Global-to-Pixel Regression for Human Mesh Recovery

Yabo Xiao[1,2], Mingshu He[*,1], Dongdong Yu[3]

[1] BUPT [2] Huawei Inc. [3] AISphere Tech.

(* Corresponding Author)

Motivations

- Existing HMR methods commonly leverage the global or dense–annotations–based local features to produce a single prediction. The compressed global and local features disrupt the spatial geometry, resulting in visual–mesh misalignment.

- Futhermore, dense annotations (e.g., IUV or part segmentation maps) are labor–intensive and expensive.

Contributions

- We propose a global–to–local wise HMR network, named GLNet, which can capture local details while maintaining spatial information to improve visual–mesh alignments without dense labels and heuristic rules.
- We introduce a **2D Keypoint–Guided Local Encoding Module** to drive each pixel feature to fuse local semantic–rich body parts’ information for global prediction refinement.
- Furthermore, we propose an **Adaptive Matching Strategy** by calculating the 2D components’ match costs between per–pixel predictions and ground–truth for assigning positive/negative samples.
- **GLNet** achieves SOTA performance and outperforms previous HMR methods significantly.

Framework

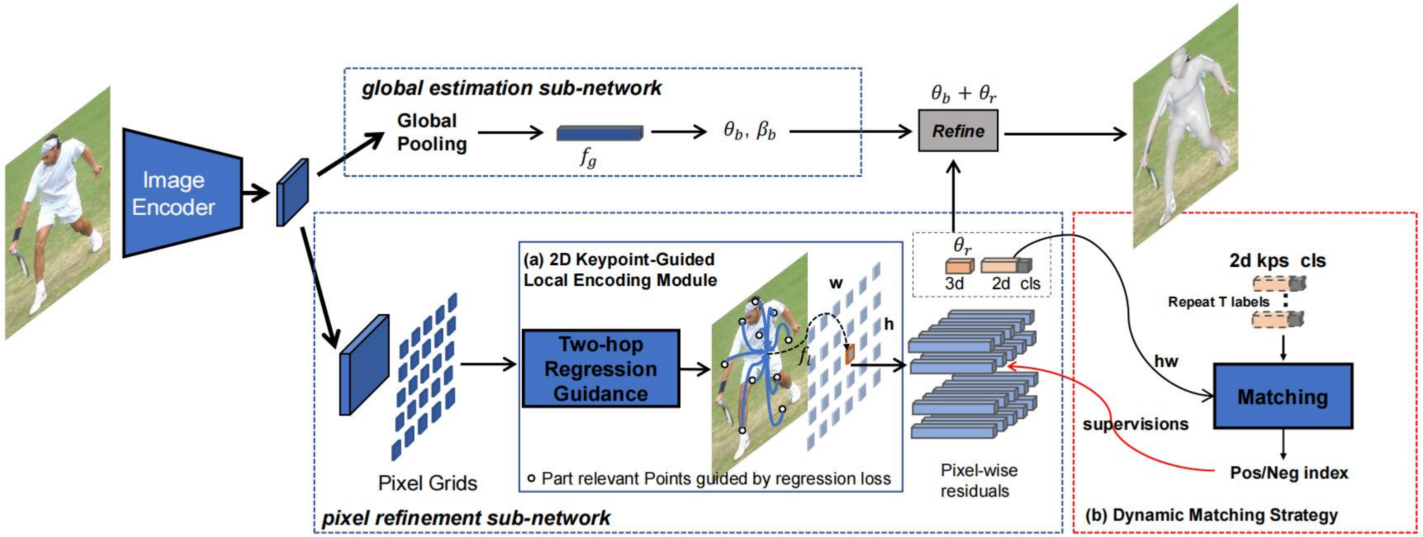


Image Encoder
global estimation sub-network
Global Pooling
f_g
θ_b, β_b
$\theta_b + \theta_r$
Refine
θ_r
3d
2d
cls
(a) 2D Keypoint-Guided Local Encoding Module
Two-hop Regression Guidance
w
h
f_l
Pixel Grids
Part relevant Points guided by regression loss
Pixel-wise residuals
pixel refinement sub-network
2d kps cls
Repeat T labels
hw
Matching
supervisions
Pos/Neg index
(b) Dynamic Matching Strategy

Fig. 2: Overview of GLNet. GLNet consists of a global estimation sub-network and pixel refinement sub-network. We leverage the global estimation sub-network to estimate camera parameters, highly abstract shape parameters, and base rotations. Afterwards, we use the pixel-wise feature output from 2D Keypoint-Guided Local Encoding Module to predict the residual rotations for rectifying base rotations. In the training stage, we leverage the Dynamic Matching Strategy that only considers the 2D elements to assign positive and negative samples.

$$L_{match}(y_i, \hat{y}_{\sigma(i)}) = -\alpha(1-\hat{P}_{\sigma(i)}(c))^{\beta} * log\hat{P}_{\sigma(i)}(c) + L_{kpt}(\hat{kpt}^{2D}_{\sigma(i)}, kpt^{2D}_i)$$

$$L_{all} = \lambda_{cls} * L_{cls} + \lambda_{2D} * \mathbb{I}_{\{c_i \neq \emptyset\}} L_{2D} + \lambda_{3D} * \mathbb{I}_{\{c_i \neq \emptyset\}} L_{3D}$$

Ablations

EUROPEAN CONFERENCE ON COMPUTER VISION

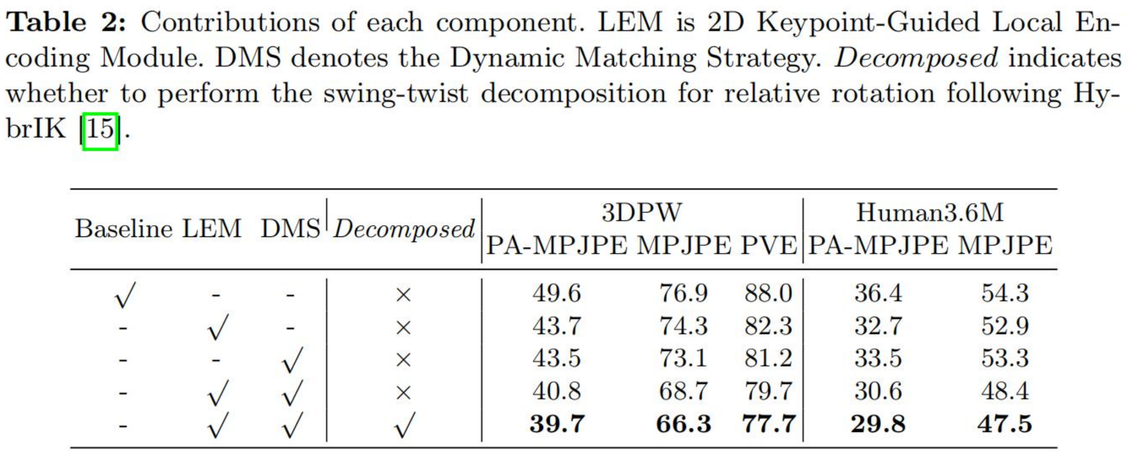


Table 2: Contributions of each component. LEM is 2D Keypoint-Guided Local Encoding Module. DMS denotes the Dynamic Matching Strategy. *Decomposed* indicates whether to perform the swing-twist decomposition for relative rotation following HybrIK [15].

Baseline	LEM	DMS	*Decomposed*	3DPW PA-MPJPE	3DPW MPJPE	3DPW PVE	Human3.6M PA-MPJPE	Human3.6M MPJPE
√	-	-	×	49.6	76.9	88.0	36.4	54.3
-	√	-	×	43.7	74.3	82.3	32.7	52.9
-	-	√	×	43.5	73.1	81.2	33.5	53.3
-	√	√	×	40.8	68.7	79.7	30.6	48.4
-	√	√	√	**39.7**	**66.3**	**77.7**	**29.8**	**47.5**

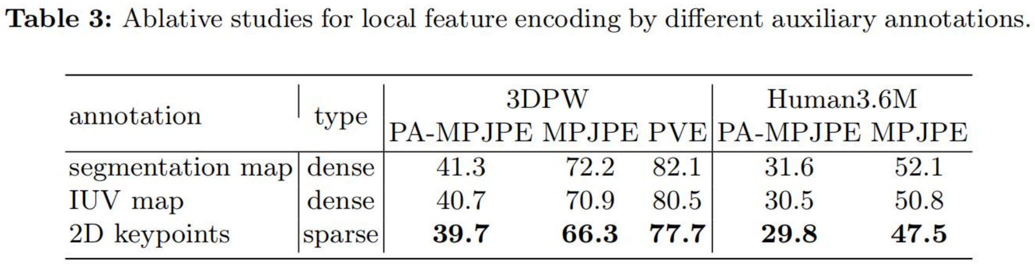


Table 3: Ablative studies for local feature encoding by different auxiliary annotations.

annotation	type	3DPW PA-MPJPE	3DPW MPJPE	3DPW PVE	Human3.6M PA-MPJPE	Human3.6M MPJPE
segmentation map	dense	41.3	72.2	82.1	31.6	52.1
IUV map	dense	40.7	70.9	80.5	30.5	50.8
2D keypoints	sparse	**39.7**	**66.3**	**77.7**	**29.8**	**47.5**

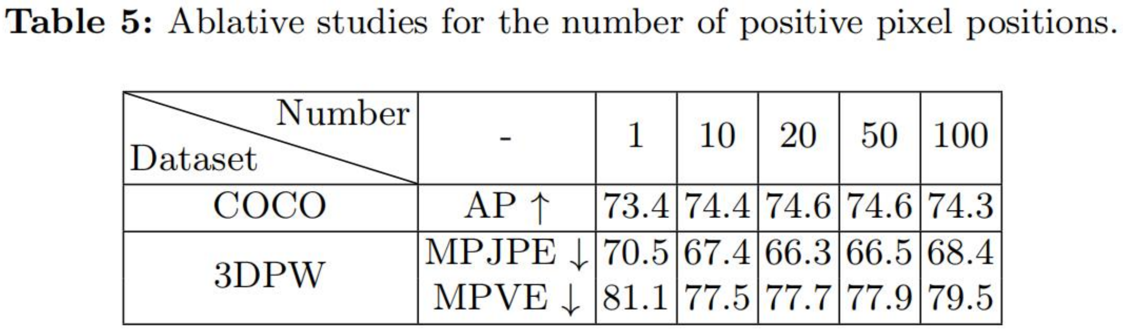


Table 5: Ablative studies for the number of positive pixel positions.

Number / Dataset	-	1	10	20	50	100
COCO	AP ↑	73.4	74.4	74.6	74.6	74.3
3DPW	MPJPE ↓	70.5	67.4	66.3	66.5	68.4
	MPVE ↓	81.1	77.5	77.7	77.9	79.5

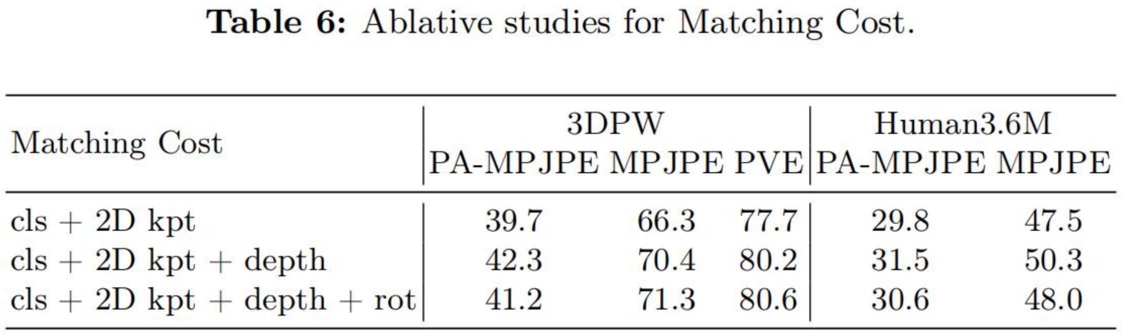
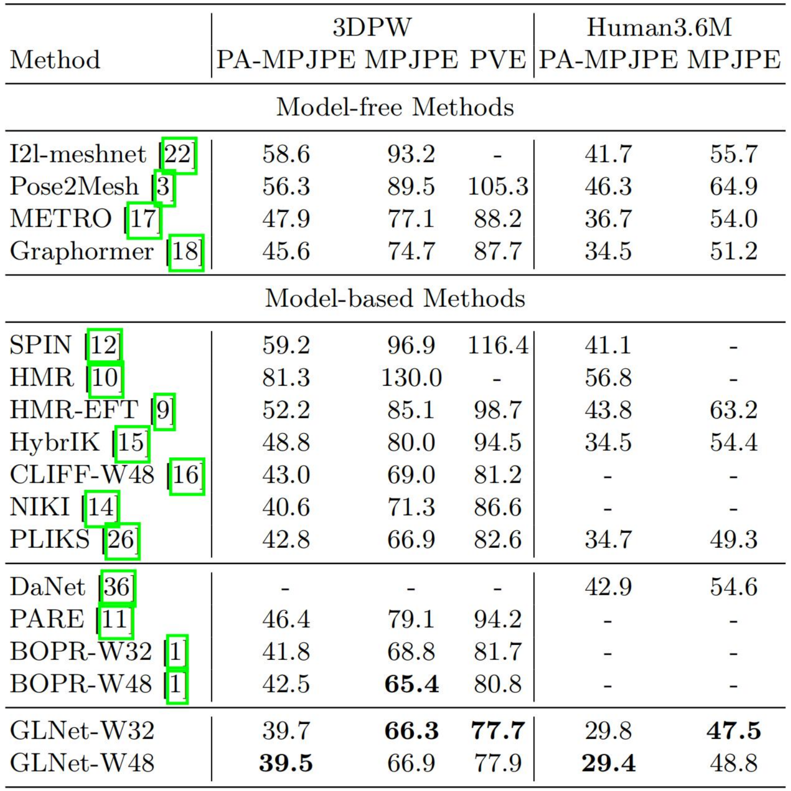


Table 6: Ablative studies for Matching Cost.

Matching Cost	3DPW PA-MPJPE	3DPW MPJPE	3DPW PVE	Human3.6M PA-MPJPE	Human3.6M MPJPE
cls + 2D kpt	39.7	66.3	77.7	29.8	47.5
cls + 2D kpt + depth	42.3	70.4	80.2	31.5	50.3
cls + 2D kpt + depth + rot	41.2	71.3	80.6	30.6	48.0

Results

Table 1: Comprehensive comparisons with previous methods on 3DPW and Human3.6M datasets.



Method	3DPW			Human3.6M	
	PA-MPJPE	MPJPE	PVE	PA-MPJPE	MPJPE
Model-free Methods					
I2l-meshnet [22]	58.6	93.2	-	41.7	55.7
Pose2Mesh [3]	56.3	89.5	105.3	46.3	64.9
METRO [17]	47.9	77.1	88.2	36.7	54.0
Graphormer [18]	45.6	74.7	87.7	34.5	51.2
Model-based Methods					
SPIN [12]	59.2	96.9	116.4	41.1	-
HMR [10]	81.3	130.0	-	56.8	-
HMR-EFT [9]	52.2	85.1	98.7	43.8	63.2
HybrIK [15]	48.8	80.0	94.5	34.5	54.4
CLIFF-W48 [16]	43.0	69.0	81.2	-	-
NIKI [14]	40.6	71.3	86.6	-	-
PLIKS [26]	42.8	66.9	82.6	34.7	49.3
DaNet [36]	-	-	-	42.9	54.6
PARE [11]	46.4	79.1	94.2	-	-
BOPR-W32 [1]	41.8	68.8	81.7	-	-
BOPR-W48 [1]	42.5	**65.4**	80.8	-	-
GLNet-W32	39.7	**66.3**	**77.7**	29.8	**47.5**
GLNet-W48	**39.5**	66.9	77.9	**29.4**	48.8

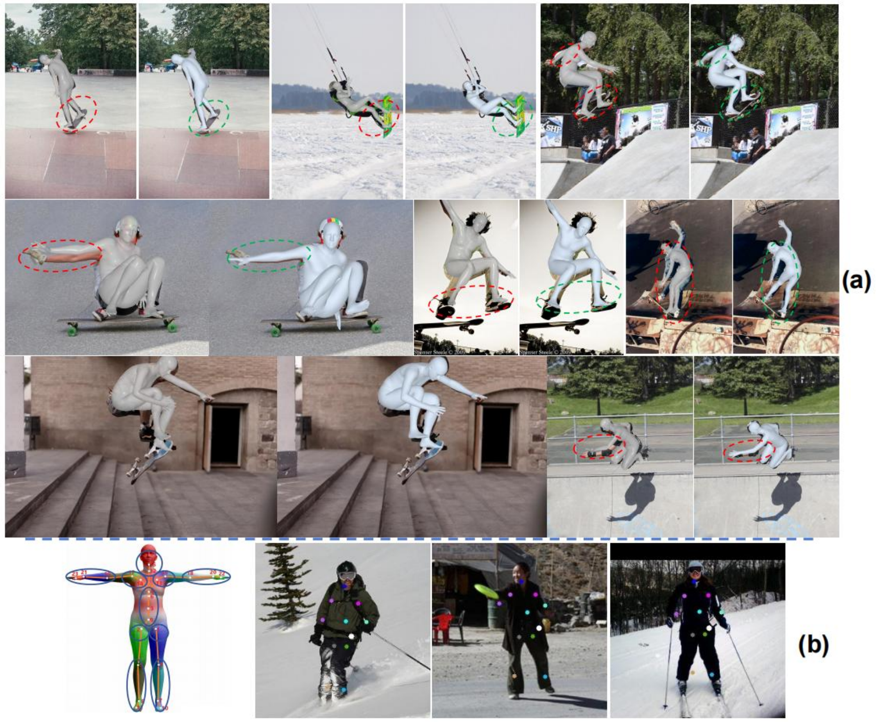


Fig. 3: (a) The images with red circles are the coarse predictions estimated by global features. The predictions with green circles are refined by local grid feature with 2D keypoint guidance. (b) The divided parts and corresponding parts relevant points. The white point is the reference point with the max confidence score.

ECCV

EUROPEAN CONFERENCE ON COMPUTER VISION

Thanks