寧夏大學
NINGXIA UNIVERSITY

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Two-stage Co-segmentation Network Based on Discriminative Representation for Recovering Human Mesh from Videos

Boyang Zhang[※]*, Kehua Ma*[※]*, Suping Wu*, Zhixiang Yuan*

School of information Engineering, Ningxia University

Yinchuan, China

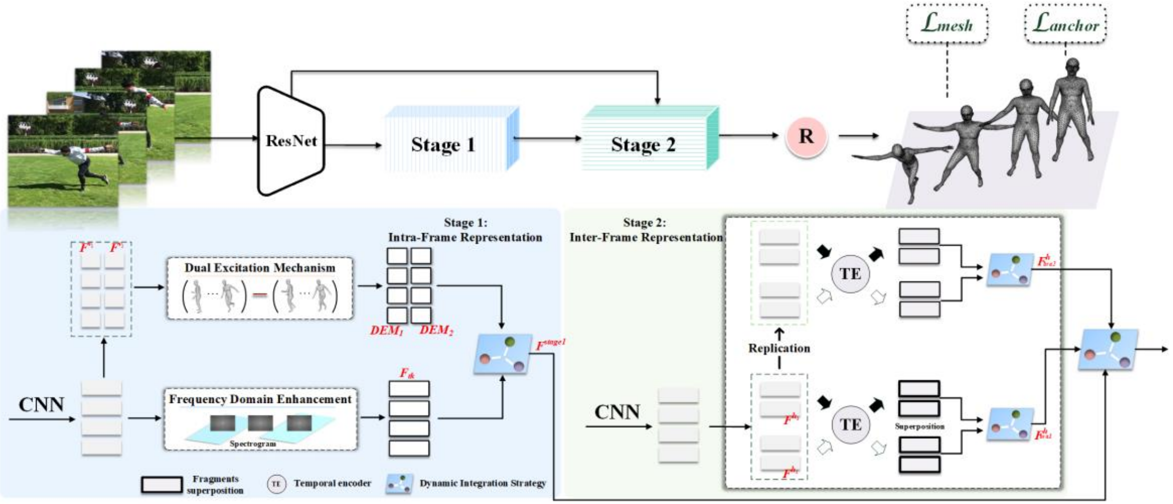
※ represents the equal contribution
* indicates the corresponding author

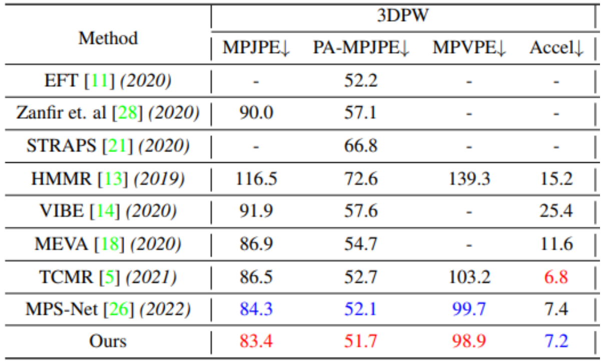
Tag: TUE-PM-146

Quick Preview

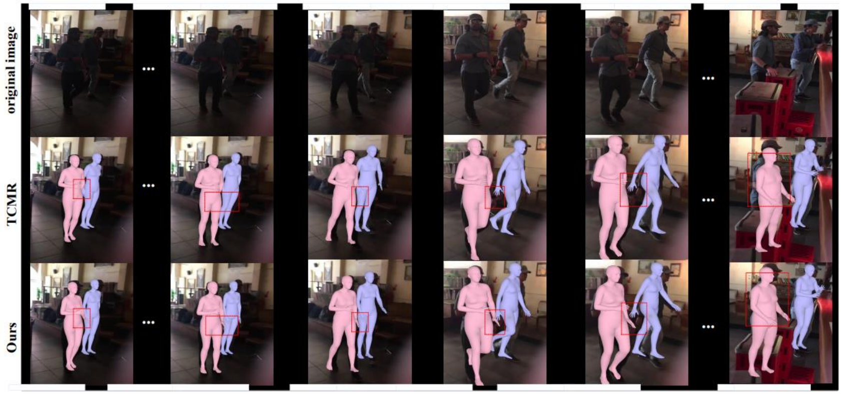
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

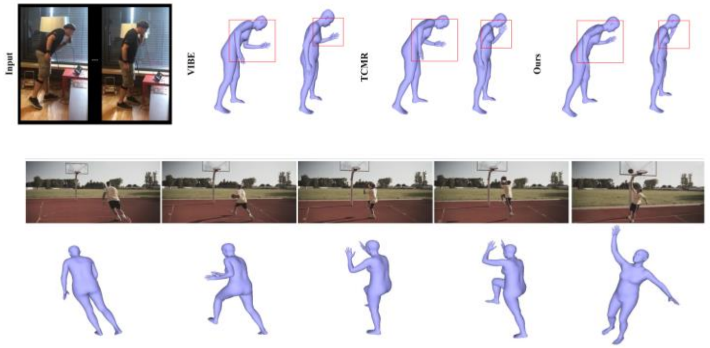
- Propose a two-stage co-segmentation network based on discriminative representation for recovering human mesh from videos. We segment spatial features to obtain multi-level spatial representation with dual excitation and dynamic integration strategy to model the spatio-temporal context, and we design a landmark anchor area loss to enhance the discriminative representation.





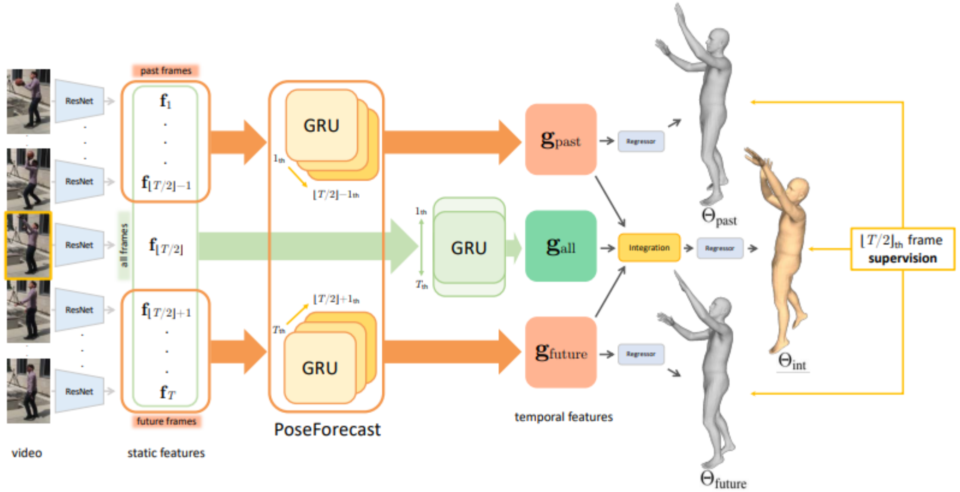
Method	3DPW			
	MPJPE↓	PA-MPJPE↓	MPVPE↓	Accel↓
EFT [11] *(2020)*	-	52.2	-	-
Zanfir et. al [28] *(2020)*	90.0	57.1	-	-
STRAPS [21] *(2020)*	-	66.8	-	-
HMMR [13] *(2019)*	116.5	72.6	139.3	15.2
VIBE [14] *(2020)*	91.9	57.6	-	25.4
MEVA [18] *(2020)*	86.9	54.7	-	11.6
TCMR [5] *(2021)*	86.5	52.7	103.2	6.8
MPS-Net [26] *(2022)*	84.3	52.1	99.7	7.4
Ours	83.4	51.7	98.9	7.2

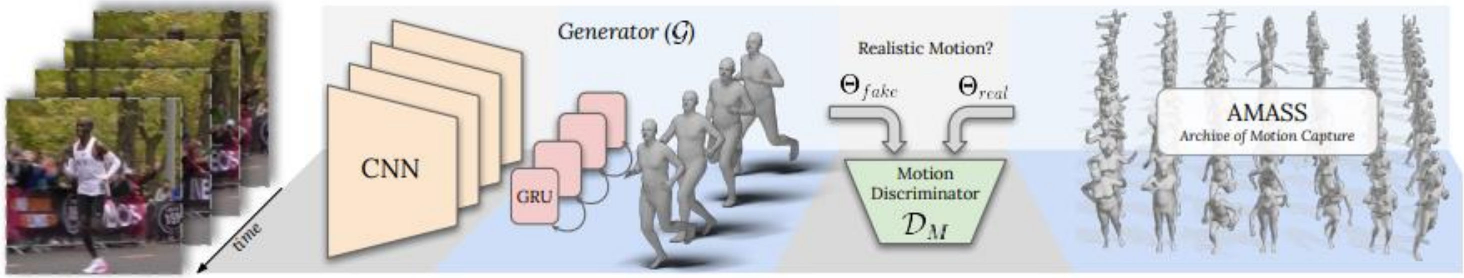


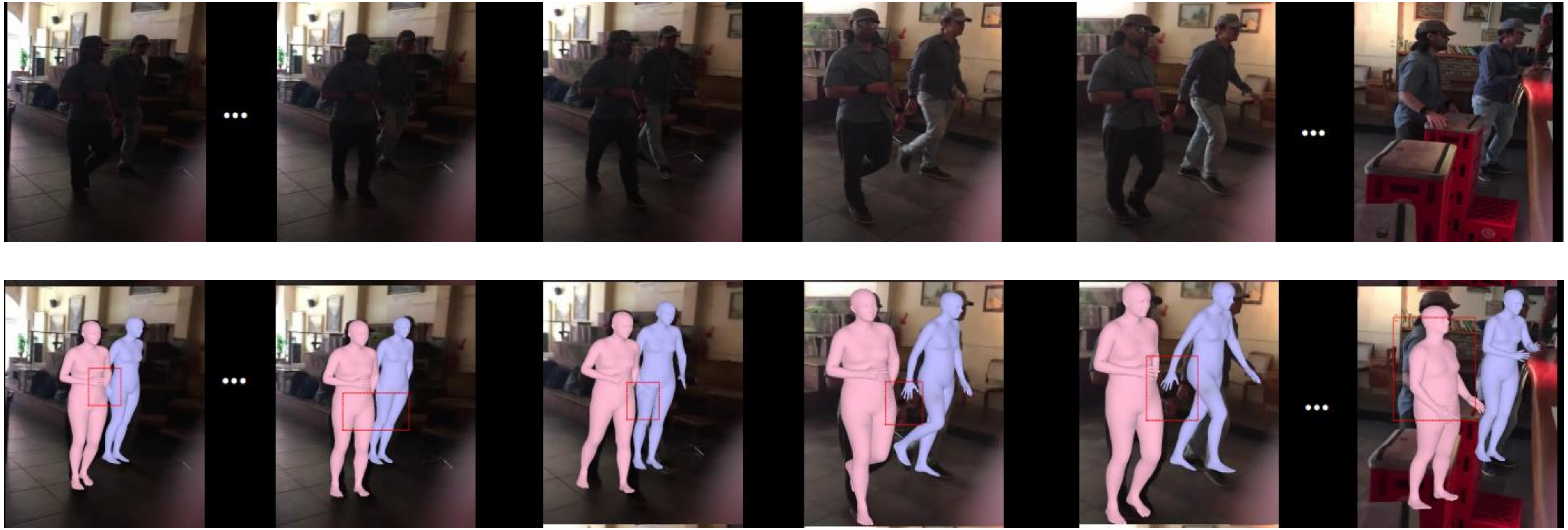


Background & Challenge

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA



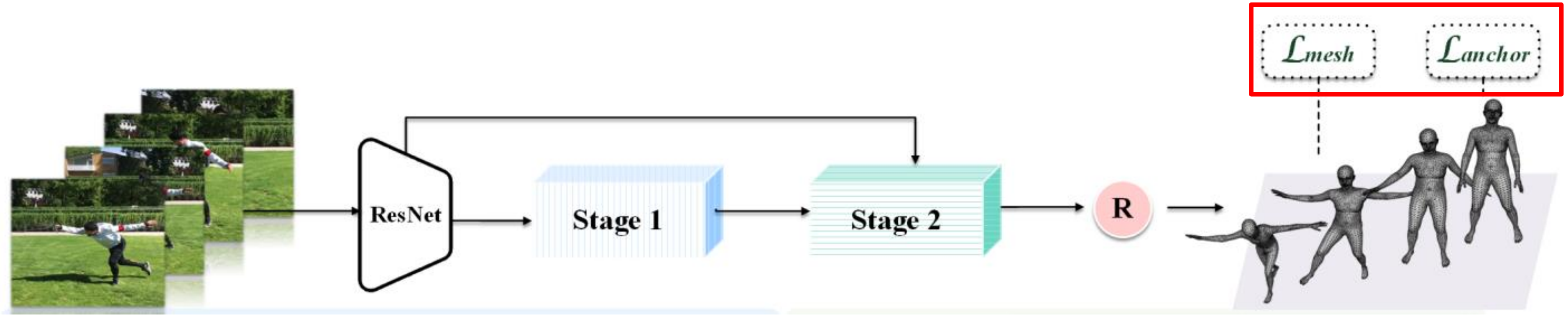


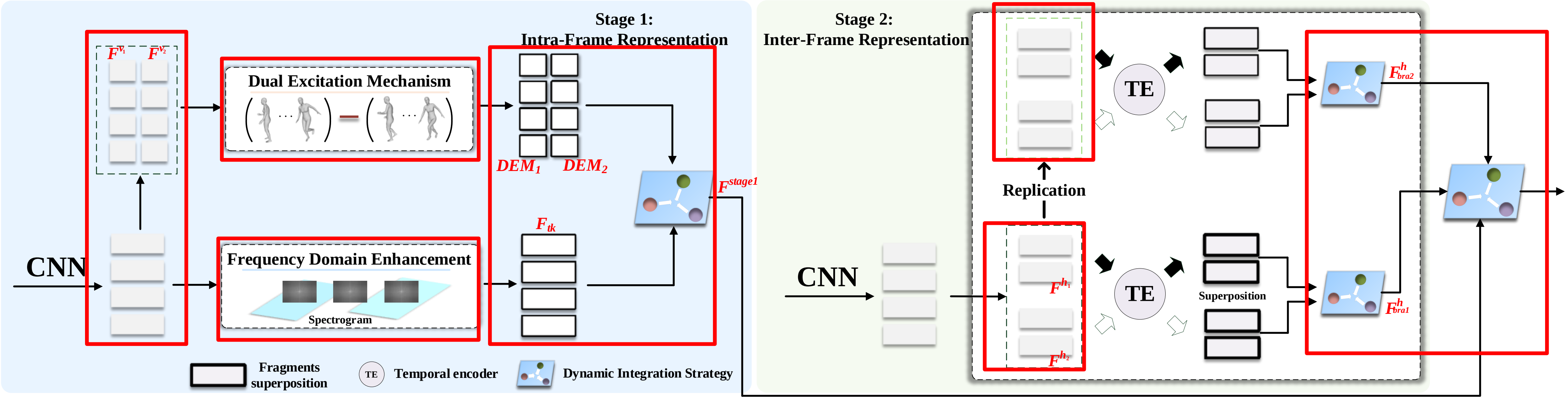


Contributions

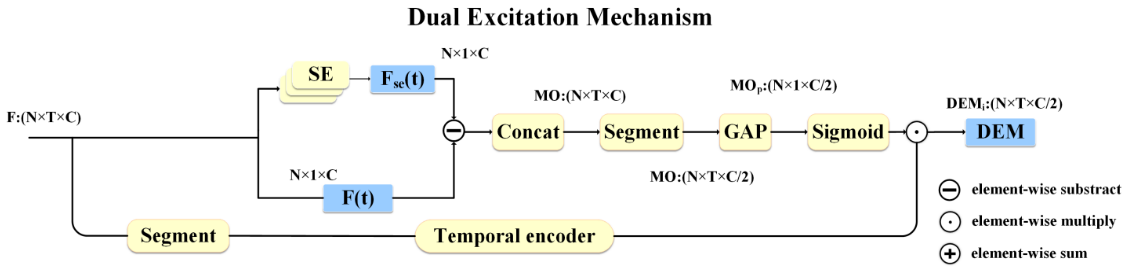
- Propose a co-segmentation network based on discriminative representation for recovering human mesh from videos. Our method motivates and learns spatiotemporal discriminative features at different stages.
- In Stage 1, our proposed dual excitation mechanism and frequency domain enhancement effectively enhance human motion features and mitigate background interference. In Stage 2, we develop a dynamic integration strategy to integrate the discriminative representations of distinct stages. We also carefully design a landmark anchor area loss to constrain the generation of the reasonable pose.
- The quantitative and qualitative results of our method show the effectiveness of the proposed method on widely evaluated benchmark datasets.

Overview

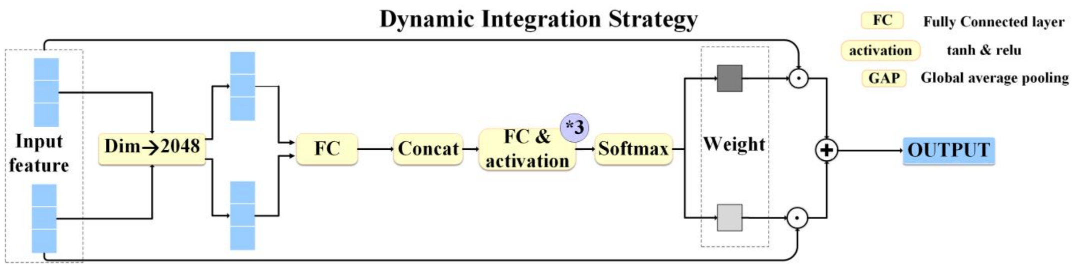

Lmesh
Lanchor
ResNet
Stage 1
Stage 2
R


Stage 1:
Intra-Frame Representation
F^v1
F^v2
Dual Excitation Mechanism
DEM1
DEM2
F^stage1
CNN
Frequency Domain Enhancement
Spectrogram
F_tk
Fragments superposition
TE
Temporal encoder
Dynamic Integration Strategy
Stage 2:
Inter-Frame Representation
TE
F^h_bra2
Replication
CNN
F^h1
F^h2
TE
Superposition
F^h_bra1

Stage 1 & Stage 2



We aim to obtain an intra-frame spatial multi-level representation to focus on spatial fine-grained information by segmenting the spatial features at Stage 1.



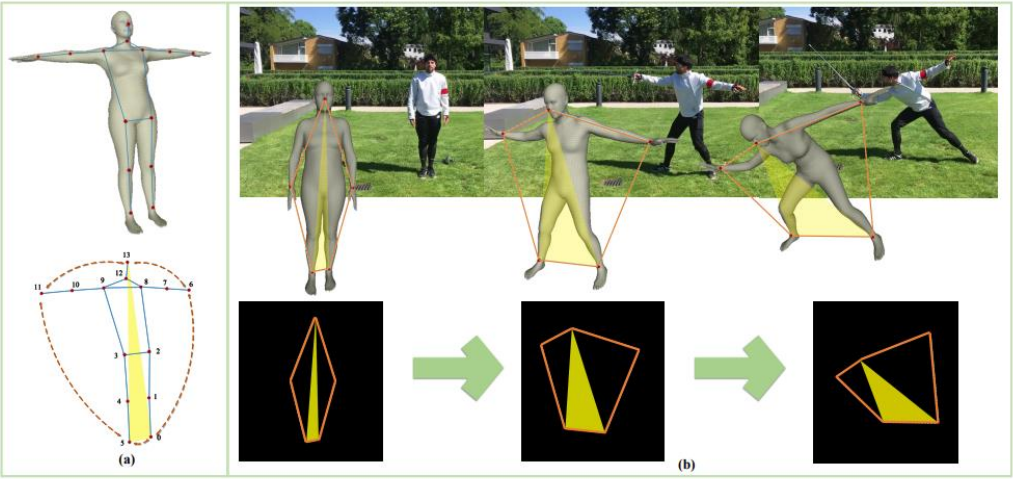
In Stage 2, Our goal is to obtain an inter-frame spatio-temporal multi-level discriminative representation by segmenting the spatial features and modeling spatio-temporal context.

When the human body and the background are present in the image, the human motion is usually reflected in the frequency spectrum in the video.

Loss Function

$$L = \lambda_{\mathcal{A}} L_{\mathcal{A}} + L_{\mathcal{M}}$$

Landmark Anchor Area Loss.



Mesh Parameter Loss.

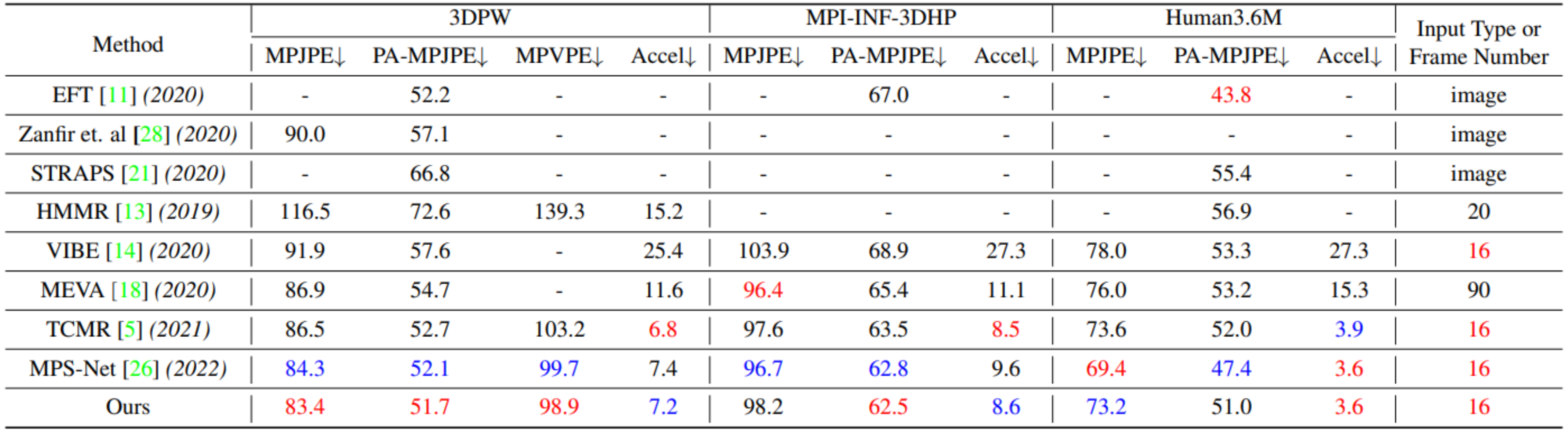
$$L_{\mathcal{M}} = \omega_{3dj} \sum_{t=1}^{T} \|X_t - \hat{X}_t\|_2 + \omega_{2dj} \sum_{t=1}^{T} \|x_t - \hat{x}_t\|_2$$

$$+ \omega_{shape} \|\beta - \hat{\beta}\|_2 + \omega_{pose} \sum_{t=1}^{T} \|\theta_t - \hat{\theta}_t\|_2$$

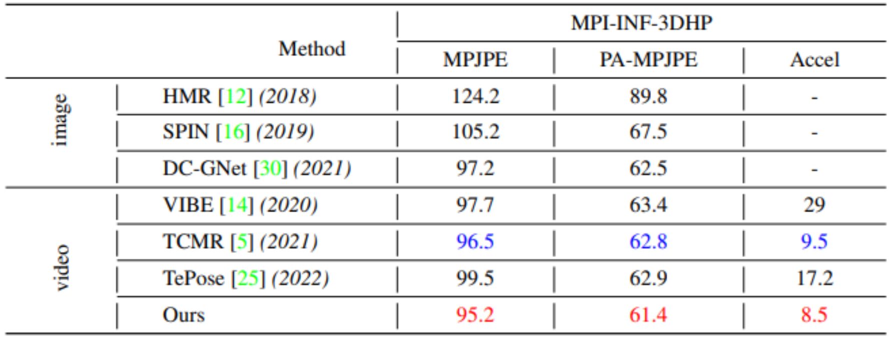
It is worth noting that our landmark anchor area loss is based on the same gesture translational invariant in the scene, so it focuses on the overall motion pose reasonably. And it is more insensitive to unalignment between 3D human joints and ground truth joints than 3D joint loss. Compared to previous works, our method not only has a joint loss as the previous method, but also constrains the perimeter and area geometry to regulate the overall human motion pose representation.

Experiments

Quantitative Comparison



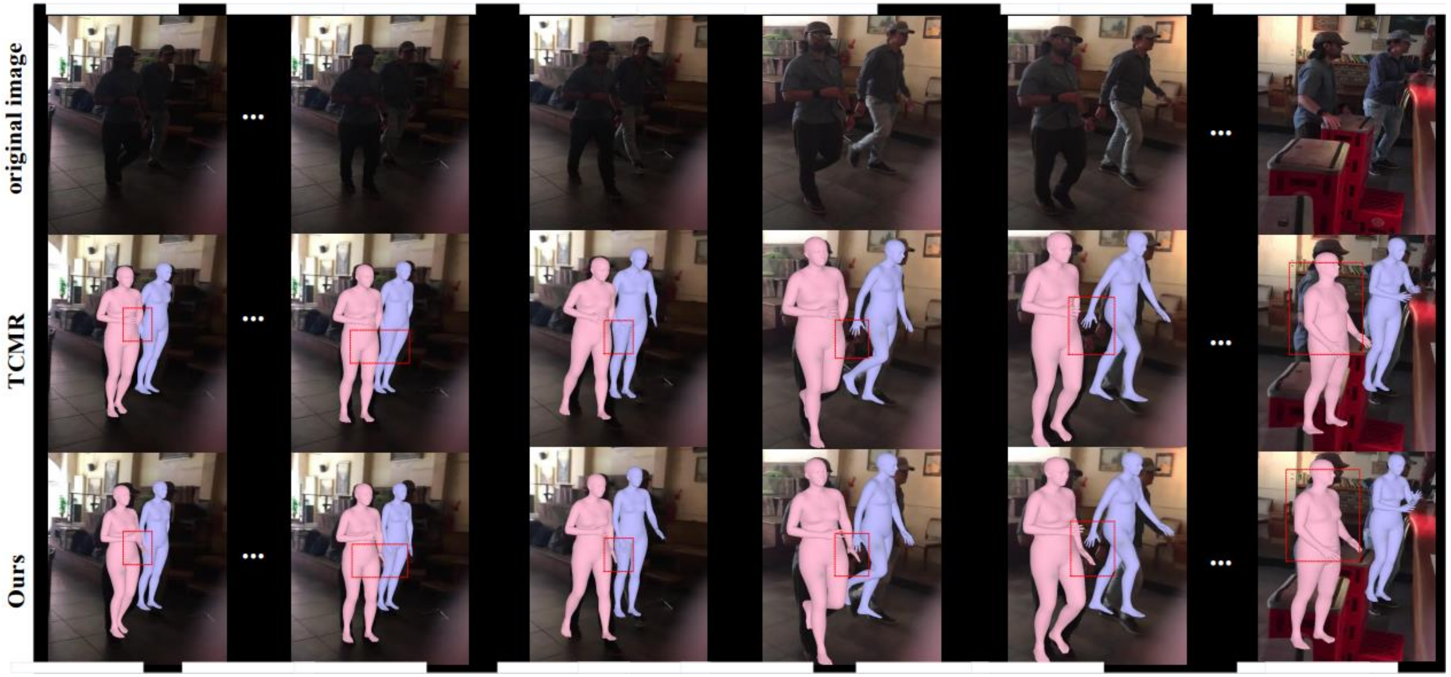
Method	3DPW				MPI-INF-3DHP			Human3.6M			Input Type or Frame Number
	MPJPE↓	PA-MPJPE↓	MPVPE↓	Accel↓	MPJPE↓	PA-MPJPE↓	Accel↓	MPJPE↓	PA-MPJPE↓	Accel↓	
EFT [11] *(2020)*	-	52.2	-	-	-	67.0	-	-	43.8	-	image
Zanfir et. al [28] *(2020)*	90.0	57.1	-	-	-	-	-	-	-	-	image
STRAPS [21] *(2020)*	-	66.8	-	-	-	-	-	-	55.4	-	image
HMMR [13] *(2019)*	116.5	72.6	139.3	15.2	-	-	-	-	56.9	-	20
VIBE [14] *(2020)*	91.9	57.6	-	25.4	103.9	68.9	27.3	78.0	53.3	27.3	16
MEVA [18] *(2020)*	86.9	54.7	-	11.6	96.4	65.4	11.1	76.0	53.2	15.3	90
TCMR [5] *(2021)*	86.5	52.7	103.2	6.8	97.6	63.5	8.5	73.6	52.0	3.9	16
MPS-Net [26] *(2022)*	84.3	52.1	99.7	7.4	96.7	62.8	9.6	69.4	47.4	3.6	16
Ours	83.4	51.7	98.9	7.2	98.2	62.5	8.6	73.2	51.0	3.6	16



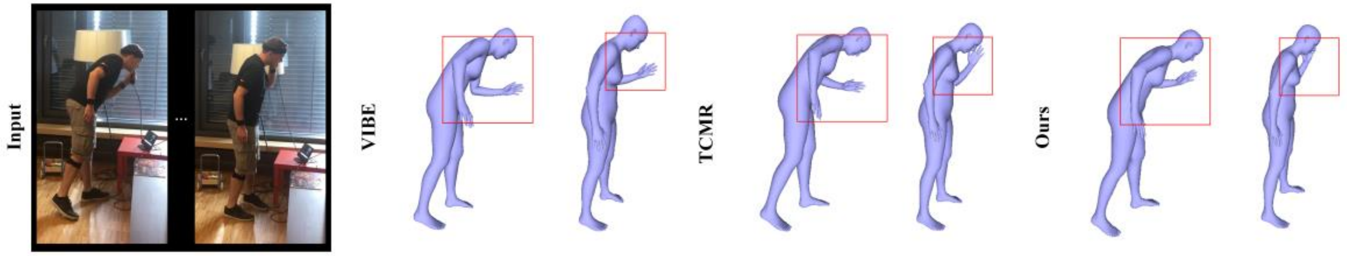
	Method	MPI-INF-3DHP		
		MPJPE	PA-MPJPE	Accel
image	HMR [12] *(2018)*	124.2	89.8	-
	SPIN [16] *(2019)*	105.2	67.5	-
	DC-GNet [30] *(2021)*	97.2	62.5	-
video	VIBE [14] *(2020)*	97.7	63.4	29
	TCMR [5] *(2021)*	96.5	62.8	9.5
	TePose [25] *(2022)*	99.5	62.9	17.2
	Ours	95.2	61.4	8.5

Experiments

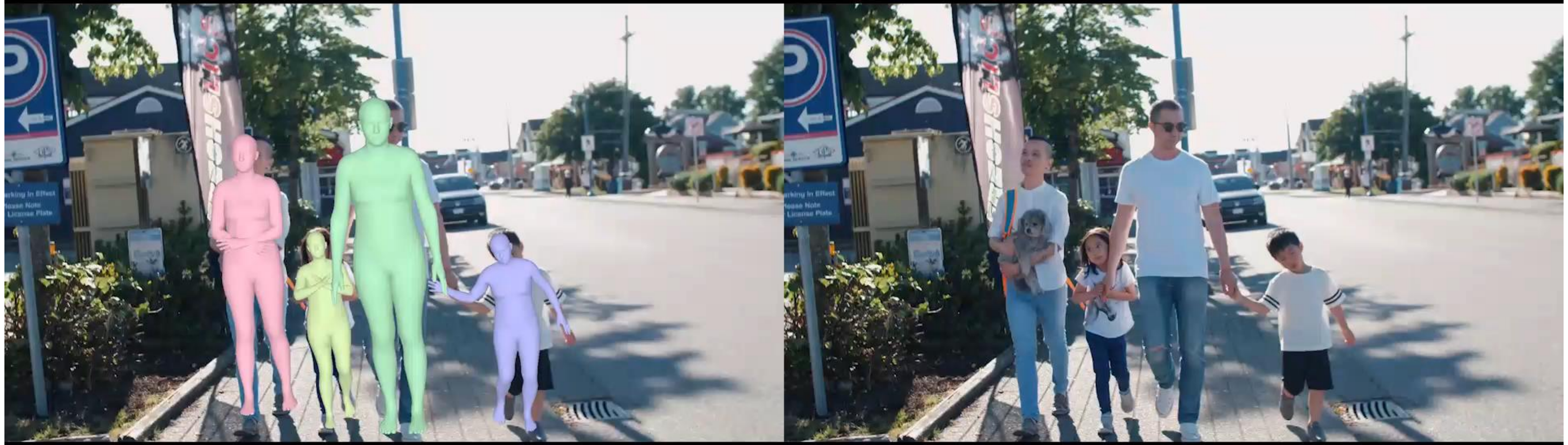
Qualitative comparison under extreme illumination



Qualitative comparison under chaotic backgrounds



Experiments



THANKS