

Noise-in, Bias-out: Balanced and Real-time MoCap Solving

Georgios Albanis ^{1,2}

Nikolaos Zioulis ¹

Spyridon Thermos ¹

Anargyros Chatzitofis ¹

Kostas Kolomvatsos ²

¹Moverse {giorgos,nick,spiros,argyris}@moverse.ai

²Dept. of Informatics and Telecommunications, University of Thessaly kostasks@uth.gr

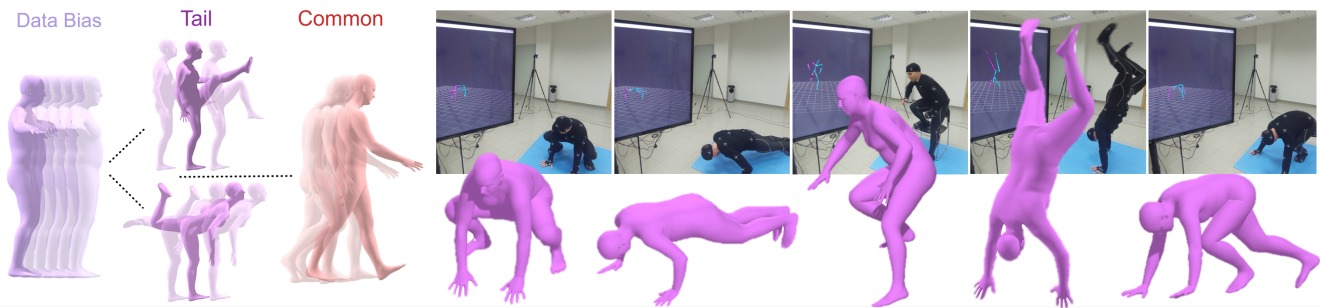


Figure 1: Effective use of human motion data needs to overcome their inherent temporal bias and long-tailed distribution (left). Our model uses a novel balanced regression technique to improve robustness and accuracy to challenging poses, denoise markers and solve joints using raw unstructured marker positions as input. It runs in real-time and can handle higher noise levels (right), producing high-quality body fits even when deployed in a system using just 3 consumer-grade sensors.

Abstract

Real-time optical Motion Capture (MoCap) systems have not benefited from the advances in modern data-driven modeling. In this work we apply machine learning to solve noisy unstructured marker estimates in real-time and deliver robust marker-based MoCap even when using sparse affordable sensors. To achieve this we focus on a number of challenges related to model training, namely the sourcing of training data and their long-tailed distribution. Leveraging representation learning we design a technique for imbalanced regression that requires no additional data or labels and improves the performance of our model in rare and challenging poses. By relying on a unified representation, we show that training such a model is not bound to high-end MoCap training data acquisition, and exploit the advances in marker-less MoCap to acquire the necessary data. Finally, we take a step towards richer and affordable MoCap by adapting a body model-based inverse kinematics solution to account for measurement and inference uncertainty, further improving performance and robustness. Project page: moverseai.github.io/noise-tail.

1. Introduction

Human Motion Capture (MoCap) technology has benefited from the last decade’s data-driven breakthroughs mostly due to significant research on the human-centric visual understanding that focuses on unencumbered capture using raw color inputs. The golden standard of MoCap technology – referred to as “optical” – still uses markers attached to the body, often through suits, for robust and accurate captures, and has received little attention in the literature. These scarce works [22, 19, 18, 13, 26, 12] mainly focus on processing (raw) archival MoCap data for direct marker labeling [19, 18] or labeling through regression [22], solving the skeleton’s joints [13, 12] or transforms [26], while [12] also addressing the case of commodity sensor captures and the noise levels associated with it.

As even high-end systems produce output with varying noise levels, be it either information- (swaps, occlusions, and ghosting), or measurement-related (jitter, positional shifts), these works exploit the plain nature of raw marker representation to add synthetic noise during training. Still, for data-driven systems, the variability of marker placements comprises another challenge that needs to be addressed. Some works [12, 26] address this implicitly,

relying on the learning process, while others [13] address this quasi-explicitly, considering them as input to the model. Another way to overcome this involves fitting the raw data to a parametric model after manually [22, 37, 41], or automatically [19, 18] labeling and/or annotating correspondences, standardizing the underlying representation.

In this work, we explore the next logical step stemming from prior work, bridging standardized representations and consumer-grade sensing, and delivering real-time data-driven MoCap that is robust to tracking errors. Most works [18, 13, 19, 26, 12] leverage high-end MoCap to acquire training data, a process that is expensive, laborious and difficult to scale, apart from [22] that used low-cost sensor acquired data, but nonetheless, applied the model to a high-end capturing system. Instead, by relying on a standardized representation using a parametric human body model, we benefit from modern markerless capture technology, greatly increasing data acquisition rates at a fraction of the costs and labor. Still, there are certain challenges that need to be addressed, such as the distribution of MoCap data and the input optical sensing noise.

The nature of human motion, albeit high-dimensional, instills a significant level of data redundancy in MoCap datasets. Indeed, standing still or walking poses dominate most captures and affect the training data distribution in two ways. First by introducing bias in the learning process, and second, by further skewing the long-tailed distribution. The latter is an important problem [58] that data-driven methods need to overcome as rare poses exist, not only due to their reduced appearance frequency, but also due to biomechanical limitations of the captured subjects in fast movements, body balancing, and striking challenging poses. Prior work crucially neglects this, resorting to uniform temporal down-sampling, which only helps in reducing data samples, yet not redundancy nor long-tailed distribution.

Another typical assumption is that the raw marker data are relatively high quality, most common to labeling works [18, 19] that solve using the raw positions. Even though synthetic noise is added during training, this is mostly to regularize training as the noisy nature of inputs is not taken into account post-labeling. Those works that directly infer solved estimates [12, 13, 26] solely rely on the model’s capacity to simultaneously denoise the inputs and solve for the joints’ positions. Nonetheless, even the models’ outputs are uncertain, a situation that will be increasingly magnified when the raw marker input is affected by higher noise levels, as common when relying on consumer-grade sensors. This lack of solutions that increase noise robustness hinders the adoption of more accessible sensing options.

To that end, we present techniques to address MoCap dataset challenges as well as noisy inputs, resulting in a MoCap framework that **does not** necessarily require data from high-end MoCap systems, **does not** require additional data

to boost long-tail performance, and **does not** require specialised hardware. More specifically we:

- Leverage representation learning to jointly oversample and perform utility-based regression, addressing the redundancy and long-tailed MoCap data distribution.
- Introduce a noise-aware body shape and pose solver that models the measurement uncertainty region during optimization.
- Demonstrate a real-time inference capable and artifact-free MoCap solving model, running at 60Hz on a system comprising just 3 consumer-grade sensors.
- Harness a human parametric representation to cold-start data-driven optical MoCap models using data through markerless acquisition methods.

2. Related Work

2.1. MoCap Solving

Solving the joints’ positions or transforms from marker data is a cascade of numerous (sometimes optional) steps. The markers need to be labeled, ghost markers need to be removed, occluded markers should be predicted and then an articulated body structure needs to be fit to the observed marker data. Various works address errors at different stages of MoCap solving, with contemporary ones relying on smoothness and bone-related (angles, offsets and lengths) constraints [24, 57, 28, 6, 16, 44, 64]. Recent approaches started resorting to existing data for initialization [60] or marker cleaning [5]. MoSh [37] moved one step ahead and instead of relying on plain structures employed a parametric human body to solve labeled marker data and estimate pose articulation and joint positions, even accounting for marker layout inconsistencies and/or soft tissue motion.

Nonetheless the advent of modern – deep – data-driven technologies have stimulated new approaches for MoCap solving. A label-via-regression approach was employed in [22] where a deep model was used to regress marker positions and then perform maximum assignment matching for labeling the input. Labeling was also formulated as permutation learning problem [19], albeit with constraints on the input, which were then relaxed in [18] by adding a ghost category. However, labeling assumes that the raw data are of a certain quality as the raw measurements are then used to solve for the joints’ transforms or extra processing steps are required to denoise the input.

Consequently, end-to-end data-driven approaches that can simultaneously denoise and solve have been a parallel line of research. While end-to-end cleaning and solving is possible using solely a single feed-forward network [26], the process naturally benefits from using two cascaded autoencoders [53], the first operating on marker data and

cleaning them for the subsequent joint regressor. The staging from markers to joints was also shown to be important from a performance perspective in [12] which trained a convolutional network with coupled noisy and clean data captures to address noisy inputs. Recently, graph convolutional models were employed in [13] allowing for the explicit encoding of marker layout and skeleton hierarchy, two crucial factors of variation that were only implicitly handled in prior end-to-end solvers.

2.2. MoCap Data

Learning to solve MoCap marker data requires supervision provided by collecting data using professional high-end MoCap systems [26, 18, 13, 12]. SOMA [18] standardized the representation using the AMASS dataset [41] which, in turn, relied on an extension of MoSh [37] to fit a parametric human body model to markers. All other works suffer from inconsistent marker layouts which is a problem that was either implicitly addressed [26, 12] or quasi-explicitly [13] using the layouts as inputs. Marker data can be (re-)synthesized in different layouts when higher-level information is available (*e.g.* marker-to-joint offsets, meshes) [26, 18]. Yet, it has been also shown that fitting a synthetic hand model to depth data acquired by consumer-grade sensors can also produce usable training data [22] for deploying a model to a high-end marker capturing system for data-driven MoCap. Compared to [22], we experimentally demonstrate this feasibility and even extend it to noisy inputs at run-time, something not considered in [22] as it relied on a high-end system for live capture.

Statistical parametric models [38, 52, 49, 50, 75, 77, 4, 78] are more expressive alternatives than the skinned mesh [73] used in [22] as, apart from realistic shape variations, deformation corrective factors can also be employed. They have been used to synthesize standardized training data before [72, 25, 33] but crucially rely on preceeding high-end MoCap acquisition. We also explore this path using multi-view markerless capture [29, 14, 82] to produce parametric model fits and synthesize marker positions as a solution to the cold-start problem of data-driven MoCap solving. Even though such data can be fit to marker data as done in AMASS [41] and Fit3D [17], the potential of acquiring them using less expensive capture solutions is very important, as long as it is feasible to train high quality models.

Still, one also needs to take into account the nature of human performance data and their collection processes. As seen in AMASS [41] and Fit3D [17], both contain significant redundancies and suffer from the long-tail distribution effect. Rare poses are challenging for regression models to predict, mainly stemming from the combined effect of the selected estimators and stochastic optimization with mini-batches. Various solutions have been surfacing in the literature, some tailored to the nature of the problem [58], lever-

aging a prototype classifier branch to initialize the learned iterative refinement, and others adapting works from imbalanced classification to the regression domain. Traditional approaches fall into either the re-sampling or re-weighting category, with the former focusing on balancing the frequency of samples and the latter on properly adjusting the parameter optimization process. Re-sampling strategies involve common sample under-sampling [69], rare sample over-sampling by synthesizing new samples via interpolation [71], re-sampling after perturbing with noise [9], and hybrid approaches that simultaneously under- and over-sample [10]. Yet interpolating high-dimensional samples like human pose is non-trivial or even defining the rare samples that need to be re-sampled.

Utility-based [70] – or otherwise, cost-sensitive – regression assigns different weights – or relevance – to different samples. Defining a utility function is also essential to re-sampling strategies for regression [69]. Recent approaches employ kernel density estimation [65], adapt evaluation metrics as losses [63], or resort to label/feature smoothing and binning [79]. Another family of methods that are now explored can be categorized as contrastive, with [20] regularizing training to enforce feature and output space proximity. BalancedMSE [55] is also a contrastive-like objective that employs intra-batch minimum error sample classification using a cross-entropy term that corresponds to an L2 error from a likelihood perspective. However, most approaches rely on stratified binning of the output space using distance measures that lose significance in higher dimensions. Further, binning can only be used with specific networks/architectures (proper feature representations for classifying bins or feature-based losses). It has not been shown to be applicable in high-performing dense networks relying on heatmap representations. Instead, we introduce a novel technique that can jointly over-sample and assign higher relevance to rare samples by leveraging representation learning and its synthesis and auto-encoding traits.

3. Approach

The MoCap representation we use is a parametric human body model \mathcal{B} . Different variants exist, all data-driven, some relying on stochastic representations [77], others on explicit ones [38, 49], with a notable exception using an artist-made one [78] and all typically employ linear blend skinning [30] and pose corrective factors [38, 77] to overcome its artifacts. Generally, we consider it as a function $(\mathbf{v}, \mathbf{f}) = \mathcal{B}(\beta, \theta, \mathbf{T})$, where (\mathbf{v}, \mathbf{f}) are the vertices $\mathbf{v} \in \mathbb{R}^{V \times 3}$ and faces $\mathbf{f} \in \mathbb{N}^{F \times 3}$ of a triangular mesh surface that is defined by S blendshape coefficients $\beta \in \mathbb{R}^S$, articulated by P pose parameters $\theta \in \text{SO}(3)^P$, and globally positioned by the transform $\mathbf{T} \in \text{SE}(3)$. Using linear functions r expressed as matrices \mathbf{R} it is possible to extract L different body landmarks $\ell := r(\mathbf{v}) = \mathbf{R} \times \mathbf{v}$, with $\ell \in \mathbb{R}^{L \times 3}$

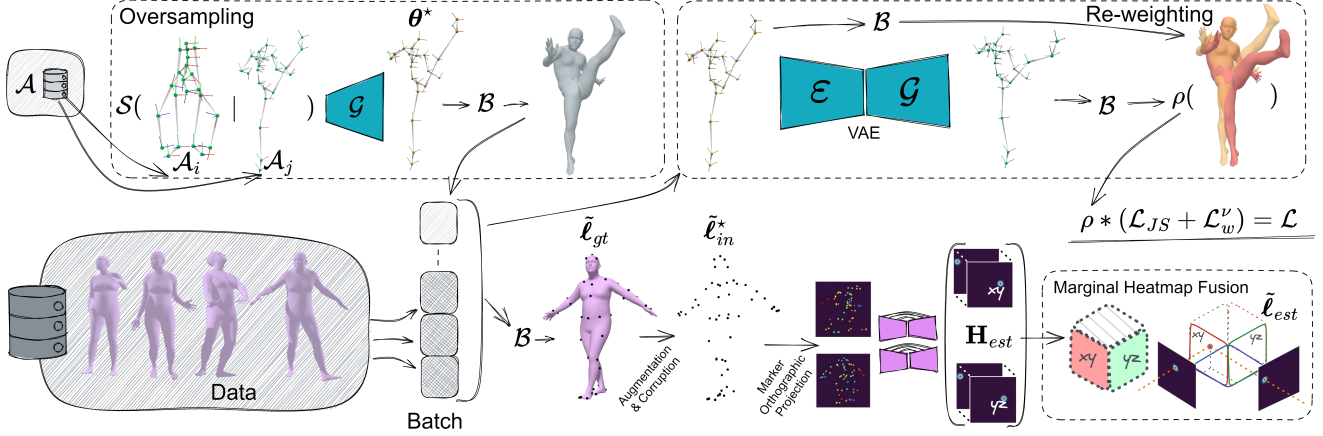


Figure 2: Overview of the balanced and real-time MoCap solving training model. Starting from an existing data corpus (*bottom left*), a set of encoded tail anchor poses \mathcal{A} are selected (Sec. 3.1 - *top left*) and randomly blended via \mathcal{S} and a generator \mathcal{G} . This oversamples the tail, adding extra synthetic rare samples during training. A UNet model (Sec. 3.2 - *bottom middle*) receives two orthographic depth map renders (xy and yz planes) of augmented and corrupted marker 3D positions ℓ_{in}^* extracted from the body’s \mathcal{B} surface, producing 2 orthogonal heatmaps which are marginally fused along the y coordinate, producing 3D positions $\tilde{\ell}_{est}$ (Sec. 3.2 - *bottom right*). The loss for each batch item is re-weighted by its relevance ρ , computed after calculating the joint reconstruction error of its pose’s θ generative autoencoder reconstruction (Sec. 3.1 - *top right*).

and $\mathbf{R} \in \mathbb{R}^{L \times V}$. This way, surface points ℓ^v can be extracted using delta (vertex picking) or barycentric (triangle interpolation) functions and joints ℓ^j using weighted average functions. Since markers are extruded by the marker radius d they correspond to $\ell^m = \ell^v + d(\mathbf{R} \times \mathbf{n})$, with \mathbf{n} being the vertices’ normals.

Following prior art [18], the input data are the parameters of a body model that synthesize markers, which due to their synthetic nature can be augmented, and corrupted with artifacts and noise [26, 13, 12]. Fig. 2 illustrates our model’s training framework which is followingly explained starting with the technique addressing the redundancy and long-tailed nature of the data (Sec. 3.1), the marker denoising and joint solving model’s design choices (Sec. 3.2), and finally the noise-aware body parameter solver (Sec. 3.3).

3.1. Balancing Regression

Relevance functions drive utility regression and guide the re-/over-/inter-sample selection/generation [10, 69, 70, 71]. Instead of defining relevance or sample selection based on an explicit formula or set of rules, we employ representation learning to learn it from the data. Autoencoding synthesis models [36, 56] jointly learn a reconstruction model as well as a generative sampler:

$$\theta^\dagger = \mathcal{G}(\mathcal{E}(\theta)), \quad \theta^* = \mathcal{G}(\mathcal{S}(\cdot)), \quad (1)$$

with varying constraints on the input θ and latent $\mathbf{z} = \mathcal{E}(\theta)$, $\mathbf{z} \in \mathbb{R}^Z$ spaces. An encoder $\mathcal{E}(\theta)$ maps input θ to a latent space \mathbf{z} which gets reconstructed to θ^\dagger by a generator $\mathcal{G}(\mathbf{z})$. Using a sampling function \mathcal{S} to sample the latent

space it is also possible to generate novel output samples θ^* . We exploit the hybrid nature of such models to design a novel imbalanced regression solution that simultaneously over-samples the distribution at the tail and adjusts the optimization by re-weighting rarer samples. Our solution is based on a deep Variational AutoEncoder (VAE) [36].

Relevance via Reconstructability. Autoencoding models are expected to reflect the bias of their training data, with redundant/rare samples being easier/harder to properly reconstruct respectively. This bias in reconstructability can be used to assign relevance to each sample as those more challenging to reconstruct properly are more likely to be tail samples. We define a relevance function ρ (see Fig. 2 re-weighting) using a reconstruction error ϵ :

$$\rho(\theta) = 1 + \exp(\epsilon/\sigma), \quad \epsilon = \sqrt{\frac{1}{J} \sum_{i=1}^J \|\bar{\ell}_i^j - \bar{\ell}_i^{\dagger j}\|_2}, \quad (2)$$

with $(\bar{\cdot})$ denoting unit normalization using the input joints’ bounding box diagonal, ϵ the normalized-RMSE over the reconstructed and original joints, and σ a scaling factor controlling the relevance ρ . Using landmark positions we can preserve interpretable semantics in ρ and σ as they are unidirectionally interchangeable (linear mapping) with the pose θ given fixed shape β . Fig. 3 shows exemplary poses as scored by our relevance function.

Balance via Controlled Synthesis. Even though the tail samples are not reconstructed faithfully, the generative and disentangling nature of modern synthesis models shape manifolds that map inputs to the underlying factors of data

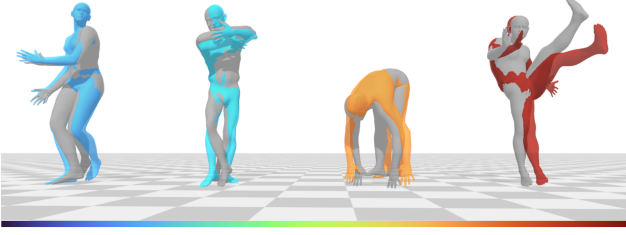


Figure 3: Color-coded (turbo colormap [45] at the bottom) autoencoding relevance ρ of various poses.

variation, effectively mapping similar poses to nearby latent codes which can be traversed across the latent space dimensions. Based on this, we define a controlled sampling scheme for synthesizing new tail samples (see Fig. 2 oversampling). Using the relevance function from Eq. (2), it is possible to identify tail samples θ^\dagger via statistical thresholding that serve as anchor latent codes $\mathcal{A} = \{\mathbf{z}^\dagger \mid \mathbf{z}^\dagger = \mathcal{E}(\theta^\dagger)\}$. This process adapts to the training data distribution instead of risking a mismatch via empiric manual picking when using a purely generative model (e.g. [68]). We then sample using the following function:

$$S_{i,j}(\cdot) = \varsigma(\mathcal{N}(\mathbf{a}_i, \mathbf{s}), \mathcal{N}(\mathbf{a}_j, \mathbf{s}), b), \quad \mathbf{a}_{i,j} \in_R \mathcal{A}. \quad (3)$$

Specifically, we sample from a normal distribution centered around two random anchors i and $j, i \neq j$, from \mathcal{A} using a standard deviation \mathbf{s} , and blend them using spherical linear interpolation [61] ς with a uniformly sampled blending factor $b \in \mathcal{U}(0, B)$, $B \in [0, 1]$. Non-linear interpolation between samples avoids dead manifold regions as not all directions lead to meaningful samples [31, 32] and increases our samples' plausibility [76], as illustrated in Fig. 4.

3.2. Real-time Landmark Estimation

Compared to pure labeling [18, 19] or pure solving approaches [26, 13] we design our model around simultaneous denoising, solving and hallucination. While some approaches use the raw marker positions as input, we opt to leverage the maturity of structured heatmap representations and employ a convolutional model, similar to [22, 12] instead of relying on unstructured regression [13, 26] using MLPs. This improves the convergence of the model and by using multi-view fusion we can also improve accuracy via robust regression. First, we augment and corrupt the input markers ℓ_{gt} into $\tilde{\ell}_{in}^*$. Then, we normalize and render $\tilde{\ell}_{in}^*$ from two orthographic viewpoints as in [12], but with a notable difference when processing the model's output; instead of predicting the 3rd dimension, we manage to predict normalized 3D coordinates by learning to solve a single 2D task. To achieve that, we use the two rendered views as input to the model, predict the corresponding

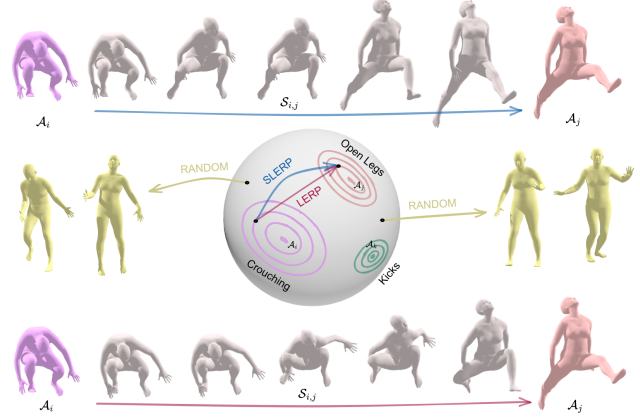


Figure 4: Tail oversampling using latent anchors \mathcal{A} . Random latent vector blending using **non-linear** interpolation generates diverse and realistic tail samples, compared to the **linear** one which produces less diverse or unrealistic samples, or to **random** sampling which produces more biased samples.

view's heatmaps, and fuse them with a variant of marginal heatmap regression [47, 80] (see Fig. 2 fusion). We assume the gravity direction along the y axis and use the orthogonal and orthographic views denoted as xy and yz which share the y axis. To estimate the landmarks' normalized positions $\tilde{\ell}_{est}$, we employ center-of-mass regression [40, 66, 46, 67] taking the average expectation [47, 80] for y from the two views. The model is supervised by:

$$\mathcal{L} = \rho(\lambda \mathcal{L}_{JS}(\mathbf{H}_{gt}, \mathbf{H}_{est}) + \mathcal{L}_w^\nu(\tilde{\ell}_{gt}, \tilde{\ell}_{est})), \quad (4)$$

where \mathcal{L}_{JS} is the λ -weighted Jensen-Shannon divergence [43] between the normalized ground truth and soft-max normalized predicted heatmaps, while \mathcal{L}_w^ν is the robust Welsch penalty function [27, 15], with the support parameter ν , between the normalized landmark ground-truth $\tilde{\ell}_{gt}$ and estimated $\tilde{\ell}_{est}$ coordinates. Overall, \mathcal{L}_{JS} accelerates training while \mathcal{L}_w^ν facilitates higher levels of sub-pixel accuracy since even though we reconstruct the heatmaps \mathbf{H} using the normalized – un-quantized – coordinates [83], discretization artifacts can never be removed entirely.

Note that the fusion outcome $\tilde{\ell}_{est}$ comprises both marker and joint estimations, essentially estimating a complete, labeled, and denoised marker set, as well as solving for the joints' positions. Finally, we use U-Net [59] as a regression backbone for its runtime performance and its efficiency in high-resolution regression.

3.3. Noise-aware Fitting

Given the denoised and complete set of landmarks $\tilde{\ell}_{est} \in \mathbb{R}^{L \times 3}$, we can fit the body to these estimates and obtain the pose θ and shape β which implies an articulated skeleton

and mesh surface. This is a non-linear optimization problem with the standard solution being MoSh [37] and its successor MoSh++ [41]. However, MoSh(++) also solves for the marker layout which in our case is known apriori as the model was trained with a standard 53 marker configuration. Compared to prior works that assume the estimates are of high-quality or low signal-to-noise ratios, we seek to relax this assumption to support additional sensing options. The solution to this is robust optimization but typical approaches that involve robust kernels/estimators require confident knowledge about the underlying data distribution. This is not easily available in practice, and moreover, it varies with different sensing options but more importantly, when involving a data-driven model, it is skewed by another challenging-to-model distribution. The Barron loss [7] is a robust variant that also adapts to the underlying distribution and interpolates/generalizes many known variants by adjusting their shape and scale jointly.

Following likelihood-based formulations [34, 21] that have been presented for multi-task/robust stochastic optimization, we formulate a noise-aware fitting objective that is adaptive and optimizes the Gaussian uncertainty region $\sigma \in \mathbb{R}^L$ jointly with the data and prior terms:

$$\underset{\theta^*, \beta^*, \mathbf{T}^*, \sigma^*}{\operatorname{argmin}} \mathcal{E}_{data} + \mathcal{E}_{prior}. \quad (5)$$

We use standard prior terms [37, 41, 52] $\mathcal{E}_{prior} = \lambda_\beta \sum \|\beta\|_2 + \lambda_z \sum \|\mathbf{z}\|_2$, and a data term formulated as:

$$\mathcal{E}_{data} = \sum_i^L \frac{1}{2\sigma_i^2} \|\tilde{\ell}_{est,i} - \tilde{\ell}_i^*\|_2 + \log \sigma_i. \quad (6)$$

As in MoSh(++) we perform staged annealed optimization but with only 2 stages as there is no marker layout optimization. The first stage optimizes over $\beta^*, \theta^*, \mathbf{T}^*$, while the second stage fixes β and \mathbf{T} and optimizes θ^*, σ^* .

4. Results

We base our implementation on the SMPL(-X) body model \mathcal{B} [38, 52]. Our models are implemented using PyTorch [51], optimized with Adam [35], initialized with Kaiming init. [23], and trained for a fixed number of epochs and with a fixed seed, with the best parameters selected using the performance indicators presented in Sec. 5 of the supplement. UNet receives 160×160 depth maps and outputs heatmaps of the same resolution for all landmarks (53 markers and 18 joints in all cases apart from the experiments in Tab. 3 where 56 markers and 24 joints are used for consistency). The autoencoding generator is implemented as a robust variant of VPoser [52]¹. To fit the body to the estimated landmarks we use quasi-Newton optimization [48].

¹Description and comparison can be found in Sec. 7.1 of the suppl.

	RMSE↓	PCK1↑	PCK3↑	PCK7↑
Optical#1	50.4 mm	36.14%	84.89%	90.90%
Optical#2	89.9 mm	41.11%	81.18%	86.24%
Optical#3	92.9 mm	39.16%	79.74%	86.08%
Markerless	59.4 mm	21.70%	79.96%	90.08%

Table 1: Markerless vs optical data tested on ACCAD.

For the evaluation, the $\tilde{\ell}_{est}$ are denormalized to ℓ_{est} . Finally, the Tables are color-coded with the best result being visualized in pink and bolded, the second in green, and the third (where it is needed) in yellow.

We use a variety of datasets that provide corresponding parametric body \mathcal{B} parameters from which we can extract input (markers) and ground truth (joints and markers). We additionally curate a custom test set comprising 4 categories of tail samples. Note that all models' performance is validated using *unseen* data comprising entire datasets, thus, ensuring different capturing contexts. For a lack of space, we moved all preprocessing (see supp. Sec. 3), datasets (see supp. Sec. 4), and metrics (see supp. Sec. 5) details in the supplement, as well as an in-the-wild [supp. video](#).

Are high-end MoCap data necessary?

Relying on an intermediate body model \mathcal{B} representation opens up new opportunities for data acquisition. We seek to validate the hypothesis that training an optical MoCap model does not necessarily require data acquired by high-end optical MoCap systems. Recent multi-view datasets [82, 14, 54] rely on markerless capturing technology to fit parametric body models to estimated keypoint observations. We train our model (without the imbalanced regression adaptation) on the combined GeneBody [14] and THuman2.0 (TH2) [82] multi-view marker-less data (*Markerless*), and on 3 high-end MoCap dataset combinations from AMASS [41], specifically, EKUT [42], HumanEva [62], MoSh [37], and SOMA [18] (*Optical #1*); CNRS and HumanEva (*Optical #2*); and, solely HumanEva (*Optical #3*) to progressively reduce the diversity of the samples. We equalize the different markerless and optical training data via temporal downsampling to a total of 9mins of MoCap. By evaluating these models using ACCAD [2] (see Tab. 1), we observe a correlation between pose diversity and performance, and that the markerless data result in comparable performance to the high-end MoCap data. The latter indicates that it is possible to acquire data for optical MoCap without having access to any high-end system.

		RMSE↓	PCK1↑	PCK3↑	PCK7↑
TH2	Base	21.4 mm	28.69%	92.08%	98.60%
	[55]	22.0 mm	25.51%	91.90%	98.62%
	Ours	19.1 mm	32.38%	93.55%	99.11%
Tail	Base	35.8 mm	22.04%	80.27%	94.31%
	[55]	32.9 mm	27.66%	81.98%	94.92%
	Ours	29.3 mm	23.42%	84.70%	97.24%

Table 2: Imbalanced regression results.

	RMSE↓	JPE↓	PCK1↑	PCK3↑	PCK7↑
[13]	21.1 mm	17.4 mm	38.11%	84.70%	99.17%
[12]	27.0 mm	17.5 mm	51.08%	89.39%	97.24%
Ours	20.1 mm	15.9 mm	50.14%	92.23%	98.14%

Table 3: Direct joint solving on CMU test set [11].

Addressing the bias and long-tail

To evaluate our novel imbalanced regression discussed in Sec. 3.1, we design an experiment simulating a progressive data collection process by aggregating the Dfaust [8], EYES [39], EKUT, HumanEva, MoSh, PosePrior [3], SFU [81], SOMA, SSM, and Transitions parts from AMASS, captured with varying acquisition protocols and settings. Tab. 2 presents the results compared to a baseline model trained without re-weighting/oversampling, and the BMSE [55] imbalanced regression loss, which is properly adapted to consider joint distances and not scalars.

Tab. 2 (top) presents the results on TH2, a dataset of diverse static poses that also includes challenging poses (e.g. extreme bending, inversion, etc.), where our approach improves overall performance compared to BMSE that presents inferior results to the baseline model. Tab. 2 (bottom) presents the results on our “tail” (rare) poses that include “high kicks”, “crouching”, “crossed arms”, and “crossed legs”. Both imbalanced regression approaches improve the long-tail performance, with our oversampling and re-weighting method achieving the best results almost horizontally. These results highlight that our approach overcomes the known weakness of the BMSE balancing the data distribution at the expense of performance on more common poses. Ablation experiments showcasing the orthogonality of oversampling and re-weighting can be found in the supplementary material (Sec. 7.2, Tab. 4).

Direct joint solving

We proceed with evaluating our model’s ability to accurately estimate the skeleton joints ℓ^j from the

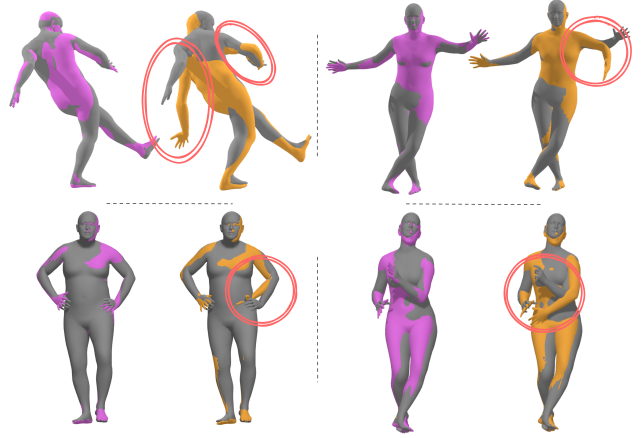


Figure 5: Fits to our regressed vs SOMA labeled markers. Incorrect labeling results in highly erroneous fits.

input markers (i.e. joint-solving). We compare our model against two SotA joint-solving approaches: a) MoCap-Solver [13] that uses graph convolutions and temporal information, and b) DeMoCap [12] that employs an HRNet [74] backbone and frontal-back fusion. All models are trained and evaluated on the CMU [11] dataset as in [13]. For MoCap-Solver we rerun the evaluation without normalizing the markers and the skeletons as this information should be unknown during testing. At the same time, we employ the joint position error (JPE) from [13] for a more fair comparison. From the results in Tab. 3 we observe that our model outperforms the SotA in both positional metrics (RMSE, JPE) while having the best or the second-best accuracy for different PCK.

Explicit vs implicit labeling

Our next experiment aims to showcase the advances of fitting a parametric body model on landmarks estimated with regression instead of explicitly labeling them. We compare our model that de-noises, completes, and implicitly labels landmarks via regression with SOMA, a SotA explicitly labeling method, by fitting the body to the markers similar to [37]. Note that in order to have a fair comparison we solve **only for markers** and not for markers & joints (as discussed in Sec. 3.2). We train our model using the same datasets that SOMA was trained on, and then test on TH2 and our “Tail” test set using the clean body-extracted markers, and the same MoSh-like fitting without uncertainty region optimization and without considering latent markers as the marker layout is fixed to the nominal one. Tab. 4 showcases

		RMSE↓	MAE↓	PCK1↑	PCK3↑	PCK7↑
TH2	[18]	29.7 mm	3.49°	28.33%	87.78%	96.11%
	Ours ($\ell^{m,*}$)	19.1 mm	2.68°	26.49%	93.72%	99.26%
	Ours (ℓ)	17.6 mm	-	33.92%	95.13%	99.35%
Tail	[18]	68.6 mm	6.76°	11.78%	60.87%	84.84%
	Ours ($\ell^{m,*}$)	30.1 mm	2.89°	12.11%	73.13%	96.87%
	Ours (ℓ)	28.3 mm	-	27.31%	83.12%	95.35%

Table 4: Explicit (SOMA [18]) vs implicit (Ours) labeled marker fits and direct landmarks' ℓ solving comparison.

	RMSE↓	MAE↓	PCK1↑	PCK3↑	PCK7↑
[37, 41]	30.1 mm	3.49°	11.79%	66.85%	98.34%
[7]	30.8 mm	3.10°	12.71%	67.06%	97.71%
Ours (ℓ^m)	28.9 mm	2.98°	14.71%	69.86%	98.18%

Table 5: Noisy landmark fits comparison on TH2.

that the fits on our model's markers ℓ^m deliver better performance, a fact that is mainly attributed to the robustness of regression compared to the larger error margin of fitting to incorrectly labeled markers. This is evident in all test sets but more pronounced in the tail (rare) poses. Indicative qualitative examples are depicted in Fig. 5. For completion (not direct comparison with SOMA), we include the results for solving both markers and joints (ℓ) estimated by our model, which clearly achieves the best overall performance.

Addressing input noise

Finally, we design an experiment for showcasing our model's fitting robustness to noisy marker input as discussed in Sec. 3.3. Tab. 5 presents results when fitting to noisy landmarks between the uncertainty optimization method and MoSh(++) like fitting (ignoring the latent marker optimization as the markers are extracted from the body's surface and placed using the nominal layout). The TH2 dataset is used for evaluation, with the body extracted input markers corrupted with high levels of noise (see Sec. 3.2 of the supp. for the applied types of noise) prior to fitting the body model to them. Naturally, optimizing the uncertainty region improves fitting performance to noisy observations. Compared to a more complex optimization objective that also considers the shape of the data distribution [7] we find that the proposed Gaussian uncertainty region optimization delivers improved fits. This can be attributed to the complexity of tuning it, as well as the increased parameter count. Fig. 6 depicts qualitative examples with body fits in the noisy inputs acquired with just 3 viewpoints (same capture session as Fig. 1) and shows that jointly optimizing the uncertainty region allows for robustness to input-

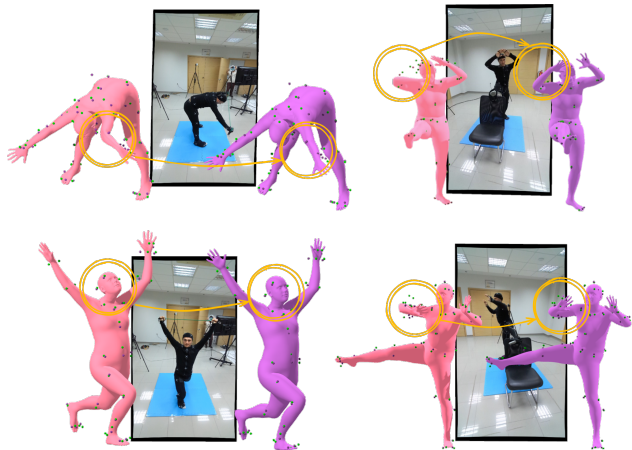


Figure 6: Plain vs uncertainty-based fit. Input markers from the consumer-grade system and the model inferred ones are colored with green, and violet respectively.

related measurement noise, as well as model-related information noise. Some interesting noise-aware fitting ablations along with visualizations can be found in Sec. 9 of the supplementary material.

Real-time performance. We validate our end-to-end method by implementing a real-time system using sparse consumer-grade sensors (see details in Sec. 11 of the supp.). Leveraging the orthogonal view two-pass approach we deploy an optimized ONNX [1] model where we flatten the two passes across the batch dimension, performing only the light-weight marginal heatmap fusion in a synchronized manner. Our system achieves under 16ms inference even on a laptop equipped with a mobile-grade RTX 2080. Nonetheless, we understand that high-quality MoCap requires greater efficiency to achieve processing rates of at least 120Hz and we set this rate as the next goal.

5. Conclusion

MoCap data are highly imbalanced and in this work we have presented a novel technique for imbalanced regression. Still we believe we have but scratched the surface of exploiting representation learning for addressing the long-tail and bias, as different architectures, samplers and relevance functions can be explored. At the same time, this work contributes to integrating machine learning in real-time optical MoCap, while also making it more accessible. However, there is room for improvements, as temporal information is not integrated in our approach, and a single, fixed marker layout is only supported.

References

- [1] Open Neural Network Exchange (ONNX). <https://github.com/onnx/onnx>. 8
- [2] Advanced Computing Center for the Arts and Design. AC-CAD MoCap Dataset. 6
- [3] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455, 2015. 7
- [4] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imGHUM: Implicit generative models of 3d human shape and articulated pose. In *Proc. IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 5461–5470, 2021. 3
- [5] Andreas Aristidou, Daniel Cohen-Or, Jessica K Hodgins, and Ariel Shamir. Self-similarity analysis for motion capture cleaning. In *Computer Graphics forum*, volume 37, pages 297–309. Wiley Online Library, 2018. 2
- [6] Andreas Aristidou and Joan Lasenby. Real-time marker prediction and cor estimation in optical motion capture. *The Visual Computer*, 29:7–26, 2013. 2
- [7] Jonathan T Barron. A general and adaptive robust loss function. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4331–4339, 2019. 6, 8
- [8] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 7
- [9] Paula Branco, Rita P Ribeiro, and Luis Torgo. Ubl: an r package for utility-based learning. *arXiv preprint arXiv:1604.08079*, 2016. 3
- [10] Paula Branco, Luis Torgo, and Rita P Ribeiro. SMOGN: A pre-processing approach for imbalanced regression. In *First international workshop on learning with imbalanced domains: Theory and applications*, pages 36–50. PMLR, 2017. 3, 4
- [11] Carnegie Mellon University. CMU MoCap Dataset. 7
- [12] Anargyros Chatzitofis, Dimitrios Zarpalas, Petros Daras, and Stefanos Kollias. Democap: low-cost marker-based motion capture. *International Journal of Computer Vision (IJCV)*, 129(12):3338–3366, 2021. 1, 2, 3, 4, 5, 7
- [13] Kang Chen, Yupan Wang, Song-Hai Zhang, Sen-Zhe Xu, Weidong Zhang, and Shi-Min Hu. Mocap-solver: A neural solver for optical motion capture data. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021. 1, 2, 3, 4, 5, 7
- [14] Wei Cheng, Su Xu, Jintan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. *arXiv preprint arXiv:2204.11798*, 2022. 3, 6
- [15] John E Dennis Jr and Roy E Welsch. Techniques for nonlinear least squares and robust regression. *Communications in Statistics-simulation and Computation*, 7(4):345–359, 1978. 5
- [16] Yinfu Feng, Mingming Ji, Jun Xiao, Xiaosong Yang, Jian J Zhang, Yueting Zhuang, and Xuelong Li. Mining spatial-temporal patterns and structural sparsity for human motion data denoising. *IEEE Transactions on Cybernetics*, 45(12):2693–2706, 2014. 2
- [17] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9919–9928, 2021. 3
- [18] Nima Ghorbani and Michael J Black. Soma: Solving optical marker-based mocap automatically. In *Proc. IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 11117–11126, 2021. 1, 2, 3, 4, 5, 6, 8
- [19] Saeed Ghorbani, Ali Etemad, and Nikolaus F Troje. Auto-labelling of markers in optical motion capture by permutation learning. In *Advances in Computer Graphics: 36th Computer Graphics International Conference, CGI 2019, Calgary, AB, Canada, June 17–20, 2019, Proceedings 36*, pages 167–178. Springer, 2019. 1, 2, 5
- [20] Yu Gong, Greg Mori, and Frederick Tung. RankSim: Ranking similarity regularization for deep imbalanced regression. *arXiv preprint arXiv:2205.15236*, 2022. 3
- [21] Mark Hamilton, Evan Shelhamer, and William T Freeman. It is likely that your loss should be a likelihood. *arXiv preprint arXiv:2007.06059*, 2020. 6
- [22] Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher D Twigg, and Kenrick Kin. Online optical marker-based hand tracking with deep labels. *ACM Transactions on Graphics (TOG)*, 37(4):1–10, 2018. 1, 2, 3, 5
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNET classification. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 6
- [24] Lorna Herda, Pascal Fua, Ralf Plänkers, Ronan Boulic, and Daniel Thalmann. Using skeleton-based tracking to increase the reliability of optical motion capture. *Human movement science*, 20(3):313–341, 2001. 2
- [25] David T Hoffmann, Dimitrios Tzionas, Michael J Black, and Siyu Tang. Learning to train with synthetic humans. In *Pattern Recognition: 41st DAGM German Conference, DAGM GCPR 2019, Dortmund, Germany, September 10–13, 2019, Proceedings 41*, pages 609–623. Springer, 2019. 3
- [26] Daniel Holden. Robust solving of optical motion capture data by denoising. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 1, 2, 3, 4, 5
- [27] Paul W Holland and Roy E Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, 6(9):813–827, 1977. 5
- [28] Alexander Hornung, Sandip Sar-Dessai, and Leif Kobbelt. Self-calibrating optical motion tracking for articulated bodies. In *Proc. IEEE Virtual Reality (IEEEVR)*, pages 75–82. IEEE, 2005. 2
- [29] Karim Isakov, Egor Burkov, Victor Lempitsky, and Yuriy Malkov. Learnable triangulation of human pose. In *Proc. ICCV*, pages 7717–7726, 2019. 3
- [30] Alec Jacobson, Zhigang Deng, Ladislav Kavan, and John P Lewis. Skinning: Real-time shape deformation (full text not available). In *ACM SIGGRAPH 2014 Courses*, pages 1–1. 2014. 3

- [31] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2020. 5
- [32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 5
- [33] Manuel Kaufmann, Yi Zhao, Chengcheng Tang, Lingling Tao, Christopher Twigg, Jie Song, Robert Wang, and Otmar Hilliges. Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11510–11520, 2021. 3
- [34] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, 2018. 6
- [35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [36] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proc. International Conference on Learning Representations, ICLR*, 2014. 4
- [37] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: motion and shape capture from sparse markers. *ACM Trans. Graph.*, 33(6):220–1, 2014. 2, 3, 6, 7, 8
- [38] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3, 6
- [39] EYES JAPAN Co. Ltd. Eyes. <http://mocapdata.com>, 2018. 7
- [40] Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, 2019. 5
- [41] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proc. IEEE/CVF international conference on computer vision (CVPR)*, pages 5442–5451, 2019. 2, 3, 6, 8
- [42] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *Proc. International Conference on Advanced Robotics (ICAR)*, pages 329–336. IEEE, 2015. 6
- [43] ML Menéndez, JA Pardo, L Pardo, and MC Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997. 5
- [44] Johannes Meyer, Markus Kuderer, Jörg Müller, and Wolfram Burgard. Online marker labeling for fully automatic skeleton tracking in optical motion capture. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 5652–5657, 2014. 2
- [45] Anton Mikhailov. Turbo, An Improved Rainbow Colormap for Visualization. <https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html>, 2019. 5
- [46] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018. 5
- [47] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. 3d human pose estimation with 2d marginal heatmaps. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1477–1485. IEEE, 2019. 5
- [48] Jorge Nocedal and Stephen J Wright. Nonlinear equations. *Numerical Optimization*, pages 270–302, 2006. 6
- [49] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *Proc. European Conference on Computer Vision (ECCV)*, pages 598–613. Springer, 2020. 3
- [50] Ahmed AA Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Supr: A sparse unified part-based human representation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 568–585, 2022. 3
- [51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6
- [52] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 3, 6
- [53] Dario Pavllo, Mathias Delahaye, Thibault Porssut, Bruno Herbelin, and Ronan Boulic. Real-time neural network prediction for handling two-hands mutual occlusions. *Computers & Graphics: X*, 2:100011, 2019. 2
- [54] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9054–9063, 2021. 6
- [55] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7926–7935, 2022. 3, 7
- [56] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proc. International Conference on Machine Learning (ICML)*, pages 1530–1538. PMLR, 2015. 4
- [57] Maurice Ringer and Joan Lasenby. A procedure for automatically estimating model parameters in optical motion capture. *Image and Vision Computing*, 22(10):843–850, 2004. 2
- [58] Yu Rong, Ziwei Liu, and Chen Change Loy. Chasing the tail in monocular 3d human reconstruction with prototype memory. *IEEE Transactions on Image Processing (TIP)*, 31:2907–2919, 2022. 2, 3

- [59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. **5**
- [60] Tobias Schubert, Alexis Gkoggidis, Tonio Ball, and Wolfram Burgard. Automatic initialization for skeleton tracking in optical motion capture. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 734–739. IEEE, 2015. **2**
- [61] Ken Shoemake. Animating rotation with quaternion curves. In *Proc. Conference on Computer Graphics and Interactive Techniques*, page 245–254, 1985. **5**
- [62] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1-2):4, 2010. **6**
- [63] Aníbal Silva, Rita P Ribeiro, and Nuno Moniz. Model optimization in imbalanced regression. In *Discovery Science: 25th International Conference, DS 2022, Montpellier, France, October 10–12, 2022, Proceedings*, pages 3–21. Springer, 2022. **3**
- [64] Jannik Steinbring, Christian Mandery, Florian Pfaff, Florian Faion, Tamim Asfour, and Uwe D Hanebeck. Real-time whole-body human motion tracking based on unlabeled markers. In *2016 IEEE International Conference on Multi-sensor Fusion and Integration for Intelligent Systems (MFI)*, pages 583–590. IEEE, 2016. **2**
- [65] Michael Steininger, Konstantin Kobs, Pádraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110:2187–2211, 2021. **3**
- [66] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proc. European conference on computer vision (ECCV)*, pages 529–545, 2018. **5**
- [67] Christopher Tensmeyer and Tony Martinez. Robust keypoint detection. In *Proc. International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 5, pages 1–7, 2019. **5**
- [68] Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *Proc. European Conference on Computer Vision (ECCV)*, pages 572–589. Springer, 2022. **5**
- [69] Luís Torgo, Paula Branco, Rita P Ribeiro, and Bernhard Pfahringer. Resampling strategies for regression. *Expert Systems*, 32(3):465–476, 2015. **3, 4**
- [70] Luis Torgo and Rita Ribeiro. Utility-based regression. In *PKDD*, volume 7, pages 597–604. Springer, 2007. **3, 4**
- [71] Luís Torgo, Rita P Ribeiro, Bernhard Pfahringer, and Paula Branco. Smote for regression. In *Progress in Artificial Intelligence: 16th Portuguese Conference on Artificial Intelligence, EPIA 2013, Angra do Heroísmo, Azores, Portugal, September 9-12, 2013. Proceedings 16*, pages 378–389. Springer, 2013. **3, 4**
- [72] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 109–117, 2017. **3**
- [73] Marin Šarić. Libhand: A library for hand articulation, 2011. Version 0.9. **3**
- [74] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(10):3349–3364, 2020. **7**
- [75] Yupan Wang, Guiqing Li, Huiqian Zhang, Xinyi Zou, Yuxin Liu, and Yongwei Nie. Panoman: Sparse localized components-based model for full human motions. *ACM Transactions on Graphics (TOG)*, 40(2):1–17, 2021. **3**
- [76] Tom White. Sampling generative networks: Notes on a few effective techniques. *arXiv:1609.04468*, 2016. **5**
- [77] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & Ghuml: Generative 3d human shape and articulated pose models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6184–6193, 2020. **3**
- [78] Haonan Yan, Jiaqi Chen, Xujie Zhang, Shengkai Zhang, Nianhong Jiao, Xiaodan Liang, and Tianxiang Zheng. Ultrapose: Synthesizing dense pose with 1 billion points by human-body decoupling 3d model. In *Proc. IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 10891–10900, 2021. **3**
- [79] Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *Proc. International Conference on Machine Learning (ICML)*, pages 11842–11851. PMLR, 2021. **3**
- [80] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In *Proc. European Conference on Computer Vision (ECCV)*, pages 142–159. Springer, 2022. **5**
- [81] KangKang Yin and Goh Jing Ying. SFU motion capture database. <http://mocap.cs.sfu.ca>. **7**
- [82] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5746–5756, 2021. **3, 6**
- [83] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7093–7102, 2020. **5**