Unsupervised Cross-Dataset Adaptation via Probabilistic Amodal 3D Human Pose Completion

Jogendra Nath Kundu* Rahul Venkatesh* Jay Patravali* R. Venkatesh Babu

Video Analytics Lab, Indian Institute of Science, Bangalore, India (*equal contribution)

jogendrak@iisc.ac.in, {rahulvenkk, jaypatravali}@gmail.com, venky@iisc.ac.in

WACV 2020

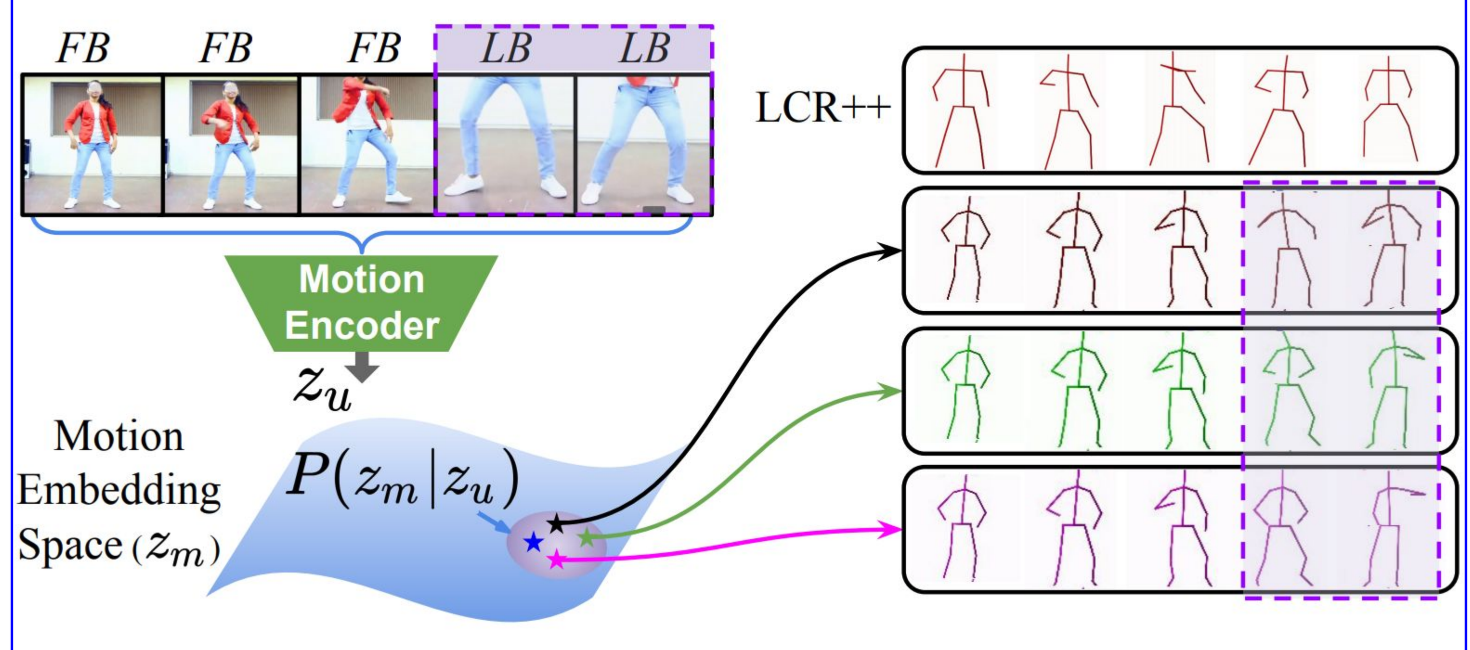
Overview

Motivations

- Performance of human 3D pose estimation approaches highly relies on availability of annotated training samples.
- Such models not only exhibit an alarming level of dataset bias, but also fail to operate on unconstrained videos in the presence of external variations such as camera motion, partial body visibility, occlusion, etc.
- we plan to formalize a self-supervised learning framework for human pose estimation particularly targeting unconstrained videos with no access to pose annotations (e.g videos collected from Youtube).

What's new?

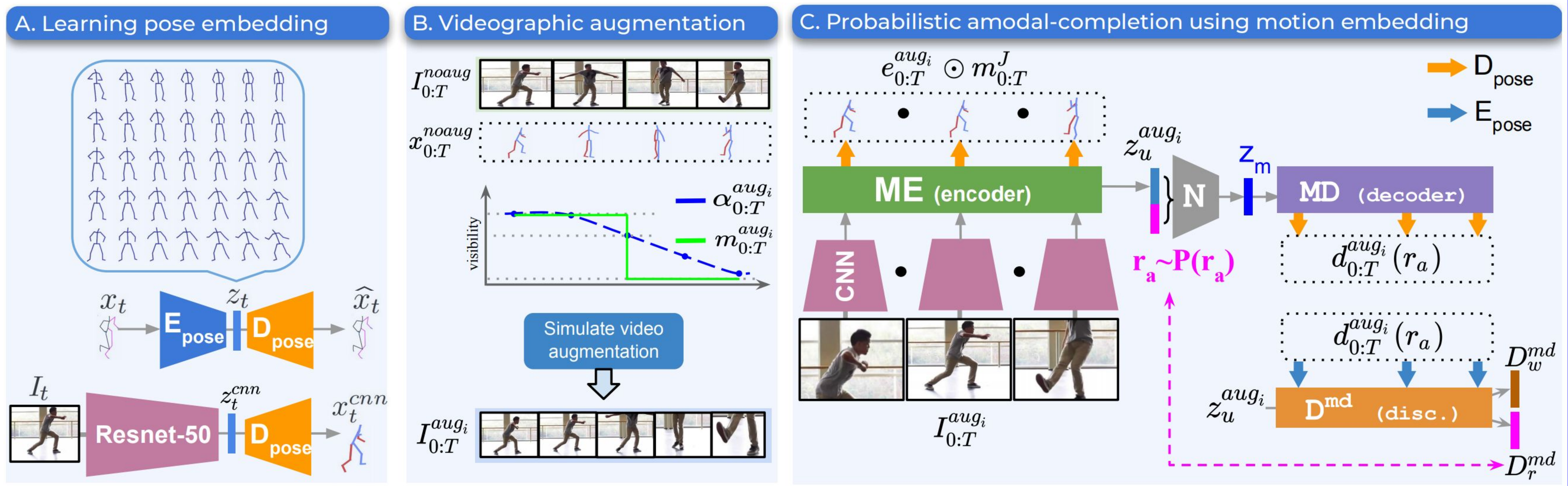
- Firstly, we aim to formalize a motion representation learning framework by effectively utilizing both constrained and artificially generated unconstrained video samples for datasets with 3D pose annotation.
- In contrast to the prior arts, our probabilistic model generates multiple plausible pose sequences (specifically for the invisible body-joints, i.e. the upper-body) for a given unconstrained video with partial-visibility.



- Secondly, to address dataset bias, the probabilistic amodal framework is re-utilized to design novel self-supervised objectives.

Approach

- Stage-wise training steps to enable probabilistic amodal motion completion

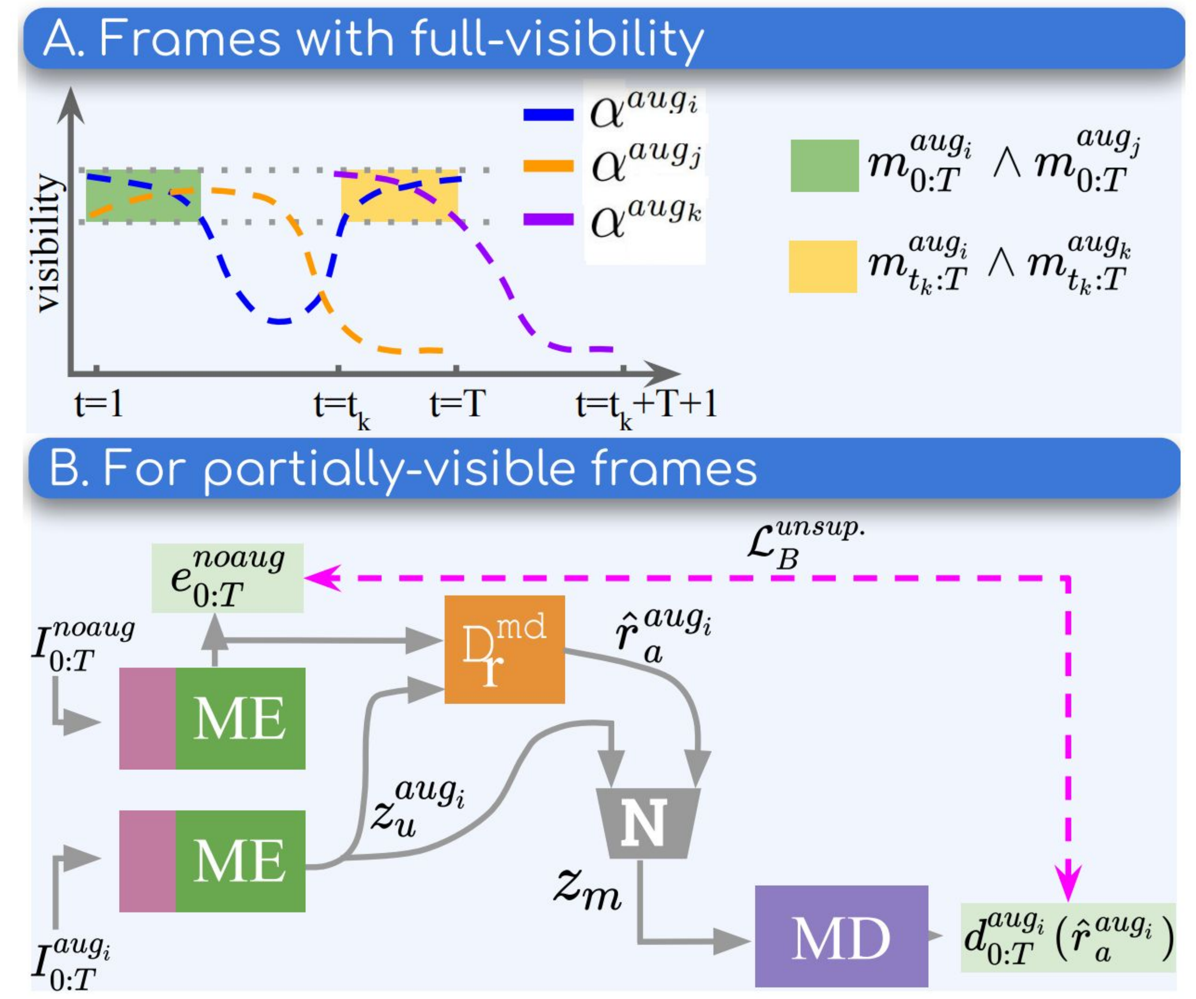


Training phase-1: The parameters of *ME*, *MD* and *N* are first trained using samples from both $\mathcal{D}^{unsim.}_{sup.}$ and $\mathcal{D}^{sim.}_{sup.}$ by enforcing $\mathcal{L}$ only for the visible time-steps i.e. $\mathcal{L}^{sup.}_A = \mathcal{L}(e^{aug_i}_{0:T}, x^{noaug}_{0:T}) \odot m^J_{0:T} + \mathcal{L}(d^{aug_i}_{0:T}(r_a), x^{noaug}_{0:T}) \odot m^J_{0:T}$

Training phase-2: After the first training phase, parameters of *N*, *MD* are finetuned along with the newly introduced discriminator D^{md} using an adversarial training framework

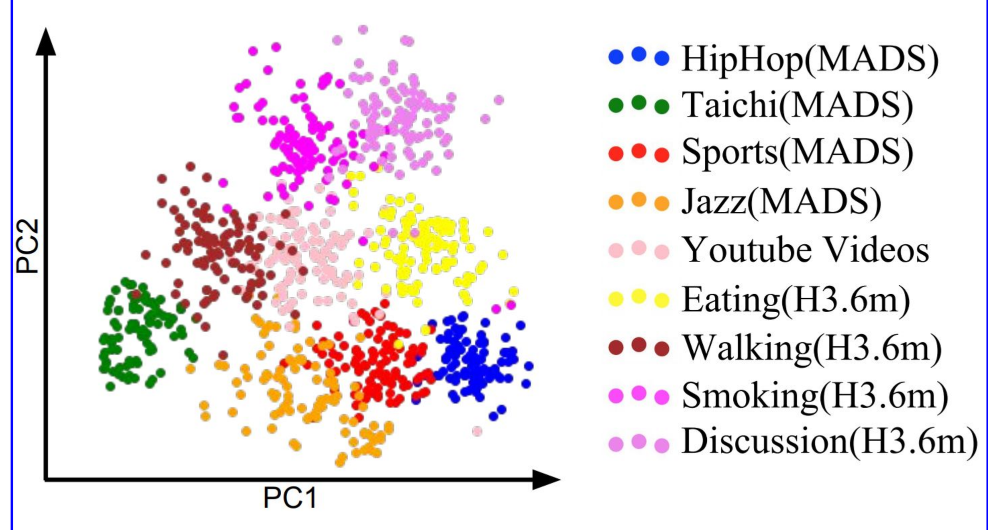
Training phase-3 (self-supervision): We define, $\mathcal{L}^{sup.}_B$ by replacing e^{noaug} with the supervised ground-truth x^{noaug} in $\mathcal{L}^{unsup.}_B$. The final unsupervised and supervised loss functions are represented as, $\mathcal{L}^{unsup.} = \mathcal{L}^{unsup.}_{A1} + \mathcal{L}^{unsup.}_{A1} + \mathcal{L}^{unsup.}_B$ and $\mathcal{L}^{sup.} = \mathcal{L}^{sup.}_A + \mathcal{L}^{X}_{content} + \mathcal{L}^{sup.}_B$

- Illustration of our self-supervised strategies



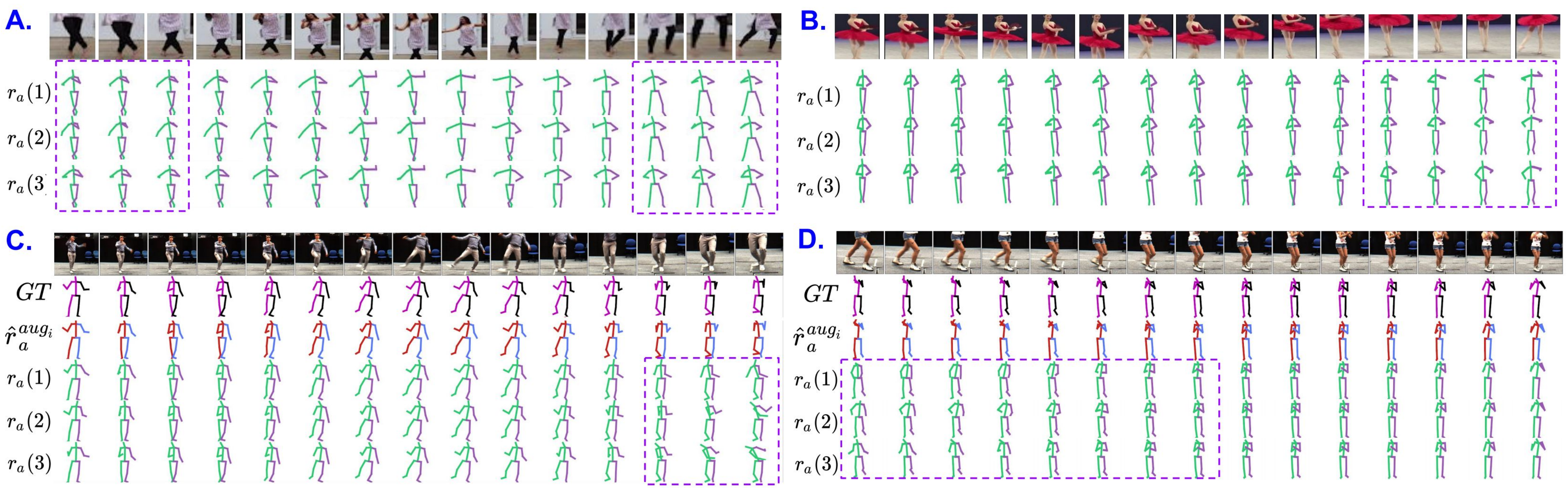
Results

- Motion embedding visualization



Cross-dataset 3D pose estimation results (PSS: Pose Structure Score)

Method	Dance Style			
	HipHop PSS(↑)	Jazz PSS(↑)	Taichi PSS(↑)	Sports PSS(↑)
Un-simulated Videos				
Vosoughi *et al.* [51]	67.4	73.9	62.5	30.3
LCR++ [43]	84.3	86.3	43.7	29.4
VNect [29]	69.2	67.5	74.5	34.5
Kocabas *et al.* [16]	74.3	72.9	66.8	39.3
Sun *et al.* [47]	77.8	78.1	69.3	40.1
Tekin *et al.* [14]	82.3	81.7	70.2	37.9
Pavllo *et al.* [39]	83.2	82.3	70.4	40.5
Ours-noSS	83.4	83.1	72.3	42.1
Ours-SSA	86.2	85.2	75.4	46.7
Ours-SSAB	**87.1**	**86.5**	**76.2**	**48.1**
Simulated Videos				
Sun *et al.* [47]	54.5	51.3	41.8	33.4
Kocabas *et al.* [16]	52.6	52.1	43.4	35.4
VNect [29]	54.5	54.8	60.1	26.6
LCR++ [43]	70.2	73.4	35.2	23.2
Vosoughi *et al.* [51]	66.4	64.2	58.9	32.1
Tekin *et al.* [14]	69.5	70.4	58.3	30.1
Pavllo *et al.* [39]	73.3	69.6	63.8	39.1
Ours-noSS	78.4	75.2	69.2	41.9
Ours-SSA	79.5	76.4	70.6	42.4
Ours-SSAB	**82.8**	**80.4**	**72.4**	**45.1**



Qualitative results on unconstrained in-the-wild videos.

A, B: On wild YouTube videos.

C,D: On MADS in-studio datasets.

Notice variations in pose-filling outcomes particularly for the non-visible joints (highlighted in dotted box).