Global-to-Pixel Regression for Human Mesh Recovery

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

[1] Beijing University of Posts and Telecommunications, Beijing, China
[2] Huawei Inc., Beijing, China
{xiaoyabo, hemingshu}@bupt.edu.cn
[3] AISphere Tech., Beijing, China
yudongdongatcasia@gmail.com

Abstract. Existing human mesh recovery (HMR) methods commonly leverage the global or dense-annotations-based local features to produce a single prediction from the input image. However, the compressed global and local features disrupt the spatial geometry of the human body and make it hard to capture the local dynamics, resulting in visual-mesh misalignment. Moreover, dense annotations are labor-intensive and expensive. Toward the above issues, we propose a global-to-local prediction framework to preserve spatial information and obtain precise visual-mesh alignments for top-down HMR. Specifically, we present an adaptive 2D Keypoint-Guided Local Encoding Module to enable per-pixel features to capture fine-grained body part information with structure and local context maintained. The acquisition of local features relies exclusively on sparse 2D keypoint guidance without dense annotations or heuristics keypoint-based ROI (Region of Interested) pooling. The enhanced pixel features are used to predict residuals for rectifying the initial estimation produced by global features. Secondly, we introduce a Dynamic Matching Strategy that determines positive/negative pixels by only calculating the classification and 2D keypoint costs to further improve visual-mesh alignments. The comprehensive experiments demonstrate the effectiveness of network design. Our framework outperforms previous local regression methods by a large margin and achieves state-of-the-art performance on Human3.6M and 3DPW datasets.

Keywords: visual-mesh alignments · Global-to-local · 2D Keypoint-Guided Local Encoding · Dynamic Matching Strategy

1 Introduction

Given an RGB image, human mesh recovery aims to reconstruct the 3D surface. It serves a key role in many downstream vision tasks and applications such as AR/VR, human-computer interaction, and so on. With the development of the parametric statistical human models, a realistic and controllable 3D mesh can be generated from the shape and rotations of articulations.

⋆ Corresponding author.

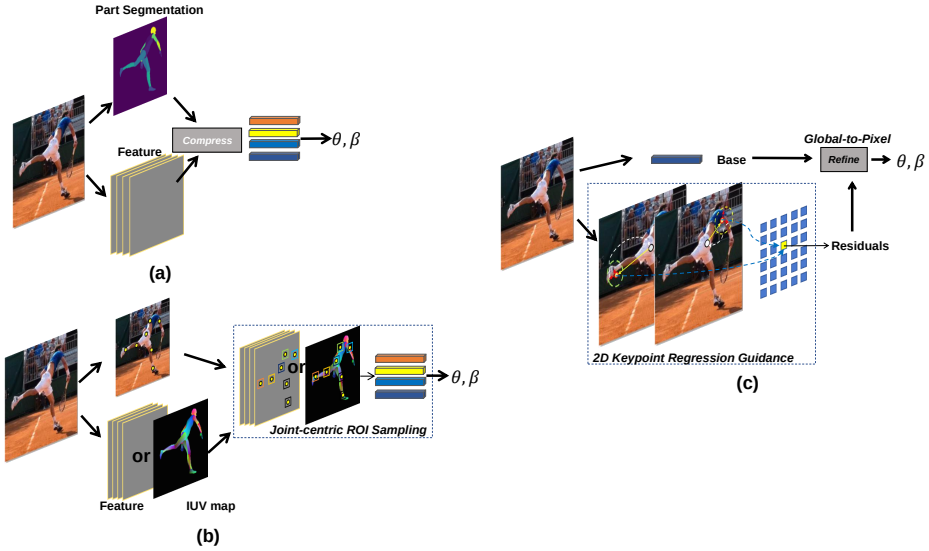

Fig. 1: (a) The framework predicts part segmentation maps to generate feature vectors corresponding to different body parts. (b) The framework leverages the predicted 2d keypoints to perform joint-centric ROI pooling on image features or IUV maps. (c) Our framework adopts a global-to-local prediction paradigm, which utilizes 2D keypoint regression guidance to extract local part features with spatial details maintained.

In general, 3D HMR methods can be classified into two categories, including optimization-based and regression-based methods. The former explicitly fits the output to 2D evidence, which can typically obtain accurate mesh-visual alignments but tends to be slow and sensitive to initialization. Regression-based methods directly predict human statistics parameters (i.e., shape parameters and relative rotations of articulations) in an end-to-end manner via the powerful modeling capacity of DNNs, which has made significant progress in recent years and has become the leading paradigm. However, most regression-based methods tend to compress image features by average pooling and leverage global features to predict shape and rotation parameters. We consider that the compression process discards local details and spatial geometry information. It is essential to maintain local spatial details for fine-grained body part perception. Several works attempt to leverage auxiliary dense representations such as part segmentation and IUV map to concentrate on body parts separately, as shown in Figure 1 (a) and (b). However, there is no 3D human pose dataset providing additional part segmentation or IUV maps generally. The preparation of the dense annotations is tedious and thus may involve uncertainty errors. Furthermore, though these works take regional evidence into consideration, the compression process based on local evidence loses spatial constraint thus leading to misalignment.

To maintain spatial information while capturing the local dynamics without additional annotation burdens, we present a novel global-to-local wise mesh re-

covery network, termed GLNet. As shown in Figure 1 (c), GLNet follows a coarse-to-fine prediction manner that uses regional grid features with spatial geometry to refine coarse global prediction. Each local grid feature leads to a proposal of the entire body for refinement. In detail, we first propose a 2D keypoint-guided Local Encoding Module to incorporate regional context into pixel-wise features. In this Module, we represent human parts as adaptive points following AdaptivePose [34] and adapt this representation into the top-down paradigm. The human body is divided into several parts, 2D keypoint position and corresponding rotation in each local part are encoded by the adaptive point's feature. We leverage 2D keypoint supervision to enable the network to focus on a set of local semantic positions corresponding to different body parts. Based on the above insight, we leverage global feature to predict initial parameters and further generate residuals by local adaptive points' features to refine initial results. Second, due to the pixel-wise framework generating plenty of predictions, we propose a Dynamic Matching Strategy to determine the positive and negative samples. The matching process only considers the classification and 2D keypoint matching costs to ensure the visual-mesh alignment. For inference, we select the final results according to the classification scores. Our proposed network achieves 29.4 mm PA-MPJPE on Human3.6M and 39.5 mm PA-MPJPE on 3DPW dataset.

Our contributions can be summarized as follows:

1. We propose a global-to-local wise human mesh recovery network, named GLNet, which can capture local details while maintaining spatial information to improve visual-mesh alignments by sparse and adaptive 2D keypoint guidance without dense labels and heuristic rules.

2. 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.

3. Equipped with the proposed 2D Keypoint-Guided Local Encoding Module and Dynamic Matching Strategy for training, GLNet achieves state-of-the-art performance and outperforms previous HMR methods significantly.

2 Related Work

In this section, we review two main paradigms for 3D human mesh recovery, including optimization-based and regression-based paradigms in Subsection 2.1 and 2.2. Then, we review the usage of 2D keypoint proxy in the HMR task and discuss the difference with our method in Subsection 2.3.

2.1 Optimization-based Methods

Optimization-based approaches primarily aim to estimate a 3D mesh that aligns with 2D image clues. The objective function generally comprises two parts: data terms and regularization terms. Data terms are used to measure the consistency

between the 2D label and its 3D re-projection. Regularization terms are crucial for achieving a physically plausible body mesh. They introduce pose priors to guard against unrealistic poses, thereby enhancing the realism and feasibility of the generated 3D mesh. For example, SMPLify [24] fits SMPL parameters to the predicted 2D keypoints iteratively. Specifically, it adopts an existing 2D pose estimator to locate the keypoints and perform gradient-based optimization. The objective function is composed of a 2D keypoint-based data term and several regularization terms, including an interpenetration error term, two pose priors, and a shape prior. Follow-up works attempt to use other 2D representations, e.g., silhouettes [13], 3D Part Orientation Fields [32], dense correspondences [6] or contact [23]. Moreover, deep learning techniques can be embedded into the gradient-based optimization process to enhance robustness and plausibility. Exemplar Fine-Tuning [9] leverages a pre-trained 3D pose estimator and performs optimization based on the strong pose priors. Song et al. [27] design an iterative algorithm based on gradient descent to generate the parameters that fit the SMPL parameters to 2D observations.

2.2 Regression-based Methods

Regression-based approaches directly predict the SMPL parameters from monocular images. We further categorize the existing regression-based works into global and local regression methods.

Global-based regression approaches compress the image feature to a global feature vector for regressing the pose and shape parameters. HMR [10] proposes to use end-to-end adversarial learning and an iterative error feedback (IEF) technique to reduce the regression errors. SPIN [12] forms a tight collaboration between regression-based and optimization-based paradigms that use the regressed pose to initialize the iterative optimization routine. CLIFF [16] reveals that the global rotations cannot be accurately inferred when only using cropped images and present to feed the model with the cropped-image feature with its bounding box information. HMR 2.0 [5] uses ViT as the image encoder and introduces SMPL query token to probe relevant visual features, then employs a standard transformer decoder with multi-head self/cross-attention to conduct SMPL parameter predictions. Besides, inverse kinematics has also been explored. HybrIK [15] designs an adaptive IK algorithm to convert 3D keypoints to the swing rotations, the shape and twist rotations are regressed by global feature. NIKI [14] further combines the FK (Forward Kinematics) and IK (Inverse Kinematics) processes using an invertible neural network to explicitly decouple errors from plausible poses.

Although the aforementioned methods attempt to use diverse intermediate estimations to facilitate the image-to-parameters mapping. The global feature is highly abstract thus is hard to capture the fine-grained high-frequency information. To alleviate this issue, several researches leverage local feature to attend to the corresponding body part with the aid of auxiliary representations. PARE [11] introduces a soft attention mechanism to enable the network to focus on local body parts by learning part segmentation masks. The image feature

maps are embedded as corresponding local feature vectors via predicted part attention masks. PyMAF [37] performs mesh alignment feedback loop by leveraging a feature pyramid to rectify the parameters explicitly from coarse to fine. FastMETRO [2] follows the model-free paradigm and utilizes joint as well as vertex tokens for non-parametric predictions via a transformer encoder-decoder architecture.

The aforementioned local regression methods rely on dense proxy representations to drive the network to perceive foreground body regions. The local features are embedded as 1D vectors while losing local spatial structure. In contrast, we design a global-to-local pipeline that uses sparse 2D keypoint guidance to encode regional context with spatial geometry maintained.

2.3 2D Keypoint Proxy

Instead of fitting SMPL parameters to 2D keypoints via re-projection loss, several methods adopt 2D keypoints as intermediate representations. Pavlakos et al. [25] utilize the 2D keypoint heatmap to predict pose parameters. Tung et al. [30] concatenates the image with the corresponding 2D heatmap as input. HaloPose [6] and DaNet [36] leverage the predicted joint positions to perform joint-centric RoI pooling, extract and fuse the partial feature or IUV map [7]. Pose2Mesh [3] and MotionBert [39] design a graph convolution and a transformer-based network that recovers 3D body mesh from the 2D keypoints respectively.

In contrast, GLNet represents compositional body parts as unconstrained relevant points with local contexts. The guidance of 2D keypoint regression enables the unconstrained points to perceive body part information respectively. The features with spatial and local context are sampled and fused to conduct global-to-local parameter rectification in the feed-forward process.

3 Method

In this section, we first briefly review the commonly-used parametric model SMPL [20] and its representative variant HybrIK [15] in Subsection 3.1, followed by elaboration on our global-to-local wise prediction framework in Subsection 3.2. In Subsection 3.3, we introduce a 2D Keypoint-guided Local Encoding Module. Finally, we describe the Dynamic Matching Strategy to assign supervisions that guarantee the visual-mesh alignments in Subsection 3.4.

3.1 Preliminary

SMPL representation. Most existing methods represent the 3D body mesh using the Skinned Multi-Person Linear (SMPL) model, which forms the human body via two characteristics including shape and pose. The shape $\beta \in \mathbb{R}^{10}$ is parameterized by the first 10 coefficients of a PCA space, indicating how individuals vary in height, weight, body proportions, etc. The pose parameters $\theta \in \mathbb{R}^{24\times3}$

consist of the global rotation of the root joint (pelvis) and 23 relative rotations of other joints relative to their parents along the hierarchical kinematic tree in axis-angle representation. SMPL present to use a differentiable function M that outputs a triangulated mesh with $N = 6980$ vertices, where $M(\theta, \beta) \in \mathbb{R}^{N\times 3}$. The 3D keypoints $J_{3D} \in \mathbb{R}^{k\times 3}$ can be generated from the linear combinations of the vertices and pretrained linear regression matrix.

Swing-Twist Decomposition. HybrIk [15] argues that directly regressing the relative rotations is highly non-linear and proposes an Inverse Kinematics-based solution via twist-and-swing decomposition. The relative rotation is factorized into twist and swing, i.e. a longitudinal rotation and an in-plane rotation. HybrIK introduces an analytical-neural inverse kinematics solution to perform 3D body mesh by 3D keypoint positions. It utilizes the predicted 3D keypoints to calculate the swing rotation analytically by the IK process. The process can be formulated as $R^{sw} = IK(P, T)$, where R is relative rotation, $P = \{p_k\}_{k=1}^{K}$ indicates input keypoints and $T = \{t_k\}_{k=1}^{K}$ denotes the rest pose. Specifically, the rotation ought to guarantee $P_k - P_{parent(k)} = R_k(t_k - t_{parent(k)})$. Then, HybrIK leverages a neural network to estimate the twist rotation R^{tw} via visual information. A accurate rotation R can be synthesized through $R = R^{tw}R^{sw}$.

Our global-to-local framework is representation-agnostic and can be compatible with direct rotation regression or twist-and-swing decomposition. For twist-and-swing decomposition, we leverage global feature to predict the initial twist rotations and use pixel features to rectify them. In parallel, we calculate the swing rotation by the predicted 2D keypoint and depth information. We report the results of two rotation representations in Table 2

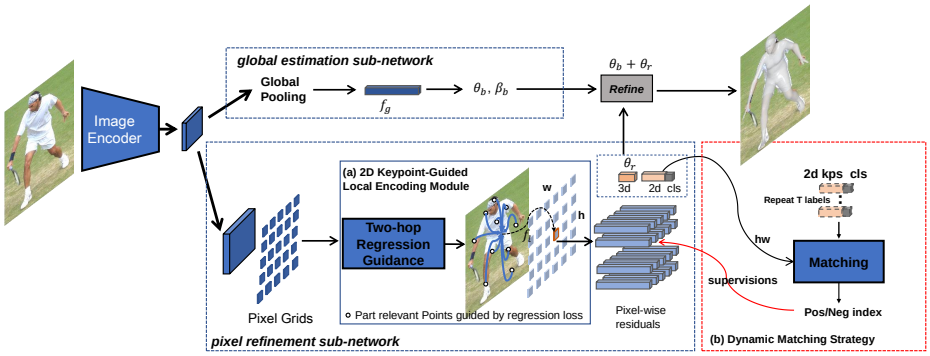

Fig. 2: Overview of GLNet. GLNet consists of a global estimation sub-network and pixel refinement sub-network. We leverage the global estimation sub-network to estimate camera parameters, highly abstract shape parameters, and base rotations. Afterwards, we use the pixel-wise feature output from 2D Keypoint-Guided 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.

3.2 Whole Framework

Previous works compress and aggregate spatial features by average pooling or soft attention masks thus probably leading to spatial misalignment. In contrast, we propose to use the global-to-local framework to maintain spatial structure and high-frequency information, pursuing the image-mesh alignments via 2D keypoint-guided regional feature.

As shown in Figure 2, we leverage widely-used HRNet [28] to encode visual information and obtain high-resolution feature $F \in \mathbb{R}^{h \times w}$ (1/4 input resolution). the output feature is fed into two sub-networks including **global estimation sub-network** and **pixel refinement sub-network**, the former follows previous works that use average pooling to generate global feature f_g, then predicts a global initial estimation for shape and rotation parameters. We consider that the global feature mainly contains low-frequency semantics, which impedes to capture local details. To tackle this issue, we propose to use per-pixel prediction fashion as CenterNet [38] in pixel refinement sub-network. Each local feature grid produces a proposal, and generates the corresponding classification score, keypoint coordinates, and rotation parameters respectively. However, we find that the pixel feature is hard to fully encode diverse body deformation with the limited receptive field and context. Thus we propose a 2D Keypoint-Guided Local Encoding Module to enable each pixel to probe a group of local features to refine the global initial estimation.

3.3 2D Keypoint-Guided Local Encoding Module

In detail, to capture fine-grained local features without dense annotations, we leverage 2D keypoint regression guidance to drive each grid feature to search a group of local features corresponding to different body parts following AdaptivePose [34]. Specifically, we divide the human body into several parts and aim to find the corresponding relevant local features f_l with spatial geometry reserved to rectify the global initial prediction.

Towards this goal, we adapt the *Parts as Adaptive points* proposed in AdaptivePose [34] from keypoint localization to 3D mesh recovery, and further build a unified keypoint and mesh estimation framework. For keypoint regression via *Parts as Adaptive Points*, as shown in Figure 3 (b), each divided part is encoded by an unconstrained relevant point. The regression route can be decomposed into two sub-routes. The former predicted by the reference points' feature, starts from a reference position to the part-relevant points. The latter offsets are regressed by part-relevant points' features from the current positions to the objective keypoints. For the training process, the supervisions are performed on the addition of two sub-routes. Due to the regression chains are differentiable, thus the supervisions can drive the unconstrained points to locate on the positions with local part context. Through the 2D keypoint regression guidance, if the keypoint position can be precisely located via the unconstrained points' feature, the encoded local context can predict other joint properties accurately without spatial information compression. Based on the above description, we represent

the human pose into several compositional parts and search the corresponding regional semantic feature with the spatial structure for each part. Instead of aiding by dense annotations, we only utilize the sparse 2D keypoint regression to facilitate the network focus on a relevant spatial position for each part.

In the forward process, We first map the global feature to generate the base estimations, including shape β_b, pose θ_b, and camera parameters. Then we aggregate all part-relevant points' features to predict the residual pose parameter θ_r to rectify the base estimation. By using the 2D keypoint guided features, the 3D SMPL parameter can implicitly fit the 2D evidence without various re-projection losses.

3.4 Dynamic Matching Strategy

In our global-to-local manner, due to the plenty of predictions with only one ground truth for each input image, a naive solution is conducting the heuristic assignments via position priors (e.g., assigning the center area pixels as positives). This idea has been verified in object perception tasks such as anchor-free object detection [29, 38] and pose estimation [4, 33–35]. In the training stage of these methods, due to the input images containing multiple objects or persons with various scales, thus roughly selecting the center positions as positives can achieve promising performance. However, directly employing the above label assignments to our GLNet only achieves unsatisfactory results. In the top-down paradigm, the inputs are cropped according to the bounding box and normalized to the unified scale, the background and the scale variance issues are greatly eliminated. To advance the training of GLNet, we present to design a Dynamic Matching Strategy only based on 2D elements matching for effective training while further ensuring the spatial alignments.

Our framework infers N predictions, where each one is generated from a pixel grid. Let us denote only one ground truth as $y = \{c, kpt^{2D}, pose^{smpl}\}$, where c is the classification label, kpt^{2D} and $pose^{smpl}$ represent 2D keypoint coordinates and SMPL parameters respectively. $\hat{y} = \{\hat{y}_i\}_{i=1}^{N=h\times w}$ represent the set of N predictions. N is the number of pixel positions in the output feature. We assume y also as a group of size N padded with $N-1$ $\emptyset$ (no object). We can perform bipartite matching between these two sets to obtain an index permutation σ with the lowest cost $argmin \sum_{i}^{N} L_{match}(y_i, \hat{y}_{\sigma(i)})$.

Due to the input image being cropped from the raw image, the foreground human body occupies most area of the input image. We observe that only assigning the supervision to one positive sample may cause slow convergence and training collapse. To facilitate the training and improve the stability, we repeat the unique ground truth for T times and pad $N-T$ $\emptyset$ to build the label sets. We denote the annotation of element i as $y_i = \{c_i, kpt_i^{2D}, pose_i^{smpl}\}$ (may be $\emptyset$). For the matching process, an intuitive solution is to consider classification cost, 2D and 3D components costs between each prediction and ground truth simultaneously. However, we found that involving the 3D components in cost computation leads to unstable matching. Ultimately, we only use classification

and 2D keypoint regression loss for calculating the matching costs, which can be formulated as:

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

where $\hat{P}_{\sigma(i)}(c)$ indicates the probability of *person* class.

Then we obtain an optimal matching via the matching cost and calculate the loss for all matched pairs. The predictions matched with $\emptyset$(empty) are regarded as negatives and only supervised by classification loss. The others are positives and supervised by classification, 2D keypoint regression, and 3D pose parameter regression loss as follows:

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

Where $\mathbb{I}_{\{c \neq \emptyset\}}$ is indicator. λ_{cls}, λ_{2D} and λ_{3D} are set to 2, 70, 0.1 experimentally. We leverage Focal loss for classification, L1 loss for 2D keypoint regression, and L2 loss for pose and shape regression. In the test stage, we pick up the final prediction with the max classification score.

4 Experiments

We first briefly introduce our experimental setups in Subsection 4.1. Then we conduct comprehensive comparisons with previous methods to verify the superiority of GLNet in Subsection 4.2. Finally, we carry out the ablation studies to investigate the effectiveness of each component in Subsection 4.3.

4.1 Experimental Setup

Datasets. We train GLNet on a mixture of MS COCO [19], Human3.6M [8], MPI-INF-3DHP [21] and 3DPW [31] with 2D/3D annotations, evaluate the 2D keypoint localization capacity on COCO validation set and 3D mesh recovery on Human3.6M and 3DPW test sets. Specifically, MS COCO [19] is a widely-used 2D pose estimation benchmark that includes 200k images with 250k human instances annotated with 17 body keypoints. We incorporate its training data into 3D data to improve the scene diversity. Human3.6M is an indoor multi-view benchmark for 3D pose estimation. Following previous methods, we use 5 subjects $S1, S5, S6, S7, S8$ for training and $S9, S11$ for evaluation. MPI-INF-3DHP is a more diverse dataset consisting of both constrained indoor and complex in-the-wile scenes. 3DPW is a challenging in-the-wild dataset with 3d pose and shape annotated by 3D IMU devices. We adopt a fixed sampling ratio of 0.35: 0.45: 0.1: 0.1 for the above datasets in the training stage.

Evaluation Metric. The three standard metrics in our experiments are briefly described below. They all measure the Euclidean distances (in millimeters) of 3D points between the predictions and ground truth. MPJPE (Mean Per Joint Position Error) first aligns the prediction and ground-truth at the pelvis and

then calculates their distances. PA-MPJPE (Procrustes-Aligned Mean Per Joint Position Error) performs Procrustes alignment before computing MPJPE, ignoring the discrepancies in scale and global rotation. PVE (Per Vertex Error) does the same alignment as MPJPE at first, then calculates the distances of vertices on the body mesh.

Augmentation. In the training stage, we carry out data augmentation via random flip with a probability of 0.5, random rotation in $[-30, 30]$ degrees, and random scaling of $[0.7, 1.3]$ to augment training samples. Each input is cropped to 256×256 pixels. The feature size in pixel refinements is $1/4$ of the input resolution.

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

Implementation Details. We train our proposed GLNet via Adam optimizer on 4 Tesla V100 GPUs, with a mini-batch size of 128 for 70 epochs. We use HRNet-W32/48 [28] as image encoder. The initial learning rate is set as 2.5e-4 and dropped to 2.5e-5 and 2.5e-6 at the 40th and 60th epochs respectively.

4.2 Comparisons with State-of-the-arts

As shown in Table 1, we report the quantitative results on 3DPW and Human3.6M datasets, then make comprehensive comparisons with previous state-

of-the-art methods. For model-free methods, our GLNet reduces 8.4mm MPJPE and 10.0mm PVE than Graphormer [18] on 3DPW, which estimates the human mesh vertices and body joints directly from the image. For model-based methods, we compare our GLNet with global regression and local regression approaches respectively. Compared with CLIFF-W48 [16], our method with HRNet-W32 reduces 2.7mm MPJPE and 3.5mm PVE without holistic location information. As for local regression methods, GLNet leverages 2D keypoint regression guidance to adaptively fuse local part features, and outperform DaNet [36] which performs joint-centric ROI sampling on proxy IUV map by a large margin. Moreover, compared with PARE [11] and BOPR [1] using the predicted part segmentation masks to obtain regional features, our approach reduces 12.8mm MPJPE and 16.5mm PVE over PARE, 2.5mm MPJPE and 4.0mm PVE than BOPR. GLNet is able to maintain local spatial geometry and dynamics without heuristic rules. Table 1 shows that GLNet achieves superior performance, especially on MPJPE and PVE metrics, which reveal more accurate visual-mesh alignment than previous local regression methods.

4.3 Ablation Analysis

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			Human3.6M	
				PA-MPJPE	MPJPE	PVE	PA-MPJPE	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**

Contributions of each component. We analyze the contribution of each component to the whole framework. The results are shown in Table 2. *Baseline* denotes using global features to predict shape and rotation directly. We first insert 2D Keypoint-Guided Local Encoding Module (body central area as positive samples) that uses part-aware local features to refine the global estimation. We observe the refinement reduces the PA-MPJPE and MPJPE by 5.9 mm, and 2.6 mm on 3DPW respectively. Dynamic Matching Strategy further decreases MPJPE and PVE by 5.6 and 2.6 mm on the 3DPW dataset. The results verify that the 2D Keypoint-Guided Local Encoding Module can capture spatial geometry and local dynamics. Equipped with the 2D evidence-based matching strategy, GLNet can effectively improve the visual-mesh alignments. We also replace the direct rotation regression with twist-swing decomposition. The swing

rotations are calculated by the predicted 2D keypoints and corresponding depth. The twist rotations are estimated by global feature and revised by the fused part-relevant points' features. The result demonstrates employing swing-twist decomposition achieves slightly better performance than direct rotation regression. In twist-and-swing decomposition, rotation denotes twist rotation. In the following studies, we adopt twist-and-swing decomposition by default.

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

annotation	type	3DPW			Human3.6M	
		PA-MPJPE	MPJPE	PVE	PA-MPJPE	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 4: Ablative studies for exploring which type of parameters should combine global estimation and pixel-wise refinements.

Pred Paradigm	Refined Param.	3DPW			Human3.6M	
		PA-MPJPE	MPJPE	PVE	PA-MPJPE	MPJPE
Global	—	45.5	73.9	84.2	33.2	53.5
G-to-P	shape + rot + cam	42.7	71.2	83.4	33.8	52.1
G-to-P	shape + rot	40.6	68.2	79.4	30.6	48.3
G-to-P	rot	39.7	66.3	77.7	29.8	47.5

Analysis of 2D Keypoint-Guided Local Encoding Module. We go deep into the design of the Local Encoding Module. Instead of obtaining local features via dense annotations, our local feature only relies on sparse 2D keypoint guidance with spatial structure kept. As shown in Table 3, we leverage the predicted part segmentation map and IUV map to produce the local features and refine the global initial estimation respectively. For the usage of the IUV map, we follow DaNet [36] that adopt joint-centric ROI pooling to obtain local features. Our 2D keypoint-guided local feature reduces the MPJPE by 5.9, 4.6 mm on 3DPW and Human3.6M compared with the segmentation map, and decreases MPJPE and MPVE by 4.6, 2.8 mm than IUV map on 3DPW. Moreover, in contrast to dense proxy labels, 2D keypoint annotations are easy to obtain and already exist in most 3D human body datasets.

We further investigate how to combine the global initial estimation and pixel-wise refinements to achieve superior performance. We utilize the part-relevant points feature to localize keypoints via the differentiable two-hop regression route. The supervision imposes on whole offsets and thus drives the part-relevant points located on body part aware positions, as shown in Figure 3 (b). Consequently, we fuse the part-relevant points' features to rectify the global prediction of rotation parameters. As shown in Table 4, we observe that the global feature is

sufficient for accurate estimation of camera and shape parameters, further refinements disrupt the initial estimation. Using local features to revise the rotation parameters achieves the best results. 2D keypoint-guided local features involve more fine-grained spatial and detailed information than global feature and thus are beneficial to encode diverse articulation deformations. As shown in Figure 3 (a), the visualizations verify the global-to-local refinements can significantly improve image-mesh alignments.

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.

Analysis of Dynamic Matching Strategy. We design a Dynamic Matching Strategy to assign positive/negative samples instead of position prior and conduct comprehensive experiments on 2D keypoint localization and 3D mesh recovery respectively to validate the effectiveness. In our framework, each pixel produces a set of parameter predictions, we perform one-to-one label assignments first. However, a large number of predictions matched with only one ground truth leading to extremely positive/negative sample imbalance and causing training instability. We repeat the ground truth for T times to increase the positive ratio in the training process. We explore the selection of T in Table 5 and find that 20 is enough to stabilize the training process and avoid collapse.

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

Dataset \ Number	-	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

Moreover, we investigate which components should be considered for the matching process. Intuitively, the cost ought to involve more factors (e.g., classification, 2D keypoints, depth, and rotation parameters) as possible. Nevertheless, we find that 3D elements (e.g., depth, rotation) disturb the matching procedure. As reported in Table 6, matching costs that encompass class prediction and 2D keypoint regression can achieve satisfactory results, further incorporating depth or rotation parameters tends to increase prediction error. We argue that due to the 3D parameters prediction from monocular image is an inherently ill-posed problem, which can lead to unstable matching.

Limitation. Using a fixed number of positive samples is sub-optimal. How to design a mechanism to adjust the number of positive samples based on training feedback remains to be explored.

Table 6: Ablative studies for Matching Cost.

Matching Cost	3DPW			Human3.6M	
	PA-MPJPE	MPJPE	PVE	PA-MPJPE	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

5 Conclusion

In this paper, we propose a global-to-local prediction framework for HMR, which leverages local features with spatial and local information to correct global prediction. First, we propose a 2D Keypoint-Guided Local Encoding Module that leverages sparse 2D keypoint guidance to extract and fuse the body part features with local spatial context. The fine-grained features are capable of tuning the rough global results. Second, we introduce a Dynamic Matching Strategy for the training stage to further improve the visual-mesh alignments. Comprehensive comparisons verify the effectiveness of the two proposed components. GLNet achieves significant improvements over previous local regression methods and obtains state-of-the-art performance on Human3.6M and 3DPW datasets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62227805, Grant 62071056 and S2021222.

References

1. Cheng, Y., Huang, S., Ning, J., Shan, Y.: Bopr: Body-aware part regressor for human shape and pose estimation. arXiv preprint arXiv:2303.11675 (2023)
2. Cho, J., Youwang, K., Oh, T.H.: Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In: European Conference on Computer Vision. pp. 342–359. Springer (2022)
3. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. pp. 769–787. Springer (2020)
4. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14676–14686 (2021)
5. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. arXiv preprint arXiv:2305.20091 (2023)
6. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10884–10894 (2019)
7. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7297–7306 (2018)
8. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013)
9. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In: 2021 International Conference on 3D Vision (3DV). pp. 42–52. IEEE (2021)
10. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7122–7131 (2018)
11. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11127–11137 (2021)
12. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2252–2261 (2019)
13. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6050–6059 (2017)
14. Li, J., Bian, S., Liu, Q., Tang, J., Wang, F., Lu, C.: Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12933–12942 (2023)
15. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3383–3393 (2021)

16. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: Cliff: Carrying location information in full frames into human pose and shape estimation. In: European Conference on Computer Vision. pp. 590–606. Springer (2022)
17. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1954–1963 (2021)
18. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12939–12948 (2021)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
20. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
21. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 international conference on 3D vision (3DV). pp. 506–516. IEEE (2017)
22. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. pp. 752–768. Springer (2020)
23. Muller, L., Osman, A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9990–9999 (2021)
24. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019)
25. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 459–468 (2018)
26. Shetty, K., Birkhold, A., Jaganathan, S., Strobel, N., Kowarschik, M., Maier, A., Egger, B.: Pliks: A pseudo-linear inverse kinematic solver for 3d human body estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 574–584 (2023)
27. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent. In: European Conference on Computer Vision. pp. 744–760. Springer (2020)
28. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019)
29. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
30. Tung, H.Y., Tung, H.W., Yumer, E., Fragkiadaki, K.: Self-supervised learning of motion capture. Advances in neural information processing systems **30** (2017)
31. Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European conference on computer vision (ECCV). pp. 601–617 (2018)

32. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10974 (2019)
33. Xiao, Y., Su, K., Wang, X., Yu, D., Jin, L., He, M., Yuan, Z.: Querypose: sparse multi-person pose regression via spatial-aware part-level query. Advances in Neural Information Processing Systems **35**, 12464–12477 (2022)
34. Xiao, Y., Wang, X.J., Yu, D., Wang, G., Zhang, Q., Mingshu, H.: Adaptivepose: Human parts as adaptive points. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2813–2821 (2022)
35. Xiao, Y., Yu, D., Wang, X.J., Jin, L., Wang, G., Zhang, Q.: Learning quality-aware representation for multi-person pose regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2822–2830 (2022)
36. Zhang, H., Cao, J., Lu, G., Ouyang, W., Sun, Z.: Learning 3d human shape and pose from dense body parts. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2610–2627 (2020)
37. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11446–11456 (2021)
38. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
39. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15085–15099 (2023)