Preliminary

Background

- Human Pose Estimation Applications

  : Autonomous driving, Robotics, Game, Sports etc.

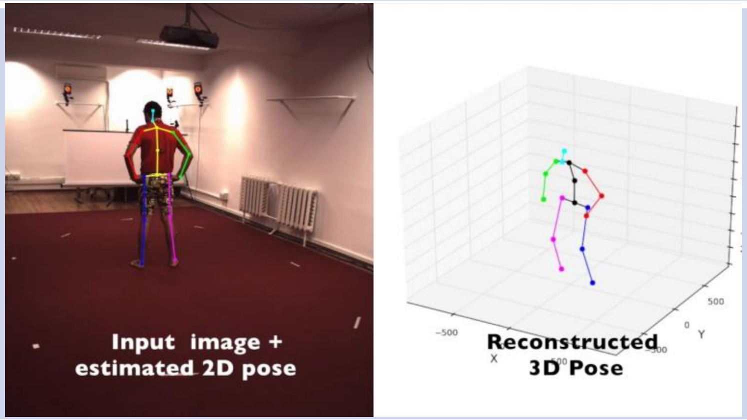
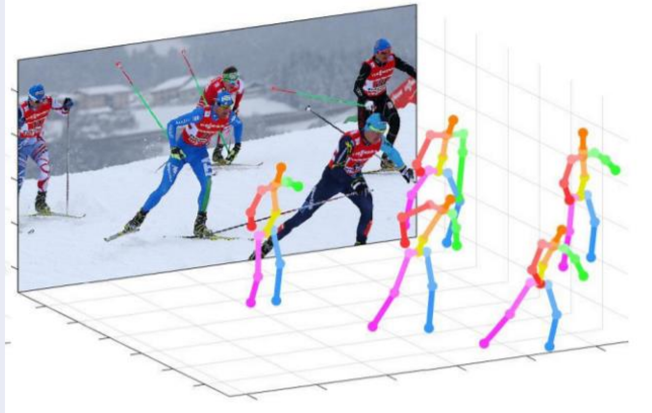




Preliminary

Background

- Techniques of Human Pose Estimation

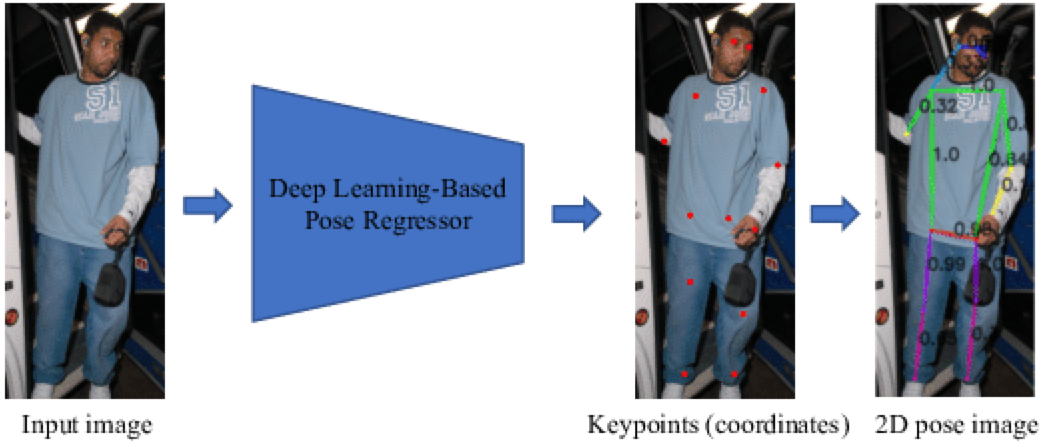
	2D Pose	3D Pose
Single-Person	Input : image 	Input : image, depth 
Multi-Person	Input : image 	Input : image, depth 

Preliminary

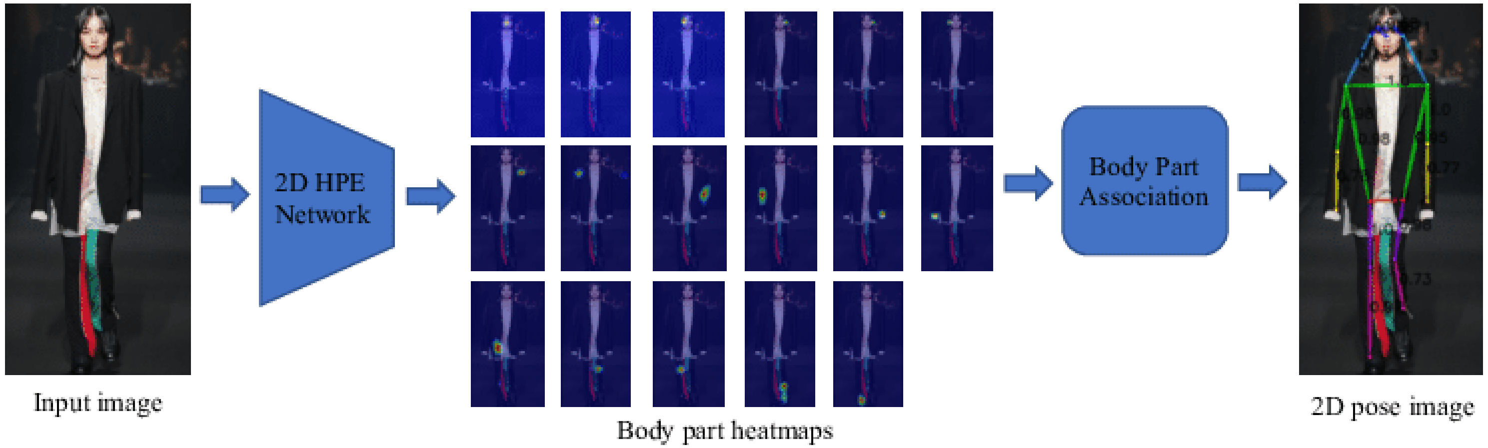
Background

- Approaches of **single-person**

Direct regression



Heatmap based



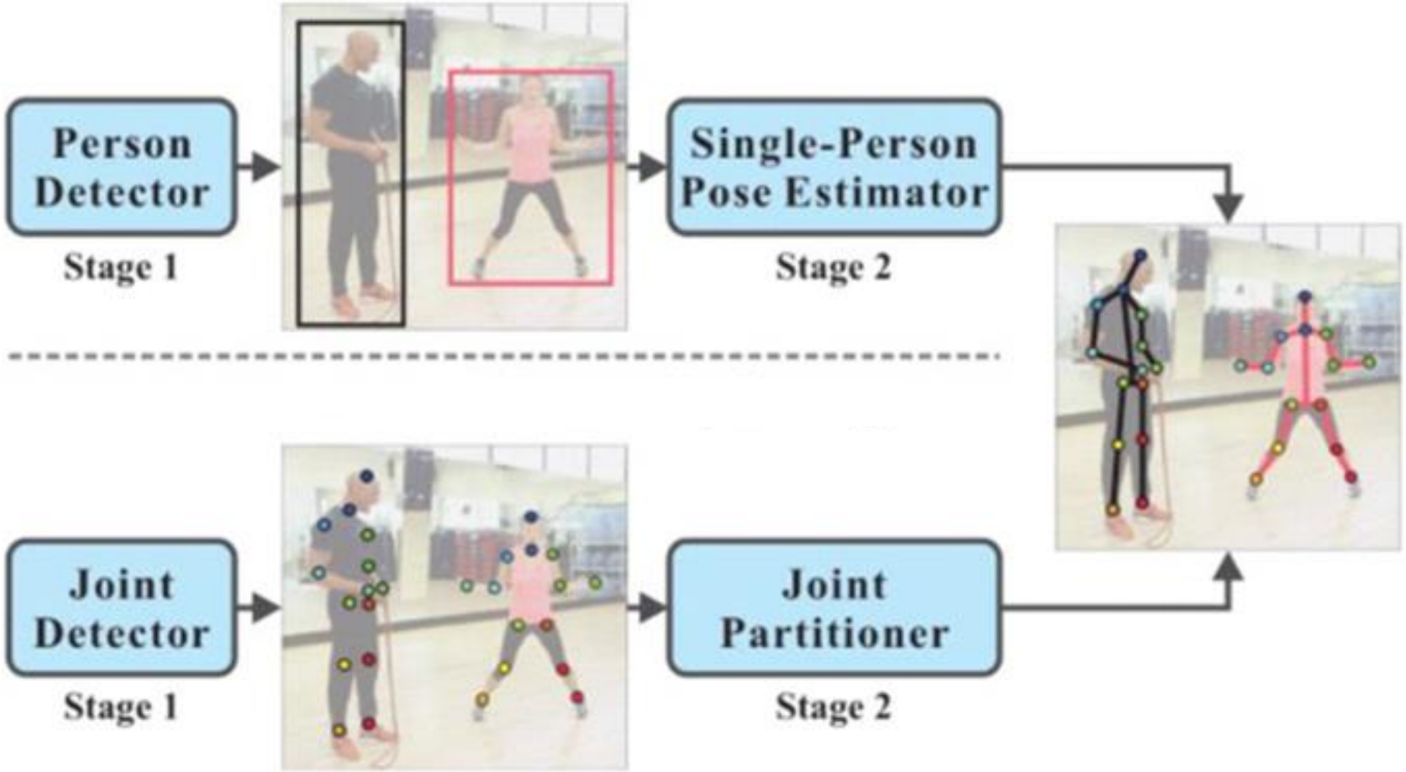
Preliminary

Background

- Approaches of **multi-person**

Top-down

Bottom-up



Preliminary

Background

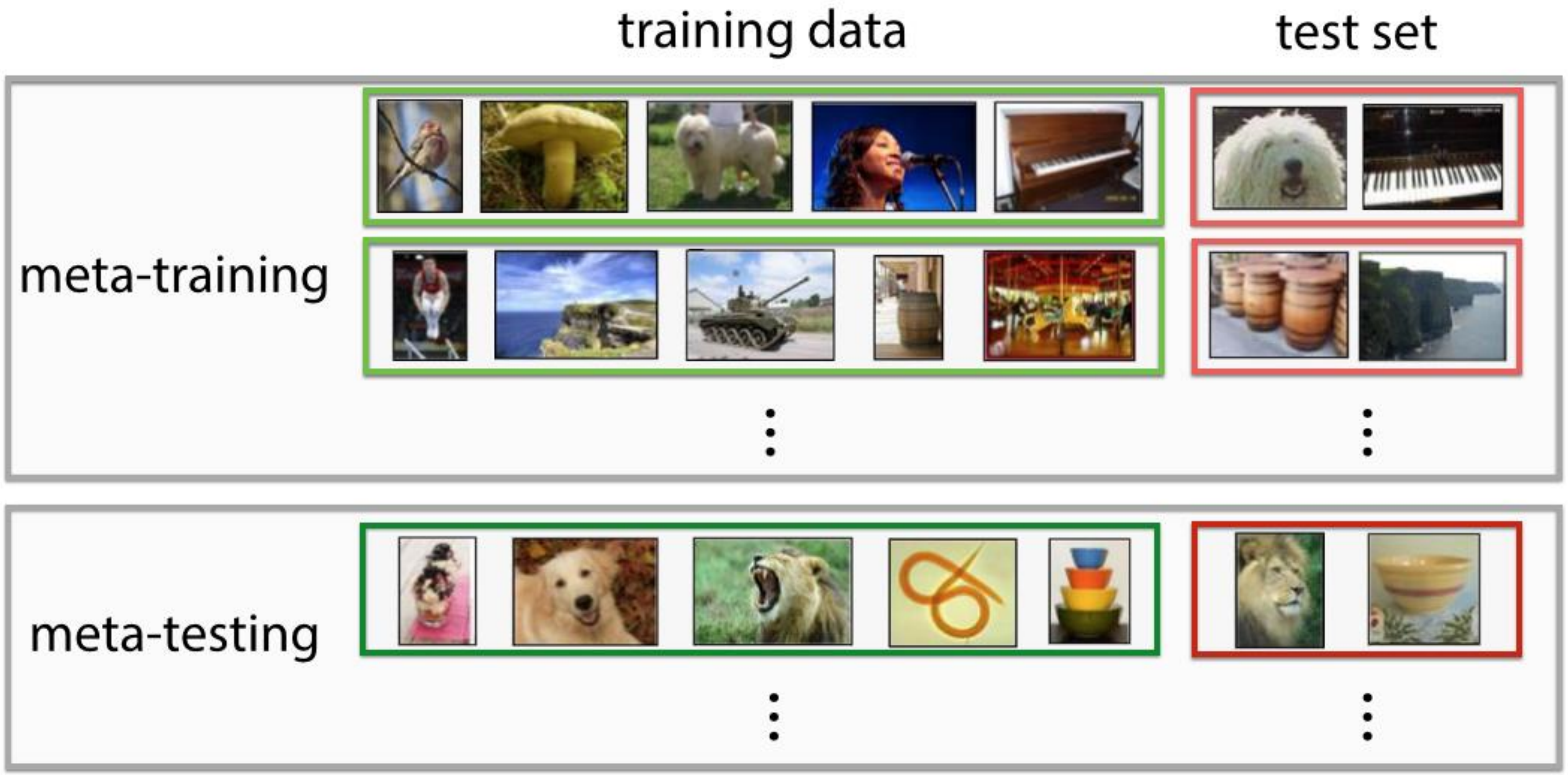
- What is Meta-Learning?
  - If you’ve learned 100 tasks already, can you figure out how to *learn* more efficiently?
  - Meta-learning = *learning to learn*

- Why is meta-Learning a good idea?
  - Deep learning algorithms require a huge number of data.
  - If we can meta-learn a learner, we can learn new tasks efficiently.

Preliminary

Background

- Meta-Learning with supervised learning



Preliminary

Background

- MAML (Model-Agnostic Meta-Learning)

  Chelsea Finn, Pieter Abbeel, Sergey Levine:
  Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks (5679 quotes)

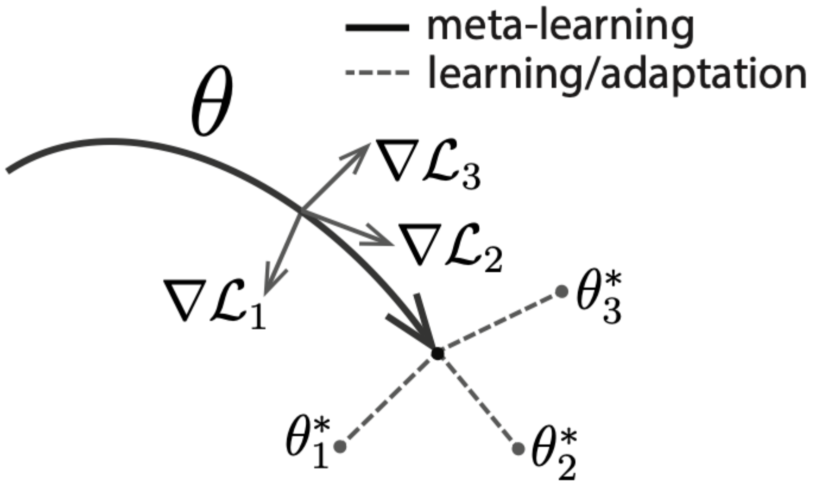


Figure 1. Diagram of our model-agnostic meta-learning algorithm (MAML), which optimizes for a representation θ that can quickly adapt to new tasks.

For each task T_i

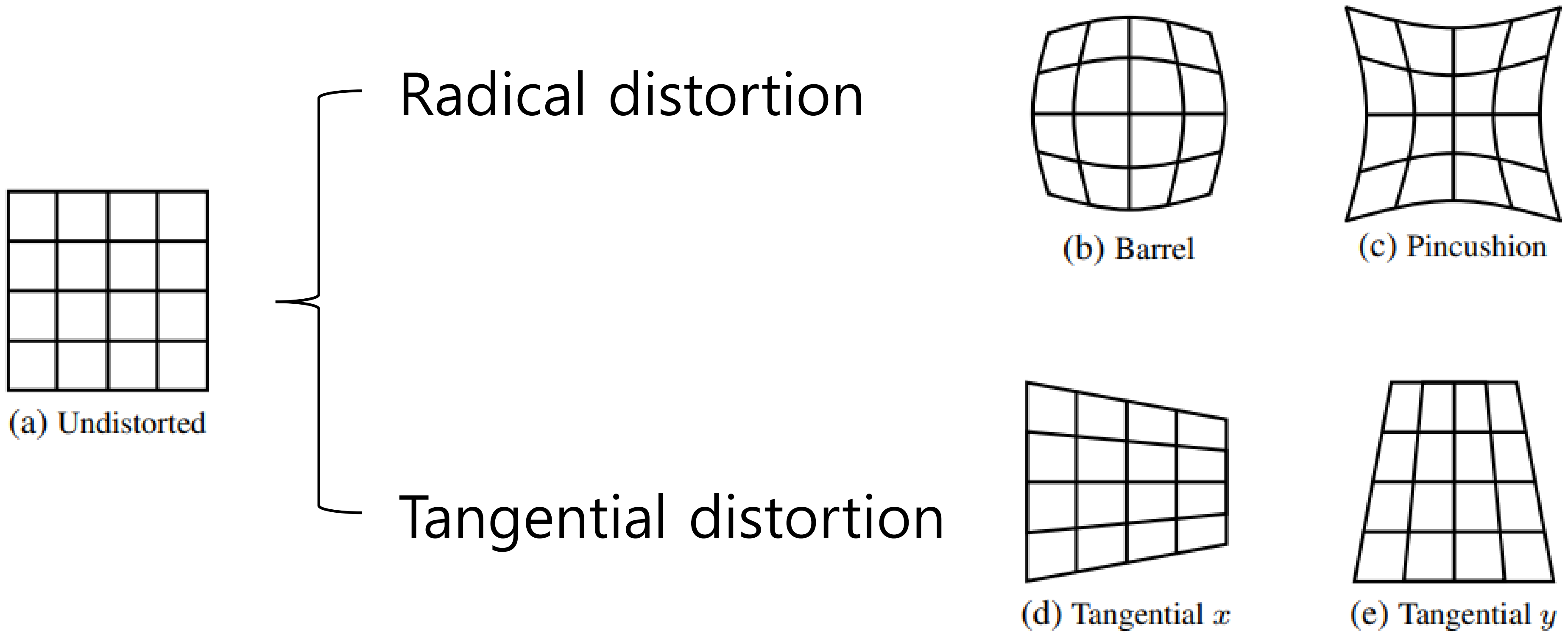
$$\theta_i' = \theta - \alpha\nabla_\theta\mathcal{L}_{T_i}(\theta, D_i)$$

Preliminary

Background

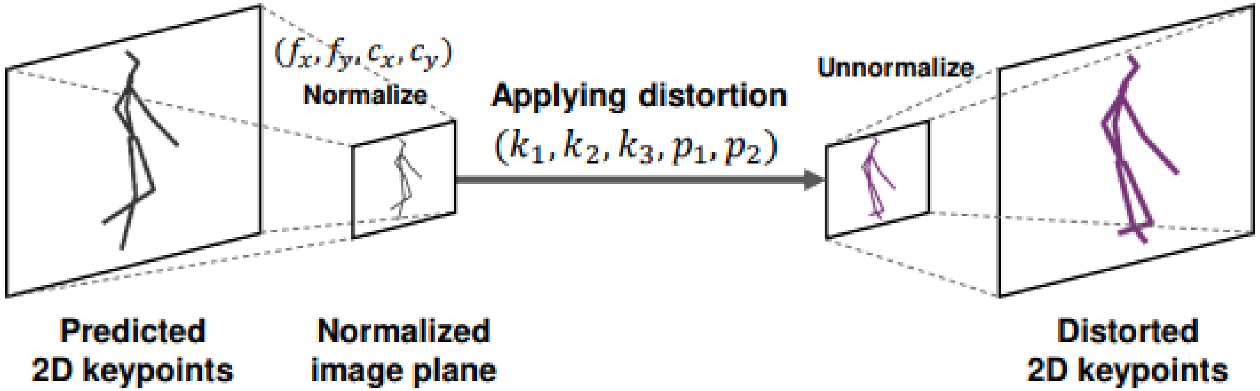
- Camera Distortion

  - There are two kinds of camera distortion.

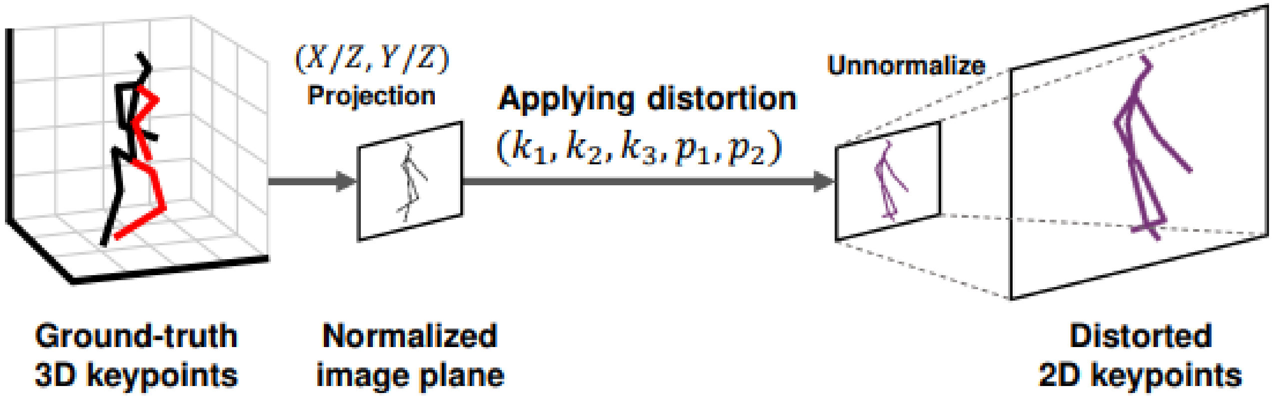


Method

Synthetic Distorted Task Generation



(a) Generating distorted 2D keypoints from predicted ones.



(b) Generating distorted 2D keypoints from 3D ground-truth.

Method

Overall Framework

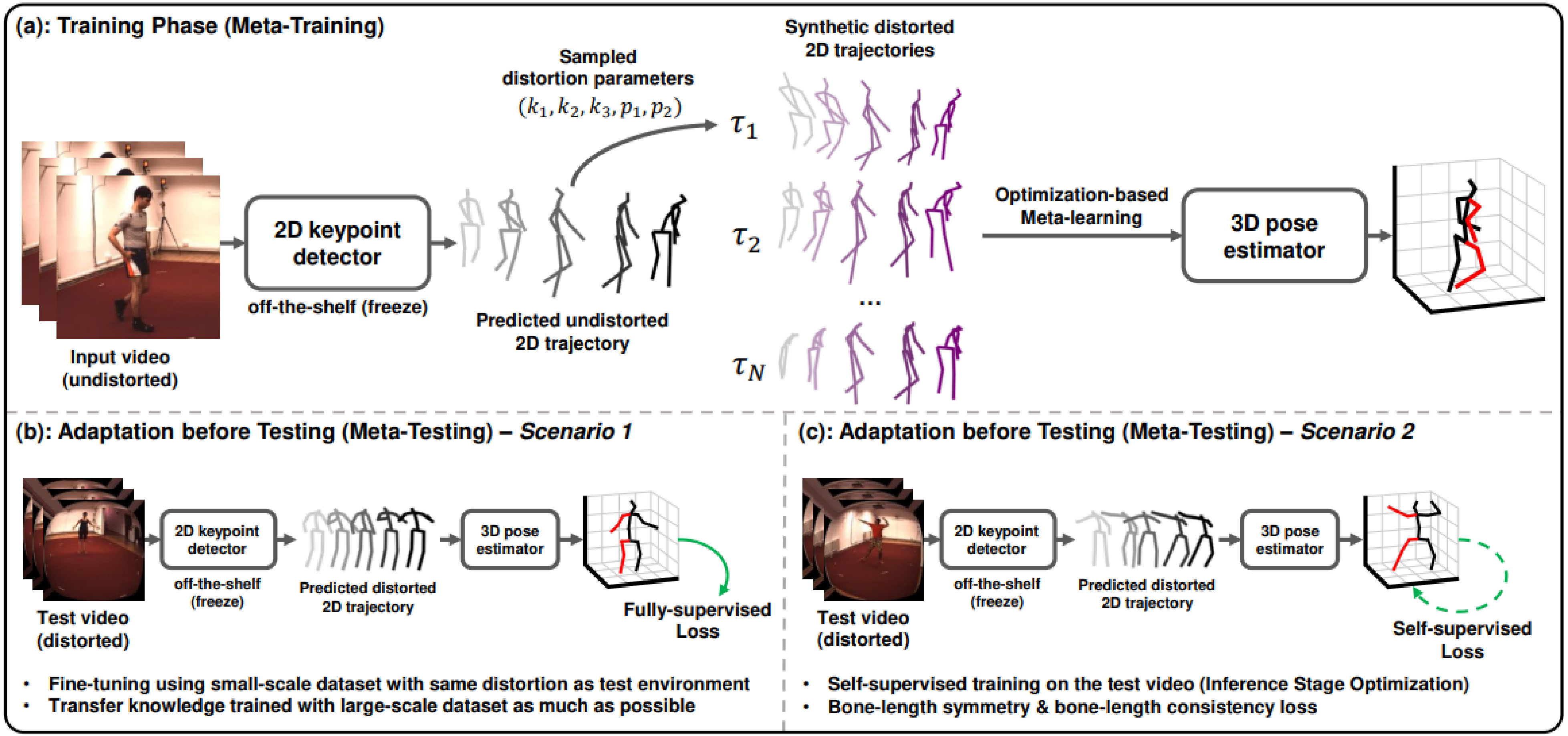


Figure 3: Overall framework of our methods. (a) We train a 2D-keypoint-conditioned 3D pose estimator that can quickly adapt to any distortions using only an undistorted large-scale dataset. Before the trained network can be used in practice, it must be adapted to a certain distortion. (b) and (c) represent adaptation method for *Scenario 1* and *Scenario 2*, respectively.

Method

Algorithm

Algorithm 1: Training Phase

Input: $\mathcal{D}$: a large-scale 3D human pose dataset
Input: α, β: learning rate hyperparameters
Output: Model parameters θ

Randomly initialize θ
while *not done* **do**
    Sample batch of tasks $\mathcal{T}_{rand,i} \sim p_{rand}(\mathcal{T})$
    for *all* $\mathcal{T}_{rand,i}$ **do**
        Calculate loss by MPJPE: $\mathcal{L}_{\mathcal{T}_{rand,i}}(g_\theta)$
        Compute updated parameters:
        $\theta = \theta - \beta \nabla_\theta \mathcal{L}_{\mathcal{T}_{rand,i}}(g_\theta)$
    end
end
while *not done* **do**
    Sample batch of tasks $\mathcal{T}_{strat,i} \sim p_{strat}(\mathcal{T})$
    for *all* $\mathcal{T}_{strat,i}$ **do**
        Calculate loss by MPJPE: $\mathcal{L}_{\mathcal{T}_{strat,i}}(g_\theta)$
        Compute updated parameters:
        $\theta'_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_{strat,i}}(g_\theta)$
    end
    Update θ with respect to average test loss:
    $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_{rand,i} \sim p_{rand}(\mathcal{T})} \mathcal{L}_{\mathcal{T}_{rand,i}}(g_{\theta'_i})$
end

random distortion pretraining

Task-level training

Task-level testing

Method

A k_1 of ith sample in the meta-batch is sampled as follow :

- $k_1, k_2, k_3 \sim u[-\lambda_1, \lambda_1]$: parameters related to radial distortion.
- $p_1, p_2, p_3 \sim u[-\lambda_2, \lambda_2]$: parameters related to tangential distortion.
- λ_1, λ_2 : the maximum value of each distribution.

$$k_{1,i} \sim -\lambda_1 + 2 \cdot \lambda_1 \cdot u\left[\frac{\mathrm{i}-1}{N}, \frac{\mathrm{i}}{N}\right]$$

Perform only one gradient descent update when the parameters θ is adapted to a new task T_i. θ_i' are obtained by :

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{T_i}(g\theta)$$

Method

The mata-objective is expressed as follows :

$$\arg\min \sum_{T_i \sim p(T)} \mathcal{L}_{T_i}(g\theta_i')$$

$$= \arg\min \sum_{T_i \sim p(T)} \mathcal{L}_{T_i}\left(g\theta - \alpha \nabla_\theta \mathcal{L}_{T_i}(g\theta)\right)$$

For the stochastic gradient descent, model parameters θ are updated as follows :

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_i \sim p(T)} \mathcal{L}_{T_i}(g\theta_i')$$

Evaluation Metrics

- MPJPE (mean per joint position error)

  : the L2 distance between ground-truth 3D joints and predicted ones

$$\mathrm{MPJPE} = \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\|(J_i^{(t)} - J_{root}^{(t)}) - (\hat{J}_i^{(t)} - \hat{J}_{root}^{(t)})\|_2$$

- P-MPJPE : calculates the error between the joints after alignment using Procrustes Analysis

- PCkh@0.5 : percentage of correct 3D joints with a threshold as 50% of the head segment length

Results

(a) B+T (d_1) (b) P+T (d_2) (c) B+T (d_3) (d) P+T (d_4)

Comparison with State-of-the-Art

- The proposed method outperforms other methods regardless of the kinds of distortions and scenarios.

	Scenario 1			*Scenario 2*		
Method	MPJPE(↓)	P-MPJPE(↓)	PCKh@0.5(↑)	MPJPE(↓)	P-MPJPE(↓)	PCKh@0.5(↑)
Martinez *et al.* [17] ICCV'17	78.3 / 63.1	58.1 / 48.7	66.6 / 76.5	128.0 / 68.3	86.8 / 49.1	47.3 / 74.1
Zhao *et al.* [36] CVPR'19	86.3 / 64.0	64.2 / 47.4	63.2 / 76.9	119.7 / 71.4	85.5 / 51.9	45.0 / 72.2
Pavllo *et al.* [21] CVPR'19	79.9 / 65.0	59.4 / 48.3	67.3 / 76.7	114.1 / 64.5	72.4 / 45.7	47.9 / 76.6
Chen *et al.* [4] TCSVT'21	89.4 / 62.7	61.9 / 46.3	59.2 / 77.8	107.3 / 65.1	71.0 / 46.3	49.0 / 77.3
Liu *et al.* [16] CVPR'20	81.5 / 68.8	60.9 / 51.0	66.4 / 74.7	110.7 / 64.0	77.5 / 46.5	49.5 / 76.8
Ours	**62.0 / 53.6**	**46.4 / 40.6**	**78.4 / 83.3**	**66.1 / 51.6**	**47.8 / 39.2**	**76.3 / 85.7**

Table 2: Comparison of average performance on (heavy) / (moderate) with other state-of-the-art models. The top two rows [17, 36] are based on a single-frame and others [21, 4, 16], including our method, are based on a video with a frame length of 27. Best in bold, second-best underlined. More results can be seen in the supplementary material (Appendix A.3).

Results

Comparison with State-of-the-Art

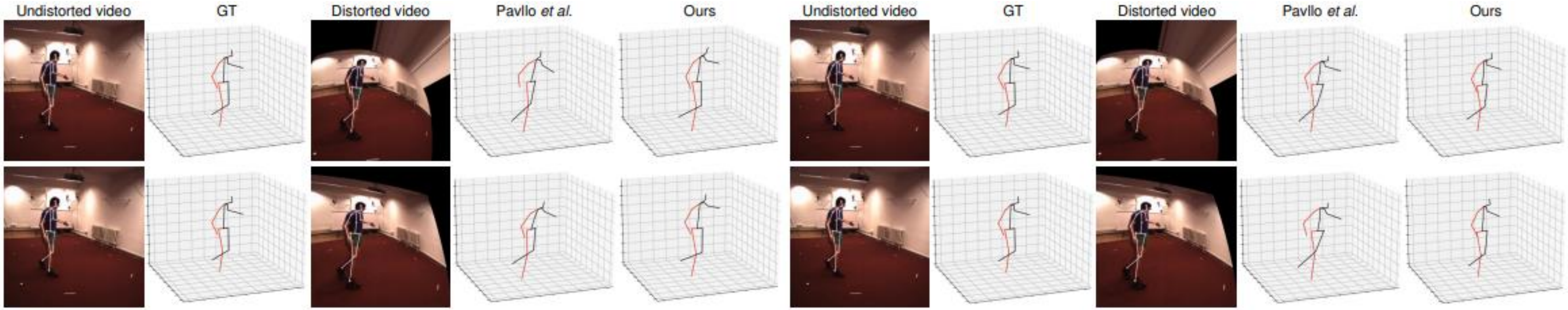


Figure 6: Qualitative results on heavily distorted videos of Human3.6M. The five columns from the leftmost are the result under the *Scenario 1* setting, while the rest columns are the result under the *Scenario2* setting. **Top row:** 3D reconstruction results on d_1. **Bottom row:** 3D reconstruction results on d_2. More results can be seen in Appendix A.4.

Results

Ablation Studies

- Notice that each method provides a positive contribution under all metrics

	MPJPE(↓)	P-MPJPE(↓)	PCKh@0.5(↑)
base model [21]	84.2 / 79.6	62.8 / 59.7	64.8 / 66.9
+ MAML (with synthetic tasks)	73.5 / 67.5	55.6 / 51.7	72.0 / 74.5
+ *stratified sampling*	71.7 / 66.2	54.3 / 50.4	72.8 / 75.2
+ *random distortion pretraining*	67.2 / 61.9	51.0 / 47.0	75.7 / 78.2

Table 3: Effectiveness of each proposed method based on input frame length of 9 under *Scenario 1* setting. Each value denotes performance on (distortion d_1) / (distortion d_2).

Results

Ablation Studies

- Notice that the former method shows better performance under all metrics and scenarios since there is less domain gap between training and testing.

Method	MPJPE(↓)	P-MPJPE(↓)	PCKh@0.5(↑)
Predicted 2D keypoints	62.0 / 53.6	46.4 / 40.6	78.4 / 83.3
Ground-truth 3D joints	64.7 / 56.1	48.2 / 42.0	77.0 / 82.0
Predicted 2D keypoints	66.1 / 51.6	47.8 / 39.2	76.3 / 85.7
Ground-truth 3D joints	71.3 / 55.6	51.9 / 42.6	72.8 / 83.5

Table 4: Comparison of average performance on (heavy) / (moderate) between the methods generating synthetic 2D keypoints. **Top rows:** *Scenario 1*. **Bottom rows:** *Scenario 2*.

Results

Ablation Studies

- No additional computational cost is required compared to the base model when testing after adaptation to the test environment

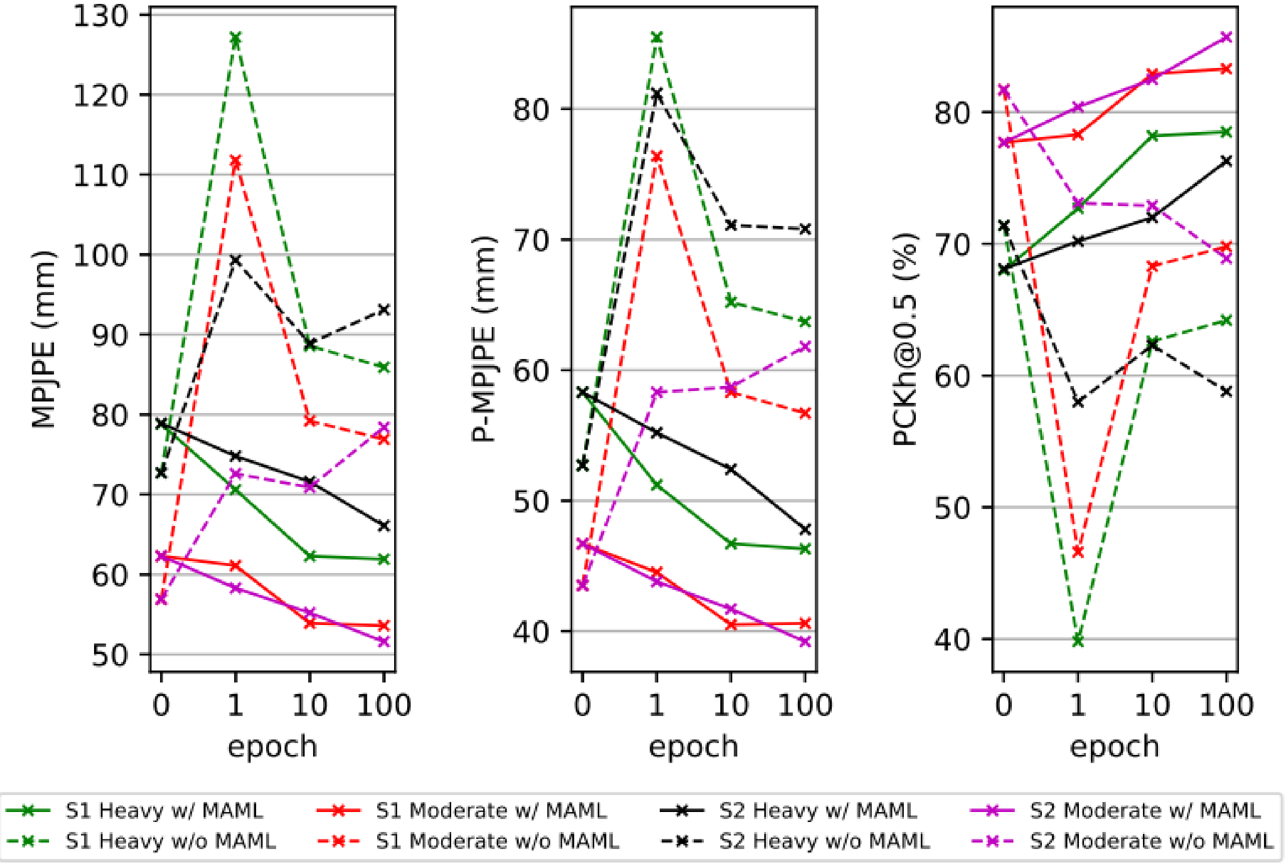
Model	Parameters	≈ FLOPs	MPJPE	P-MPJPE	PCKh@0.5
Pavllo *et al.* [21] 27f	8.56M	17.11M	72.4	53.8	72.0
Ours 3f	0.16M	0.32M	75.0	56.1	69.6
Ours 9f	4.36M	8.71M	59.8	45.4	79.5
Ours 27f	8.56M	17.11M	57.6	43.4	80.9

Table 5: Performance and computational complexity of various models under *Scenario 1*. The reported performance is the average value for all kinds of distortions.

Results

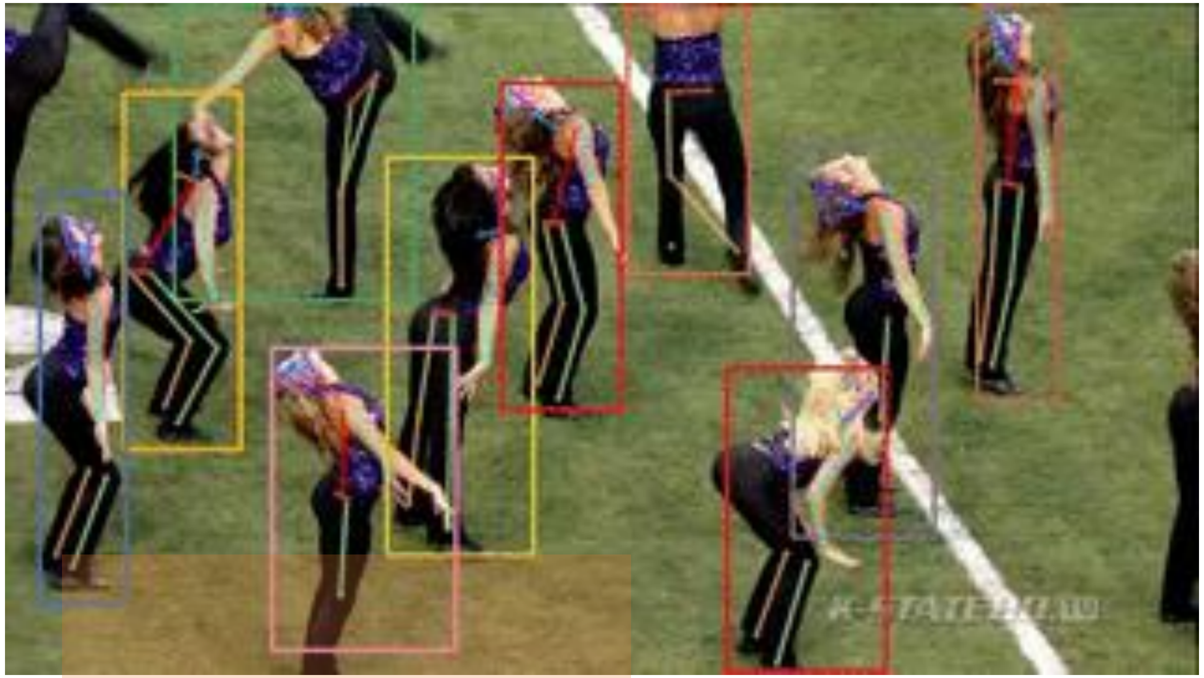
Performance Changes during Adaptation

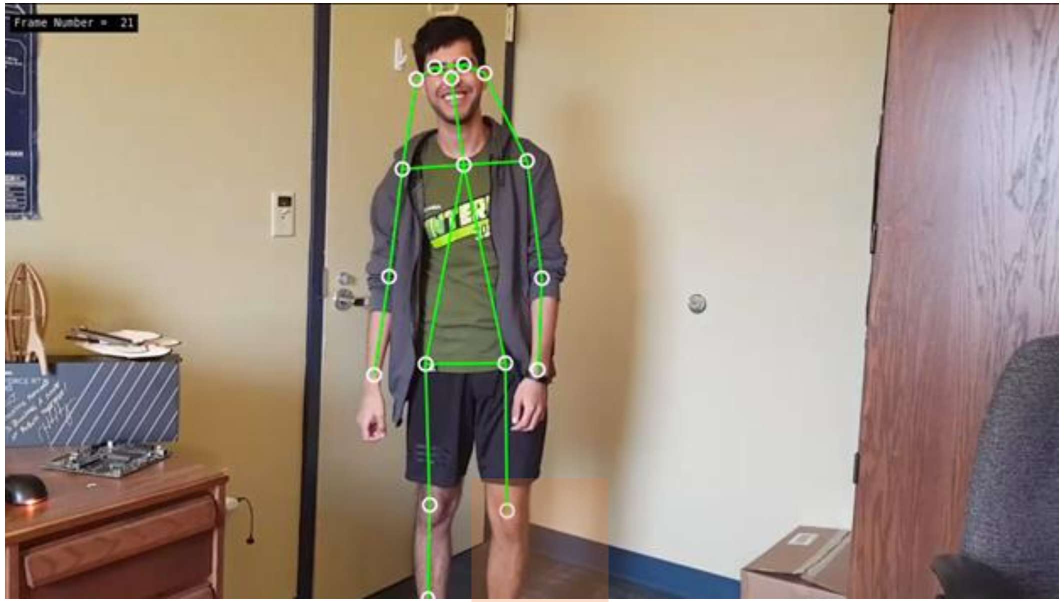
- Demonstrate the superior potential of MAML to adapt to various distortion environments.



Thank you


Kountoura
BCN
VICES




Frame Number = 21

Training size	AP	AP^M	AP^L
512	67.1	61.5	76.1
640	68.5	64.3	75.3
768	68.5	64.9	73.8

Table 5. Ablation study of HigherHRNet with different training image size on **COCO2017 val** dataset.