

# Distribution and Depth-Aware Transformers for 3D Human Mesh Recovery

Jerrin Bright\*, Bavesh Balaji\*, Harish Prakash\*, Yuhao Chen\*, David A Clausi\* and John Zelek\*

\* Vision and Image Processing Lab, University of Waterloo, Canada

{jerrin.bright, bbalaji, harish.prakash, yuhao.chen1, dclausi, jzelek}@uwaterloo.ca

**Abstract**—Precise Human Mesh Recovery (HMR) with in-the-wild data is a formidable challenge and is often hindered by depth ambiguities and reduced precision. Existing works resort to either pose priors or multi-modal data such as multi-view or point cloud information, though their methods often overlook the valuable scene-depth information inherently present in a single image. Moreover, achieving robust HMR for out-of-distribution (OOD) data is exceedingly challenging due to inherent variations in pose, shape and depth. Consequently, understanding the underlying distribution becomes a vital subproblem in modeling human forms. Motivated by the need for unambiguous and robust human modeling, we introduce Distribution and depth-aware human mesh recovery (D2A-HMR), an end-to-end transformer architecture meticulously designed to minimize the disparity between distributions and incorporate scene-depth leveraging prior depth information. Our approach demonstrates superior performance in handling OOD data in certain scenarios while consistently achieving competitive results against state-of-the-art HMR methods on controlled datasets.

**Keywords**-Human Mesh Recovery, Depth Ambiguity, Distribution Modeling, Transformers, Residual Likelihood

## I. INTRODUCTION

Monocular Human Mesh Recovery (HMR) is an approach for estimating the pose and shape of a human subject from a single image, featuring a broad spectrum of applications in various downstream tasks [1], [2], [3]. HMR can be split into two types: parametric and non-parametric approaches. The parametric approach involves the modeling of a network to generate model parameters, which are subsequently utilized for human mesh generation, as elucidated in [4], [5], [6]. Recent strides have been witnessed in nonparametric-based approaches [7], [8], which directly regresses the 3D coordinates of the human mesh.

Despite the notable progress in both paradigms, they struggle with two key challenges- the appearance domain gap and depth ambiguity. Controlled environments, often used for training, offer a setting where data collection and annotation are manageable and precise. However, the challenge arises when the trained model is applied to in-the-wild data, where real-world variability, such as lighting conditions, backgrounds, and poses, differs significantly from controlled settings. Second, depth-ambiguity issues plague single-view images. In response to the latter challenge, researchers, as exemplified in [9] and [5], have proposed solutions that leverage temporal information extracted from video inputs to enhance the understanding of human motion.

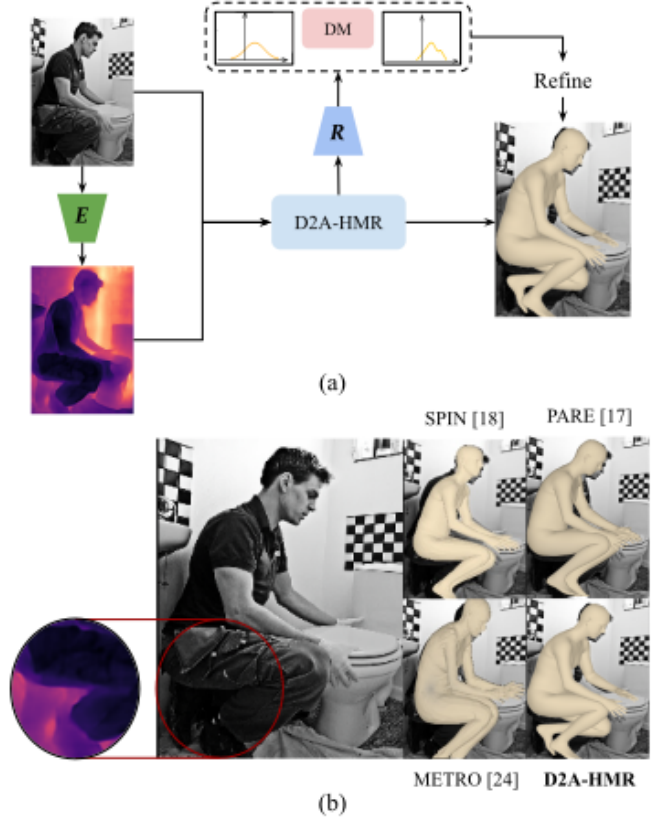


Figure 1: **Illustration of our main idea.** (a) Overview of the proposed D2A-HMR approach (b) Our method, D2A-HMR improves the mesh-image alignment (particularly as visualized in the highlighted region) when compared against SPIN [10], PARE [11] and METRO [8].

However, these temporal approaches have entailed significant computational overhead.

Obtaining ground truth mesh labels for human mesh reconstruction is a tedious task, mainly due to challenges like complexities of dynamic human motion, scene dynamics, resource constraints, and privacy concerns. In response to the inherent difficulty in obtaining accurate ground truth labels, existing works such as [8], [7], [10] resort to using pseudo ground truth to train models. Consequently, the modeling of human forms is inherently biased due to the presence of noisy labels. Moreover, the generalization of HMR for OOD poses, as discussed earlier, is an immensely challenging

problem. Prior works [12], [13] model the output as a distribution of plausible 3D pose using normalizing flows and use information such as 2D keypoints or part segments as priors to provide deterministic predictions for downstream tasks. However, since these models use normalizing flows to explicitly estimate the underlying output distribution, they fail to generalize as shown in [14] and do not solve the model’s bias to actual data, especially in scenarios with noisy labels and uncertainties.

To address the limitations of existing methods, our work introduces a novel approach to address these issues through a depth- and distribution-aware framework designed for the recovery of human mesh from monocular images. Notably, we integrate scene-depth information from monocular cameras obtained from previous depth models (termed *pseudo-depth*) into a transformer encoder via the cross-attention mechanism. In addition, we employ a log-likelihood residual approach to learn deviations in the underlying distribution, facilitating a refinement module in the training process. This distribution approach *explicitly encourages the model* to learn a more generalizable representation that can perform better on unseen data. To further refine the mesh shape and feature relationships, we introduce a dedicated silhouette decoder and a masked modeling module. As showcased in Figure 1, these contributions allow our D2A-HMR approach to excel in *handling challenging, unseen poses*. To the best of our knowledge, D2A-HMR is the only framework to explicitly incorporate depth priors and systematically learn the mesh distribution disparity between the underlying prediction and ground truth distributions. Through experimentation, we demonstrate that our method outperforms existing works on some benchmarked datasets. In summary, our contributions include:

- 1) We introduce a novel image-based HMR model named D2A-HMR that adeptly models the underlying distributions and integrates pseudo-depth priors for efficient and accurate mesh recovery.
- 2) By leveraging residual log-likelihood approach, we refine the model by learning the disparity between the underlying predicted and ground truth distribution.
- 3) Validation of the enhanced performance through the integration of pseudo-depth and distribution-aware modules in HMR, particularly in complex human pose scenarios.

## II. RELATED WORK

**Human Mesh Recovery from a Single Image.** Recent works on HMR can be split into parametric and nonparametric approaches. Parametric approaches can further be split into optimization-based and learning-based approaches. Optimization-based approaches fit a body model by minimizing the error between different prior terms. SMPLify [15] fits the parametric SMPL [16] model to minimize the error between the recovered mesh and keypoints. In addition, prior

terms including silhouettes [6], [1] or distance functions [17] are used to penalize unrealistic shapes and poses. Learning-based approaches take advantage of deep neural networks to predict model parameters [4], [10], [6], [18]. Recent works including HMR-ViT [18] use a transformer-only temporal architecture to predict the model parameters, and ImpHMR [6] uses neural feature fields to model humans in 3D space from a single image.

For directly regressing the vertices, works including GraphCMR [19], Pixel2mesh [20], and Featnet [21] use graphical neural networks to regress the vertices of RGB images, effectively by modeling neighborhood vertex-vertex interactions. Pose2Mesh [7] uses a 2D and 3D pose to regress the vertices using graphical spectral neural networks. METRO [8] uses transformers to model the global interaction between the vertices and I2LMeshNet [22] uses a heatmap-based representation called lixel to regress the human mesh.

**Normalizing Flow.** Normalizing flow is a tool for efficiently transforming a simple distribution into a complex one through a series of invertible transformations [14], [12]. It applies to probability density estimation, which can be used to estimate the likelihood. Previous work including [13] and [23] use normalizing flows to learn a priori the distributions of plausible human poses. ProHMR [12] focuses on modeling the output of the human mesh as a distribution over all the different possible meshes. However, it utilizes normalizing flows to directly predict the exact underlying distribution which is demonstrated to perform poorly for OOD data [14]. RLE [24] uses normalizing flow to minimize the difference between the distributions of the ground truth and predicted 2D poses rather than using the output distribution to sample one particular pose, thereby boosting the performance of regression-based pose estimation techniques.

Inspired by the literature on residual log-likelihood in 2D human pose estimation [24] and the shortcomings of existing HMR approaches, our approach focuses on mitigating distribution discrepancies of the output and ground truth meshes by leveraging normalizing flow techniques. This alleviates the problem of poor performance on OOD data as we use normalizing flows in the refinement module to minimize the difference between the output mesh distribution and ground truth mesh distribution instead of predicting output poses/meshes using the captured output distribution.

**Attention for Human Mesh Recovery.** Attention mechanisms have been shown to be effective for HMR by enabling models to focus on the most relevant parts of the input data. METRO [8] uses self-attention to reduce ambiguity by establishing non-local feature exchange between visible and invisible parts with progressive dimensionality reduction. SAHMR [25] uses cross-attention between image and scene contact information to improve the posture of the regressed mesh. The recently proposed JOTR [26] uses self-attention

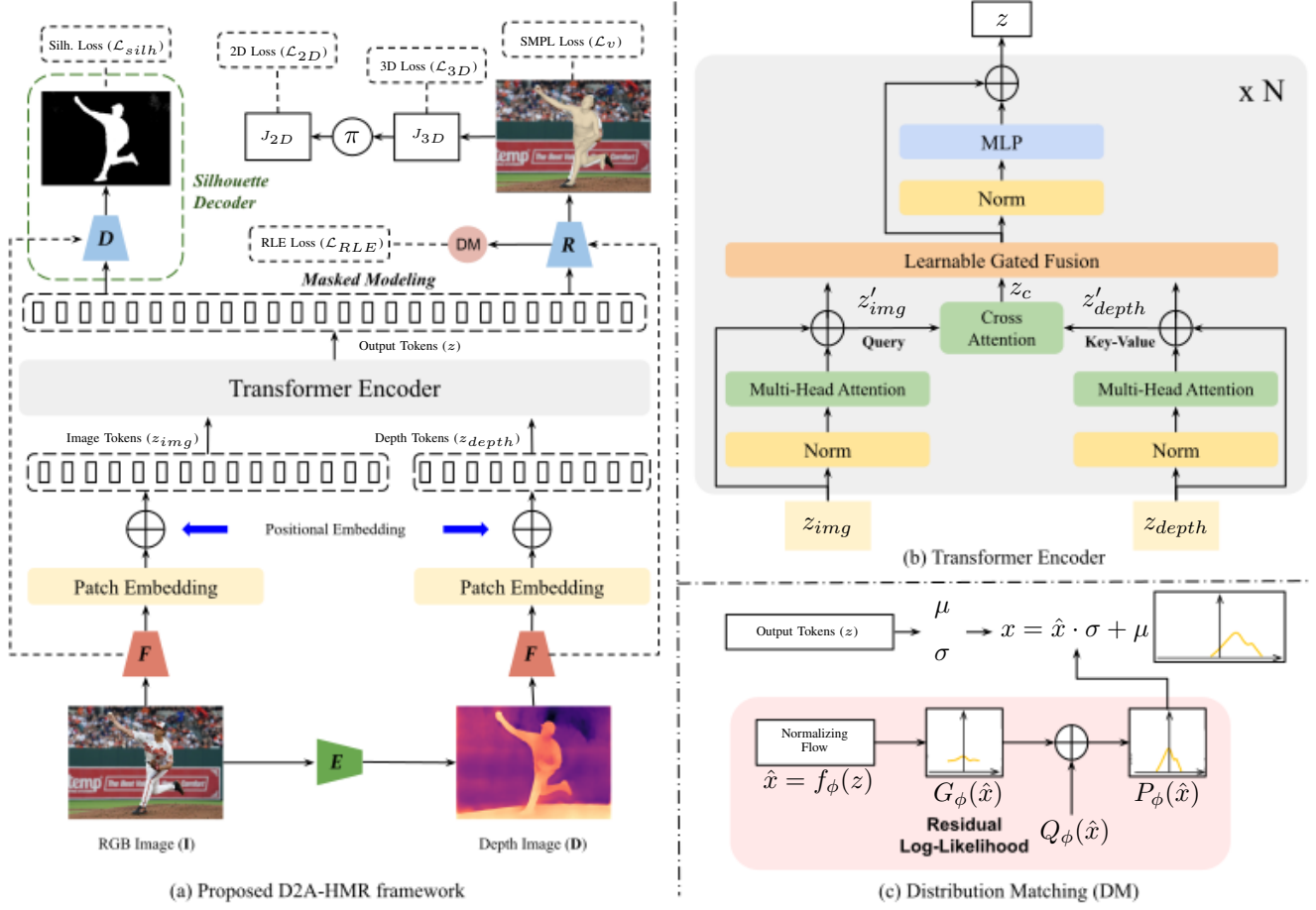


Figure 2: **D2A-HMR model architecture.** Given an image (**I**), we first incorporate a transformer backbone (**E**) to estimate the depth map (**D**) and a CNN backbone (**F**) to extract the features from the images. Positional embedding is applied to both image and pseudo-depth features, utilizing a hybrid approach for image tokens ( $z_{img}$ ) and pseudo-depth tokens ( $z_{depth}$ ). Self-attention is performed on  $z_{img}$  and  $z_{depth}$ , resulting in  $z'_{img}$  and  $z'_{depth}$ , respectively. Subsequently, cross-attention is applied between  $z'_{img}$  and  $z'_{depth}$  to produce  $z_c$ . The learnable fusion gates combine  $z'_{img}$ ,  $z'_{depth}$ , and  $z_c$ , followed by layer normalization and an MLP. The resulting gated tokens ( $z$ ) are input into three distinct refinement modules: a decoder (**D**) for silhouette estimation, a regressor head, **R** which incorporates normalizing flow (**DM**) for distribution-aware joint vertex estimation and masked modeling for enhanced semantic representation of the features.

to study the dependencies of 2D and 3D features to solve problems of occlusion. PSVT [27] uses a spatiotemporal attention mechanism to capture relations between tokens and pose/ shape queries in both temporal and spatial dimensions. Similarly, OSX [28] uses a component-aware encoder to capture the correlation between different parts of the human body to predict the whole-body human mesh.

We propose a parallel network composed of two self-attention modules to learn global dependencies within the image and pseudo-depth features, respectively, and a cross-attention module to learn inter-modal dependencies between the image and pseudo-depth features. This allows the network to learn a more comprehensive representation for accurate 3D mesh recovery.

### III. METHOD

The overview of the proposed D2A-HMR framework is presented in Figure 2. In this section, we delve into the architecture and training objective of D2A-HMR. The feature encoding process begins with the extraction of features from the image and pseudo-depth map using a convolutional neural network (CNN) backbone, followed by hybrid position encoding. These encoded features are then inputted into the transformer encoder, which engages in cross-attention with the pseudo-depth cues and the input image. Following this, the refinement module comes into play, incorporating the distribution matching, silhouette decoder, and masked modeling components to regularize the model during the training process.

### A. Architecture

**Feature Encoding.** The initial step involves passing the input image and depth map through a CNN backbone to extract pertinent features. Subsequently, to explicitly model the structure of the features, position embedding is applied to these extracted features.

Specifically, we implement a hybrid positional encoding ( $P_e$ ) illustrated in Equation (1), for the image and depth tokens. This hybrid approach capitalizes on the strengths of both learnable position embeddings ( $P_l$ ) and sinusoidal position embeddings ( $P_s$ ).  $P_l$  adapts to task-specific positional patterns, proving highly effective in capturing intricate spatial relationships. Meanwhile,  $P_s$  contributes to the globally consistent positional understanding, capturing more information about the position. This combination optimally balances adaptability and global context, yielding fine-grained spatial patterns and general positional relationships.

$$P_e = \omega_1 P_l + \omega_2 P_s \quad (1)$$

where  $\omega_1$  and  $\omega_2$  are learnable parameters controlling the position embedding contribution of both types.

**Transformer Encoder.** The utilization of the transformer encoder in D2A-HMR is driven by the overarching goal of effectively learning pseudo-depth cues from the input data. Using self-attention mechanisms on the encoded features derived from both modalities (image and pseudo-depth map), namely  $z_{img}$  and  $z_{depth}$ , the transformer encoder facilitates understanding of spatial relationships within each domain. Furthermore, we propose to use a cross-attention mechanism to establish intricate connections between the image and pseudo-depth information. The resulting fused representation, denoted as  $z$ , encapsulates rich depth cues, crucial for the subsequent regression of human vertices.

The embedded features, denoted as  $z_{img}$  and  $z_{depth}$ , serve as input tokens to the transformer encoder, embodying our pursuit of learning pseudo-depth cues. Using self-attention mechanisms, the encoder refines  $z_{img}$  and  $z_{depth}$  by capturing spatial relationships within each modality, producing updated features  $z'_{img}$  and  $z'_{depth}$ , respectively. Subsequently, the introduction of a cross-attention mechanism facilitates connections between image and pseudo-depth features. The resulting cross-attended tokens denoted as  $z_c$ , are then fused with  $z'_{img}$  and  $z'_{depth}$  from their respective attention heads, yielding a final fused representation denoted as  $z$ , as illustrated in Equation (2). To facilitate this fusion, learnable fusion gates are employed, similar to the position encoding methodology. These gates adaptively emphasize the importance of each source, enhancing the model's capacity to capture meaningful relationships between the image and pseudo-depth features.

$$z = \omega_3 z'_{img} + \omega_4 z'_{depth} + (1 - \omega_3 - \omega_4) z_c \quad (2)$$

Here, in Equation 2,  $\omega_3$  and  $\omega_4$  are the learnable parameters. Once the fusion is done,  $z$  is normalized and fed as input to an MLP to get the output tokens. This holistic approach enables our model to effectively capture intricate patterns and dependencies within the input image and the 3D information of the scene. A visual illustration of the transformer encoder is shown in Figure 2.

### B. Refinement Module

The refinement module in the D2A-HMR framework encompasses three key components, each designed to enhance the model's capabilities in capturing different aspects of human pose and shape. First, the distribution matching component aids in refining the model's representation by aligning the output mesh distribution to the ground truth mesh distribution. This adaptation enables the model to capture and adapt to inherent variations in the distribution of training data, promoting a more generalized performance that extends beyond the specific characteristics of the training data. The second component, the silhouette decoder, focuses on optimizing the model's capacity to align the shape with the input image by adeptly capturing the outlines of the human subject. This component contributes significantly to the model's ability to refine and improve its representation based on the visual cues present in the input data. Lastly, the masked modeling component serves to empower the model by learning from available information, thereby enhancing its ability to capture long-range relationships among features in the image. This integration ensures that the model can leverage relationships across the entire input, contributing to a more comprehensive understanding of the underlying human pose and shape.

**Distribution Matching.** To align the model with the underlying data distribution, we incorporate the RealNVP [29] normalizing flow mechanism within the D2A-HMR framework. This aims to refine the model by minimizing the discrepancy between predicted and ground truth mesh distributions. The transformer encoder's output tokens  $z$  are passed through a MLP regressor ( $\mathbf{R}$ ), which utilize linear layers to predict the mean  $\mu$  and standard deviation  $\sigma$ , controlling the position and scale of the initially assumed Gaussian distribution. The flow-modeled distribution ( $P_\phi(\hat{x})$ , where  $\hat{x}$  is the predicted mesh) is deconstructed into three essential terms, as expressed in the equation:

$$\log P_\phi(\hat{x}) = \log Q(\hat{x}) + \log \frac{P(\hat{x})}{c \cdot Q(\hat{x})} + \log c \quad (3)$$

The first term,  $\log Q(\hat{x})$ , quantifies the logarithmic probability of the data under the simple distribution. The second term,  $\log \frac{P(\hat{x})}{c \cdot Q(\hat{x})}$ , represents the residual log-likelihood, serving as the distinction between the log-probability of the data under the optimal underlying distribution and the log-



probability under the tractable initial density function. The third term,  $\log c$ , functions as a normalization constant.

**Silhouette Decoder.** To optimize shape alignment, we used a specialized decoder to reconstruct silhouettes. Leveraging features from the transformer encoder, this decoder employs a sequence of deconvolution layers with ReLU activation and dropout, culminating in a fully connected layer. This reconstruction process significantly augments the model’s capability to generate high-quality silhouette representations. To acquire the pseudo-ground truth silhouette of human subjects, we utilize an existing segmentation technique [30].

**Masked Modeling.** Prior works including [31], [8], and [32] have demonstrated the efficacy of masked modeling in elucidating diverse relationships within training datasets, spanning textual, vertex, and image domains respectively. In alignment with these established works, we adopt random masking of the embedded features to recover the vertex of the human body. By deliberately obscuring a percentage of embedded features during training, our model is forced to rely solely on the unmasked features extracted from the image. This enables a comprehensive understanding of both short and long-range relationships among the features, contributing to the overall performance of D2A-HMR framework.

### C. Loss Functions

In this sub-section, we present the comprehensive training objectives employed to recover the human mesh in our model. These objectives consist of a weighted combination of various loss components, each serving a specific role in refining the model’s output.

The loss function  $\mathcal{L}_v$  is computed using the loss metric  $L_1$ , with the aim of minimizing the disparities between the model’s output vertices with the ground truth vertex representation. Simultaneously,  $\mathcal{L}_{3D} = |J_{3D} - J_{3D}^g|$  leverages the same loss metric to optimize the 3D pose by regression ( $J_{3D}$ ) of the output mesh vertices following [8], seeking alignment with the ground truth pose coordinates ( $J_{3D}^g$ ). To enhance the alignment between image and mesh representations, camera parameters are employed to reproject and infer the 2D human pose coordinates ( $J_{2D}$ ) represented with  $\mathcal{L}_{2D} = |J_{2D} - J_{2D}^g|$ , where  $J_{2D}^g$  is the 2D pose ground truth. This reprojected output is refined by applying loss optimization using  $L_1$ .

As mentioned in Section III-B, a distribution matching regularizer is used to penalize the model for predicting outputs that are unlikely under the underlying ground truth distribution. Equation (4) shows the distribution regularizer ( $\mathcal{L}_{RLE}$ ) used in the D2A-HMR architecture.

$$\mathcal{L}_{RLE} = -\log Q(\bar{\mu}_g) - \log G_\phi(\bar{\mu}_g) - \log c + \log \sigma \quad (4)$$

Here, in Equation (4),  $G_\phi(\bar{\mu}_g)$  is the learned residual distribution of the predicted value  $\bar{\mu}_g$  where  $\bar{\mu}_g = (\mu_g - \mu)/\sigma$ . Here,  $\mu_g$  is the ground truth distribution and  $\phi$  is the flow model parameter. Additionally, we incorporate silhouette loss, denoted  $\mathcal{L}_{silh}$ , which regularizes the model by controlling the shape of the reconstructed mesh. The overall objective function is shown in Equation (5), which represents a combination of these individual losses.

$$\mathcal{L} = \lambda_d \mathcal{L}_{RLE} + \lambda_v \mathcal{L}_v + \lambda_{3D} \mathcal{L}_{3D} + \lambda_{2D} \mathcal{L}_{2D} + \lambda_s \mathcal{L}_{silh} \quad (5)$$

where  $\lambda_d$ ,  $\lambda_v$ ,  $\lambda_{3D}$ ,  $\lambda_{2D}$  and  $\lambda_s$  denote the weights attributed to the training objectives concerning the distribution, vertices, 3D pose coordinates, 2D pose coordinates, and silhouettes, respectively.

## IV. EXPERIMENTS

### A. Implementation Details

**Training Details.** Training was carried out on an infrastructure comprising three NVIDIA A6000 GPUs. The network was trained for 500 epochs, with a batch size of 48, and 24 parallel workers. Adam optimizer, configured with a learning rate of  $10^{-4}$  and beta values of 0.9 and 0.99, was used for optimization. The network was designed to output a coarse mesh representation containing 431 vertices. This output was subsequently upsampled [19] to the original mesh’s 6890 vertices, utilizing learnable MLP layers, resulting in the model’s ability to capture fine-grained spatial details.

**Datasets.** Following previous work, we used two prominent 3D human pose estimation datasets, namely 3D Poses in the Wild (3DPW) [38] and Human3.6M [39] to train our D2A-HMR model. For the 3DPW dataset, we follow the standard practice of splitting the dataset into a training set of 22,000 images and a test set of 35,000 images. In the case of Human3.6M, we trained our D2A-HMR model on subjects S1, S5, S6, S7, and S8 and conducted testing on subjects S9 and S11. These data configurations were aligned with the common training and evaluation settings within the domain [8], [5]. The qualitative evaluation of the model was done in Leeds Sports Pose (LSP) [40], and various dedicated sports datasets including the MLBPitchDB dataset [41] and HARPE dataset [42].

**Evaluation Metrics.** In line with established practices from previous research [11], [8], [7], we subjected our model to a comprehensive evaluation using key metrics: mean per joint position error (mPJPE), procrustes-aligned mean per joint position error (PA-mPJPE) and per vertex error (mPVE) in both the 3DPW and Human3.6M datasets. mPVE metric is ignored if the ground truth mesh is not available. All metrics were measured in millimeters (mm), providing a precise assessment of our model’s performance.

Table I: Comparison to state-of-the-art 3D pose reconstruction approaches on 3DPW and Human3.6M datasets. **Bold**: best; Underline: second best.

	Method	Human3.6M		3DPW		
		mPJPE ↓	PA-mPJPE ↓	mPVE ↓	mPJPE ↓	PA-mPJPE ↓
Video	HMMR [5]	-	58.1	139.3	116.5	72.6
	TCMR [33]	62.3	41.1	111.5	95.0	55.8
	VIBE [9]	65.6	41.4	99.1	93.5	56.5
Model-based	HMR [4]	88.0	56.8	-	130.0	81.3
	SPEC [34]	-	-	118.5	96.5	53.2
	SPIN [10]	62.5	41.1	116.4	96.9	59.2
	PyMAF [35]	57.7	40.5	110.1	92.8	58.9
	ROMP [36]	-	-	105.6	89.3	53.5
	HMR-EFT [37]	63.2	43.8	98.7	85.1	52.2
	PARE [11]	76.8	50.6	97.9	82.0	50.9
Model-free	ProHMR [12]	-	41.2	109.6	95.1	59.5
	I2LMeshNet [22]	55.7	41.1	-	93.2	57.7
	Pose2Mesh [7]	64.9	47.0	-	89.2	58.9
	METRO [8]	<u>54.0</u>	<u>36.7</u>	<b>88.2</b>	<b>77.1</b>	<b>47.9</b>
	<b>D2A-HMR (Ours)</b>	<b>53.8</b>	<b>36.2</b>	<u>88.4</u>	<u>80.5</u>	<u>48.4</u>

### B. Main Results

We assess the performance of the proposed D2A-HMR framework by comparing it with established state-of-the-art techniques for HMR. The results, presented in Table I, highlight the competitive performance of our method across various metrics on the Human3.6M and 3DPW datasets. The comparative results demonstrate that the meshes generated by the D2A-HMR framework exhibit superior alignment with the input image. Our method’s adept understanding of pseudo-depth cues and the distribution contributes significantly to improved alignment, particularly in handling challenging input scenarios characterized by depth ambiguities and extreme poses.

Table II: Comparison of D2A-HMR on the MLBPitchDB baseball dataset [41]. **Bold**: best; Underline: second best; Double Underline: third best.

Method	Acc. ↑	mPJPE ↓
HMR [8]	65.9	61.3
SPIN [10]	<u>84.7</u>	<u>32.1</u>
ProHMR [8]	76.1	48.2
ROMP [8]	77.4	48.9
METRO [8]	81.5	37.8
PARE [11]	<u>84.0</u>	<u>33.7</u>
<b>D2A-HMR (Ours)</b>	<b>87.9</b>	<b>30.6</b>

Table II presents a comprehensive comparison between our proposed method and established state-of-the-art HMR techniques, utilizing the MLBPitchDB baseball dataset [41]. Notably, D2A-HMR demonstrates *superior performance* in terms of accuracy and mPJPE on this dataset, which is characterized by *high player motion blur and instances of self-occlusion*.

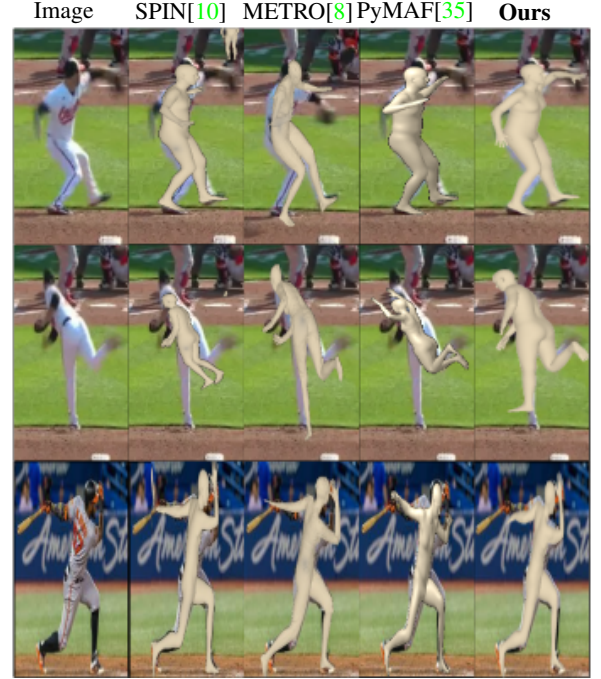


Figure 3: **Qualitative results.** Inferred SMPL mesh reconstruction on the MLBPitchDB baseball dataset [41].

Figure 3 visualizes the effectiveness of our approach in handling these complexities, highlighting its robustness to unseen poses. To further emphasize the efficacy of our proposed approach, we conducted a qualitative comparison against several state-of-the-art techniques on unseen poses, as depicted in Figure 4. This comparative analysis underscores D2A-HMR’s potential for tackling challenging real-world scenarios.



Figure 4: **Qualitative results.** Qualitative comparison of D2A-HMR with SPIN [10], PARE [11], METRO [8], ROMP [36] and PyMAF [35] on in-the-wild data from different sports dataset [41], [42], [40] and unusual poses from the internet.

### C. Ablation Studies

To verify the individual impact of each module on the proposed D2A-HMR model, comprehensive studies were conducted, as detailed in this sub-section. For consistency across all studies, the 3DPW dataset was utilized as the common benchmark.

**Integration of Multi-Modal Data.** Experimentation to assess the impact of depth and distribution matching components within the D2A-HMR are detailed in Table III.

Table III: Ablation study on pseudo-depth and distribution modeling for D2A-HMR evaluated on 3DPW dataset.

Depth	Distribution	mPJPE ↓	PA-mPJPE ↓
✓		92.7	61.8
	✓	90.0	56.9
✓	✓	<b>80.5</b>	<b>48.4</b>

Incorporation of both the pseudo-depth and distribution



modeling modules in the D2A-HMR framework is observed to lead to a substantial improvement in the overall performance of mesh recovery. This observation serves as confirmation that the underlying motivation behind the proposed framework is valid and aids in enhancing the model’s capabilities.

**Depth on mPJPE(z).** Experimentation on exclusively capturing the depth component of the regressed 3D joints in order to demonstrate its impact on the human pose is conducted in Table IV.

Table IV: Ablation study on the impact of depth modeling for D2A-HMR evaluated on 3DPW dataset.

	mPJPE(z) ↓	PA-mPJPE(z) ↓
w/o depth modeling	69.1	58.3
w/ depth modeling	<b>65.4</b>	<b>53.6</b>

A notable enhancement in the z-axis of the reconstructed mesh is evident, as highlighted in Table IV. We computed mPJPE along the z-axis denoted as mPJPE(z), disregarding the components  $x$  and  $y$  of the reconstructed mesh. This experimentation validates that the incorporation of scene-depth information contributes to an improvement in HMR.

**Silhouette and Masked Modeling.** Table V illustrates the impact of the silhouette decoder and masked modeling used within the D2A-HMR framework.

Table V: Ablation study on the silhouette decoder and masked modeling evaluated on 3DPW dataset.

Silhouette	Masked Modeling	mPJPE ↓	PA-mPJPE ↓
✓		89.5	62.2
	✓	84.7	51.4
✓	✓	<b>80.5</b>	<b>48.4</b>

The observations drawn from Table V highlight the beneficial impact of incorporating both the silhouette decoder and masked modeling modules in enhancing the model’s ability to disentangle the appearance and part-relationship of the person. While prior studies, such as [35], employ methods like explicit iterative optimization for mesh-to-image alignment, our silhouette decoder yields improved alignment outcomes compared to scenarios without the decoder. Thus, these modules are utilized during the training process of the D2A-HMR framework, contributing to its improved performance.

**Backbones.** A comprehensive analysis of D2A-HMR’s performance by investigating its behavior with various backbone architectures was conducted. To establish a strong baseline, we first trained two ResNet variants for 1000 epochs on the ImageNet dataset [43] for an image classification task. We also explored HRNet variants trained for 1000 epochs using the COCO dataset [44] for the classification task.

Table VI: Different input representations as the backbone for D2A-HMR evaluated on 3DPW dataset.

Backbone	mPJPE ↓	PA-mPJPE ↓
ResNet50	91.1	59.9
ResNet101	89.5	55.8
HRNet-w40	85.2	52.1
HRNet-w64	<b>80.5</b>	<b>48.4</b>

We observe that HRNet-w64 gives the most positive impact on feature extraction from both the image and depth maps compared to the ResNet backbones. This can be attributed to HRNet-w64’s effectiveness in capturing both local and global contexts through its multiresolution fusion representations, thereby enhancing the model’s ability to extract rich and informative features.

## V. CONCLUSION

In summary, our research introduces the Distribution and Depth-Aware Human Mesh Recovery (D2A-HMR) framework as an innovative solution to the persistent challenge of depth ambiguities and distribution disparities in monocular human mesh recovery. By explicitly incorporating scene-depth information, we have substantially reduced the inherent ambiguity, resulting in a more precise and accurate alignment of human meshes. The utilization of normalizing flows to model the output distribution has been instrumental in regularizing the model to minimize the underlying distribution disparities, enhancing its resilience against noisy labels, and mitigating biases in human-form modeling.

Our extensive experimentation on diverse datasets has demonstrated the competitive performance of the D2A-HMR method when compared to state-of-the-art HMR techniques. Furthermore, it has been noticed that our network outperforms existing work on sports datasets with OOD data. This proposed framework not only addresses depth ambiguities and mitigates noise, but also leverages the inherent 3D information present in images, providing a robust and unambiguous solution for human mesh recovery. Future work will entail training on more diverse datasets to enhance the alignment and generalizability of the HMR process.

## ACKNOWLEDGEMENT

We extend our gratitude to the Baltimore Orioles of Major League Baseball, whose generous support through the Mitacs Accelerate Program played a pivotal role in advancing this research. We also acknowledge the Digital Research Alliance of Canada for their invaluable hardware support.

## REFERENCES

- [1] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black, “Econ: Explicit clothed humans optimized via normal integration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 512–523.



- [2] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9054–9063.
- [3] H. Yi, C.-H. P. Huang, S. Tripathi, L. Hering, J. Thies, and M. J. Black, "Mime: Human-aware 3d scene generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 965–12 976.
- [4] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7122–7131.
- [5] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5614–5623.
- [6] H. Cho, Y. Cho, J. Ahn, and J. Kim, "Implicit 3d human mesh recovery using consistency with pose and shape from unseen-view," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 148–21 158.
- [7] H. Choi, G. Moon, and K. M. Lee, "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*. Springer, 2020, pp. 769–787.
- [8] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1954–1963.
- [9] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5253–5263.
- [10] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "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*, 2019, pp. 2252–2261.
- [11] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "Pare: Part attention regressor for 3d human body estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 127–11 137.
- [12] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis, "Probabilistic modeling for human mesh recovery," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 605–11 614.
- [13] A. Zanfir, E. G. Bazavan, H. Xu, W. T. Freeman, R. Suthankar, and C. Sminchisescu, "Weakly supervised 3d human pose and shape reconstruction with normalizing flows," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 465–481.
- [14] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Why normalizing flows fail to detect out-of-distribution data," *ArXiv*, vol. abs/2006.08545, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:219687356>
- [15] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*. Springer, 2016, pp. 561–578.
- [16] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.
- [17] N. Zioulis and J. F. O'Brien, "Kbody: Towards general, robust, and aligned monocular whole-body estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6214–6224.
- [18] H. Cho, J. Ahn, Y. Cho, and J. Kim, "Video inference for human mesh recovery with vision transformer," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 2023, pp. 1–6.
- [19] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4501–4510.
- [20] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 52–67.
- [21] N. Verma, E. Boyer, and J. Verbeek, "Feastnet: Feature-steered graph convolutions for 3d shape analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2598–2606.
- [22] G. Moon and K. M. Lee, "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*. Springer, 2020, pp. 752–768.
- [23] B. Biggs, D. Novotny, S. Ehrhardt, H. Joo, B. Graham, and A. Vedaldi, "3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 496–20 507, 2020.
- [24] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, and C. Lu, "Human pose regression with residual log-likelihood estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 025–11 034.
- [25] Z. Shen, Z. Cen, S. Peng, Q. Shuai, H. Bao, and X. Zhou, "Learning human mesh recovery in 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 038–17 047.

- [26] J. Li, Z. Yang, X. Wang, J. Ma, C. Zhou, and Y. Yang, "Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9110–9121.
- [27] Z. Qiu, Q. Yang, J. Wang, H. Feng, J. Han, E. Ding, C. Xu, D. Fu, and J. Wang, "Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 254–21 263.
- [28] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li, "One-stage 3d whole-body mesh recovery with component aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 159–21 168.
- [29] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," *CoRR*, vol. abs/1605.08803, 2016. [Online]. Available: <http://arxiv.org/abs/1605.08803>
- [30] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, "Robust high-resolution video matting with temporal guidance," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 238–247.
- [31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423>
- [32] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2536–2544.
- [33] H. Choi, G. Moon, J. Y. Chang, and K. M. Lee, "Beyond static features for temporally consistent 3d human pose and shape from a video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1964–1973.
- [34] M. Kocabas, C.-H. P. Huang, J. Tesch, L. Müller, O. Hilliges, and M. J. Black, "Spec: Seeing people in the wild with an estimated camera," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 035–11 045.
- [35] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun, "Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 446–11 456.
- [36] Y. Sun, Q. Bao, W. Liu, Y. Fu, B. Michael J., and T. Mei, "Monocular, One-stage, Regression of Multiple 3D People," in *ICCV*, 2021.
- [37] H. Joo, N. Neverova, and A. Vedaldi, "Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation," in *3DV*, 2020.
- [38] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 601–617.
- [39] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2014.
- [40] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *bmvc*, vol. 2, no. 4. Aberystwyth, UK, 2010, p. 5.
- [41] J. Bright, Y. Chen, and J. Zelek, "Mitigating motion blur for robust 3d baseball player pose modeling for pitch analysis," 2023.
- [42] M. Fani, H. Neher, D. A. Clausi, A. Wong, and J. Zelek, "Hockey action recognition via integrated stacked hourglass network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 85–93.
- [43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [44] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.