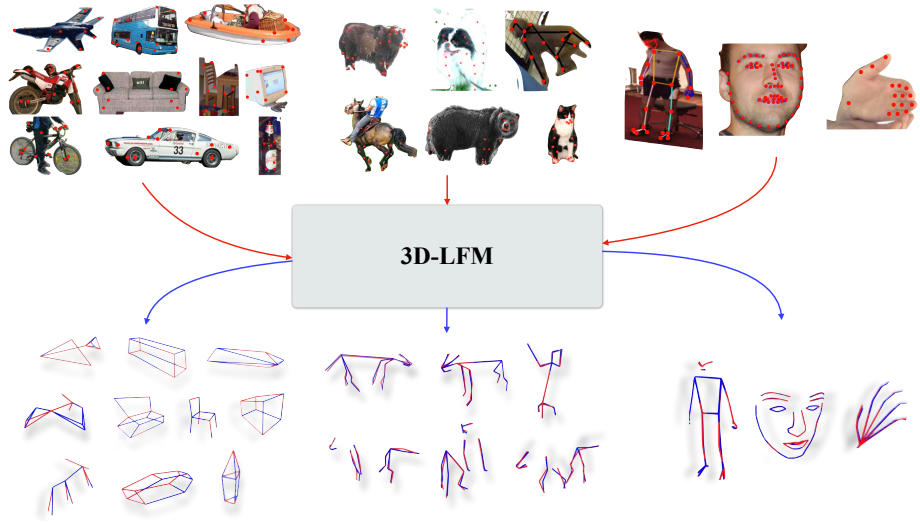
arXiv:2312.11894v2 [cs.CV] 26 Apr 2024

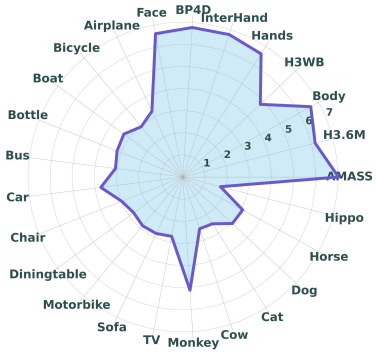
3D-LFM: Lifting Foundation Model

Mosam Dabhi[1] László A. Jeni[1*] Simon Lucey[2*]
[1]Carnegie Mellon University [2]The University of Adelaide

3dlfm.github.io



(a) Unified 2D-3D lifting for 30+ categories.



(b) Dataset diversity visualization.

Figure 1. **Overview: (a)** This figure shows the 3D-LFM's ability in lifting 2D landmarks into 3D structures across an array of over 30 diverse categories, from human body parts, to a plethora of animals and everyday common objects. The lower portion shows the actual 3D reconstructions by our model, with red lines representing the ground truth and blue lines showing the 3D-LFM's predictions. **(b)** This figure displays the model's training data distribution on a logarithmic scale, highlighting that inspite of 3D-LFM being trained on imbalanced datasets, it preserves the performance across individual categories.

Abstract

The lifting of a 3D structure and camera from 2D landmarks is at the cornerstone of the discipline of computer vision. Traditional methods have been confined to specific rigid objects, such as those in Perspective-n-Point (PnP) problems, but deep learning has expanded our capability to reconstruct a wide range of object classes (e.g. C3DPO [18] and PAUL [24]) with resilience to noise, occlusions, and perspective distortions. However, all these techniques have been limited by the fundamental need to establish correspondences across the 3D training data, significantly limiting their utility to applications where one has an abundance of "in-correspondence" 3D data. Our approach harnesses the inherent permutation equivariance of transformers to manage varying numbers of points per 3D data instance, withstands occlusions, and generalizes to unseen categories. We demonstrate state-of-the-art performance across 2D-3D lifting task benchmarks. Since our approach can be trained across such a broad class of structures, we refer to it simply as a 3D Lifting Foundation Model (3D-LFM) – the first of its kind.

1. Introduction

Lifting 2D landmarks from a single-view RGB image into 3D has long posed a complex challenge in the field of computer vision because of the ill-posed nature of the problem. This task is important for a range of applications from augmented reality to robotics, and requires an understanding of non-rigid spatial geometry and accurate object descriptions [2, 11, 25]. Historically, efforts in single-frame 2D-3D lifting have encountered significant hurdles: reliance on object-specific models, poor scalability, and limited adaptability to diverse and complex object categories. Traditional methods, while advancing in specific domains like human

*Both authors advised equally.

body [14, 16, 31] or hand modeling [3, 6], often fail when faced with the complexities of varying object types or object rigs (skeleton placements).

To facilitate such single-frame 2D-3D lifting, deep learning methods like C3DPO [18] and others [8, 11, 24, 25, 28] have recently been developed. However, these methods are fundamentally limited in that they must have knowledge of the object category and how the 2D landmarks correspond semantically to the 2D/3D data it was trained upon. Further, this represents a drawback, especially when considering their scaling up to dozens or even hundreds of object categories, with varying numbers of landmarks and configurations. This paper marks a departure from such correspondence constraints, introducing the 3D Lifting Foundation Model (3D-LFM), an object-agnostic single frame 2D-3D lifting approach. At its core, 3D-LFM addresses the limitation of previous models, which is the inability to efficiently handle a wide array of object categories while maintaining high fidelity in 3D keypoint lifting from 2D data. We propose a solution rooted in the concept of permutation equivariance, a property that allows our model to autonomously establish correspondences among diverse sets of input 2D keypoints.

3D-LFM is capable of performing single frame 2D-3D lifting for 30+ categories using a single model simultaneously, covering everything from human forms [9, 15, 32], face [29], hands [17], and animal species [1, 10, 27], to a plethora of inanimate objects found in everyday scenarios such as cars, furniture, etc. [26]. Importantly, 3D-LFM is inherently scalable, poised to expand to hundreds of categories and improve performance, especially in out-of-distribution or less-represented areas, showcasing its broad utility in 3D lifting tasks. 3D-LFM is able to achieve 2D-3D lifting performance that matches those of leading methods specifically optimized for individual categories. The generalizability of 3D LFM is further evident in its ability to handle out-of-distribution (OOD) object categories and rigs, which we refer to as OOD 2D-3D lifting, where the task is to lift the 2D landmarks to 3D for a category never seen during training. We show such OOD results: **(1)** for inanimate objects - by holding out an object category within the PASCAL dataset, **(2)** for animals - by training on common object categories such as dogs and cats found in [27] and reconstructing 3D for unseen and rare species of Cheetahs found in [10] and in-the-wild zoo captures from [5], and **(3)** by showing rig transfer, *i.e.* training 2D to 3D lifting on a Human3.6M dataset rig [7] and showing similar 2D to 3D lifting performance on previously unseen rigs such as those found in Panoptic studio dataaset rig [9] or a COCO dataset rig [13]. 3D-LFM transfers learnings from seen data during training to unseen OOD data during inference. It does so by learning general structural features during the training phase via the proposed permutation equivariance properties and specific design choices that we discuss in the following sections.

Recognizing the important role geometry plays in 3D reconstruction [4, 5, 11, 18, 24, 25], we integrate Procrustean methods such as Orthographic-N-Point (OnP) or Perspective-N-Point (PnP) to direct the model's focus on deformable aspects within a canonical frame. This incorporation significantly reduces the computational burden on the model, freeing it from learning redundant rigid rotations and focusing its capabilities on capturing the true geometric essence of objects. Scalability, a critical aspect of our model, is addressed through the use of tokenized positional encoding (TPE), which, when combined with graph-based transformer architecture, not only enhances the model's adaptability across diverse categories but also strengthens its ability to handle multiple categories with different number of keypoints and configurations. Finally, the use of skeleton information (joint connectivity) within the graph-based transformers via adjacency matrices provides strong clues about joint proximity and inherent connectivity, aiding in the handling of correspondences across varied object categories.

To the best of our knowledge, 3D-LFM is one of the only known work which is a unified model capable of doing 2D-3D lifting for 30+ (and potentially even more) categories simultaneously. Its ability to perform unified learning across a vast spectrum of object categories without specific object information and its handling of OOD scenarios highlight its potential as one of the first models capable of serving as a 2D-3D lifting foundation model.

The contributions of this paper are threefold:

1. We propose a Procrustean transformer that is able to focus solely on learning the deformable aspects of objects within a single canonical frame whilst preserving permutation equivariance across 2D landmarks.
2. The integration of tokenized positional encoding within the graph-based transformer, to enhance our approach's scalability and its capacity to handle diverse and imbalanced datasets.
3. We demonstrate that 3D-LFM surpasses state-of-the-art methods in categories such as humans, hands, and faces (benchmark in [32]). Additionally, it shows robust generalization by handling previously unseen objects and configurations, including animals ([5, 10]), inanimate objects ([26]), and novel object arrangements (rig transfer in [9])

In subsequent sections, we explore the design and methodology of our proposed 3D-LFM architecture, including detailed ablation experiments and comparative analyses. Throughout this paper, 'keypoints', 'landmarks', and 'joints' are used interchangeably, referring to specific, identifiable points or locations on an object or figure that are crucial for understanding its structure and geometry.

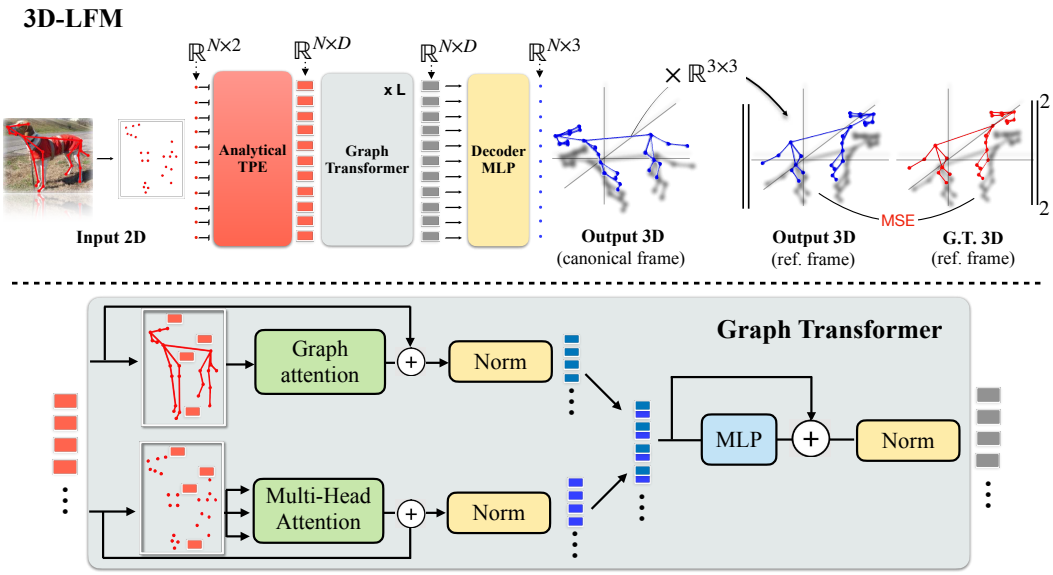


Figure 2. **Overview of the 3D Lifting Foundation Model (3D-LFM) architecture**: The process begins with the input 2D keypoints undergoing Token Positional Encoding (TPE) before being processed by a series of graph-based transformer layers. The resulting features are then decoded through an MLP into a canonical 3D shape. This shape is aligned to the ground truth (G.T. 3D) in the reference frame using a Procrustean method, with the Mean Squared Error (MSE) loss computed to guide the learning. The architecture captures both local and global contextual information, focusing on deformable structures while minimizing computational complexity.

2. Related works

The field of 2D-3D lifting has evolved substantially from classic works such as those based on Perspective-n-Point (PnP) algorithms [12]. In these early works, the algorithm was given a set of 2D landmarks and some 3D supervision, namely the known 3D rigid object. The field has since witnessed a paradigm shift with the introduction of deep learning methodologies, led by methods such as C3DPO [18], PAUL [24], and Deep NRSfM [11], along with recent transformer-based innovations such as NRSfM-Former [8]. In these approaches one does not need knowledge of the specific 3D object, instead it can get away with just the 2D landmarks and correspondences to an ensemble of 2D/3D data from the object category to be lifted. However, despite their recent success, all these methods still require that the 2D/3D data be in semantic correspondence. That is, the index to a specific landmark has the same semantic meaning across all instances (e.g. chair leg). In practice, this is quite limiting at run-time, as one needs intimate knowledge of the object category, and rig in order to apply any of these current methods. Further, this dramatically limits the ability of these methods to leverage cross-object and cross-rig datasets, prohibiting the construction of a truly generalizable 2D to 3D lifting foundation model – a topic of central focus in this paper.

Recent literature in pose estimation, loosely connected to NRSfM but often more specialized towards human and animal body parts, has also seen remarkable progress. Models such as Jointformer [14] and SimpleBaseline [16] have refined the single-frame 2D-3D lifting process, while generative approaches like MotionCLIP [19] and Human Motion Diffusion Model [20] have laid the groundwork for 3D generative motion-based foundation models. These approaches, however, are even more limiting than C3PDO, PAUL, etc. in that they are intimately wedded to the object class and are not easily extendable to an arbitrary object class.

3. Approach

Given a set of 2D keypoints representing the projection of an object's joints in an image, we denote the keypoints matrix as $\mathbf{W} \in \mathbb{R}^{N\times 2}$, where N is the predetermined maximum number of joints considered across all object categories. For objects with joints count less than N, we introduce a masking mechanism that utilizes a binary mask matrix $\mathbf{M} \in \{0,1\}^N$, where each element m_i of $\mathbf{M}$ is defined as:

$$m_i = \begin{cases} 1 & \text{if joint } i \text{ is present} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The 3D lifting function $f : \mathbb{R}^{N\times 2} \rightarrow \mathbb{R}^{N\times 3}$ maps the 2D keypoints to their corresponding 3D structure while compensating for the projection:

$$\mathbf{S} = f(\mathbf{W}) = \mathbf{W}\mathbf{R}^\top + \mathbf{b} \tag{2}$$

where $\mathbf{R} \in \mathbb{R}^{3\times3}$ is the projection matrix (assumed either weak-perspective or orthographic) and $\mathbf{b} \in \mathbb{R}^{N\times3}$ is a bias term that aligns the centroids of 2D and 3D keypoints.

Permutation Equivariance: To ensure scalability and adaptability across a diverse set of objects, we leverage the property of permutation equivariance inherent in transformer architectures. Permutation equivariance allows the model to process input keypoints $\mathbf{W}$ regardless of their order, a critical feature for handling objects with varying joint configurations:

$$f(\mathcal{P}\mathbf{W}) = \mathcal{P}f(\mathbf{W})$$

where $\mathcal{P}$ is a permutation matrix that reorders the keypoints.

Handling Missing Data: To address the challenge of missing data, we refer the Deep NRSfM++ [25] work and use a masking mechanism to accommodate for occlusions or absences of keypoints. Our binary mask matrix $\mathbf{M} \in \{0,1\}^N$ is applied in such a way that it not only pads the input data to a consistent size but also masks out missing or occluded points: $\mathbf{W}_m = \mathbf{W} \odot \mathbf{M}$, where $\odot$ denotes element-wise multiplication. To remove the effects of translation and ensure that our TPE features are generalizable, we zero-center the data by subtracting the mean of the visible keypoints:

$$\mathbf{W}_c = \mathbf{W}_m - \text{mean}(\mathbf{W}_m) \tag{3}$$

We scale the zero-centered data to the range $[-1, 1]$ while preserving the aspect ratio to maintain the geometric integrity of the keypoints. For more details on handling missing data in the presence of perspective effects, we refer the reader to Deep NRSFM++[25].

Token Positional Encoding: replaces the traditional Correspondence Positional Encoding (CPE) or Joint Embedding which encodes the semantic correspondence information (as used in works such as like [14, 31]) with a mechanism that does not require explicit correspondence or semantic information. Owing to the success of per-point positional embedding, particularly random Fourier features [30] in handling OOD data, we compute Token Positional Encoding (TPE) using analytical Random Fourier features (RFF) as follows:

$$\mathbf{TPE}(\mathbf{W}_c) = \sqrt{\frac{2}{D}}\Big[\sin(\mathbf{W}_c\boldsymbol{\omega} + b); \cos(\mathbf{W}_c\boldsymbol{\omega} + b)\Big] \tag{4}$$

where D is the dimensionality of the Fourier feature space, $\boldsymbol{\omega} \in \mathbb{R}^{2\times\frac{D}{2}}$ and $\mathbf{b} \in \mathbb{R}^{\frac{D}{2}}$ are parameters sampled from a normal distribution, scaled appropriately. These parameters are sampled once and kept fixed, as per the RFF methodology. The output of this transformation $\mathbf{TPE}(\mathbf{W}_c)$ is then fed into the graph-based transformer network as $\mathbf{X}^\ell$ where ℓ indicates the layer number (0 in the above case). This set of features is now ready for processing inside the graph-based transformer layers without the need for correspondence among the input keypoints. The TPE retains the property of permutation equivariance while implicitly encoding the relative positions of the keypoints.

3.1. Graph-based Transformer Architecture

Our graph-based transformer architecture utilizes a hybrid approach to feature aggregation by combining graph-based local attention [22]($\mathbf{L}$) with global self-attention mechanisms [21]($\mathbf{G}$) within a single layer (shown as grey block in Fig. 2. This layer is replicated L times, providing a sequential refinement of the feature representation across the network's depth.

Hybrid Feature Aggregation: For each layer ℓ, ranging from 0 to L, the feature matrix $\mathbf{X}^{(\ell)} \in \mathbb{R}^{N\times D}$ is augmented through simultaneous local and global processing. The local processing component, $\texttt{GA}(\mathbf{X}^{(\ell)}, \mathbf{A})$, leverages an adjacency matrix $\mathbf{A}$, which encodes the connectivity based on the object category, to perform graph-based attention on batches of nodes representing the input 2D data:

$$\begin{aligned}\mathbf{L}^{(\ell)} &= \texttt{GA}(\mathbf{X}^{(\ell)}, \mathbf{A}),\\ \mathbf{G}^{(\ell)} &= \texttt{MHSA}(\mathbf{X}^{(\ell)})\end{aligned} \tag{5}$$

Local and global features are concatenated to form a unified representation $\mathbf{U}^{(\ell)}$:

$$\mathbf{U}^{(\ell)} = \texttt{concat}(\mathbf{L}^{(\ell)}, \mathbf{G}^{(\ell)}) \tag{6}$$

Following the concatenation, each layer applies a normalization(LN) and a multilayer perceptron (MLP). The MLP employs a Gaussian Error Linear Unit (GeLU) as the nonlinearity function to enhance the model's expressive power

$$\begin{aligned}\mathbf{X}'^{(\ell)} &= \texttt{LN}(\mathbf{U}^{(\ell)}) + \mathbf{U}^{(\ell)},\\ \mathbf{X}^{(\ell+1)} &= \texttt{LN}(\texttt{MLP\_GeLU}(\mathbf{X}'^{(\ell)})) + \mathbf{X}'^{(\ell)}\end{aligned} \tag{7}$$

Here, GA represents Graph Attention, MHSA denotes Multi-Head Self-Attention, and MLP_GeLU indicates our MLP with GeLU nonlinearity. This architecture is designed to learn patterns in 2D data by considering both the local neighborhood connectivity of input 2D and the global data context of input 2D, which is important for robust 2D to 3D structure lifting.

3.2. Procrustean Alignment

The final operation in our pipeline decodes the latent feature representation $\mathbf{X}^{(L)}$ into the predicted canonical structure $\mathbf{S}_c$ via a GeLU-activated MLP:

$$\mathbf{S}_c = \texttt{MLP}_{\text{shape\_decoder}}(\mathbf{X}^{(L)})$$

Subsequently, we align $\mathbf{S}_c$ with the ground truth $\mathbf{S}_r$, via a Procrustean alignment method that optimizes for the rotation matrix $\mathbf{R}$. The alignment is formalized as a minimization problem:

$$\underset{\mathbf{R}}{\text{minimize}} \quad ||\mathbf{M} \odot (\mathbf{S}_r - \mathbf{S}_c\mathbf{R})||_F^2$$

where $\mathbf{M}$ is a binary mask applied element-wise, and $||\cdot||_F$ denotes the Frobenius norm. The optimal $\mathbf{R}$ is obtained via SVD, which ensures the orthonormality constraint of the rotation matrix:

$$\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^\top = \text{SVD}((\mathbf{M} \odot \mathbf{S}_c)^\top \mathbf{S}_r), \quad \mathbf{R} = \mathbf{U}\mathbf{V}^\top$$

The predicted shape is then scaled relative to the reference shape $\mathbf{S}_r$, resulting in a scale factor γ, which yields the final predicted shape $\mathbf{S}_p$:

$$\mathbf{S}_p = \gamma \cdot (\mathbf{S}_c\mathbf{R})$$

This Procrustean alignment step is crucial for directing the model's focus on learning non-rigid shape deformations over rigid body dynamics, thus significantly enhancing the model's ability to capture the true geometric essence of objects by just focusing on core deformable (non-rigid) aspects. The effectiveness of this approach is confirmed by faster convergence and reduced error rates in our experiments, as detailed in Fig. 6. These findings align with the findings presented in PAUL [24].

3.3. Loss Function

The optimization of our model relies on the Mean Squared Error (`MSE`) loss, which calculates the difference between predicted 3D points $\mathbf{S}_p$ and the ground truth $\mathbf{S}_r$:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{S}_p^{(i)} - \mathbf{S}_r^{(i)}\|^2 \tag{8}$$

Minimizing this loss across N points ensures the model's ability in reconstructing accurate 3D shapes from input 2D landmarks. This minimization effectively calibrates the shape decoder and the Procrustean alignment to focus on the essential non-rigid characteristics of the objects, helping the accuracy of the 2D to 3D lifting process.

4. Results and Comparative Analysis

Our evaluation shows the 3D Lifting Foundation Model (3D-LFM)'s capability in single-frame 2D-3D lifting across diverse object categories without object-specific data in Sec. 4.1. Following that, Sec. 4.2 highlights 3D-LFM's performance over specialized methods, especially achieving state-of-the-art performance in whole-body benchmarks[32] (Fig. 4). Additionally, Sec. 4.3 shows 3D-LFM's capability in 2D-3D lifting across 30 categories using a single unified model, enhancing category-specific performance and achieving out-of-distribution (OOD) generalization for unseen object configurations during training. In conclusion, the ablation studies in Section 4.4 validate our proposed procrustean approach, token positional encoding, and the local-global hybrid attention mechanism in the transformer model, confirming their role in 3D-LFM's effectiveness in both single- and multiple-object scenarios.

4.1. Multi-Object 3D Reconstruction

Clarifying naming convention: In 'object-specific' versus 'object-agnostic', our primary focus in this naming is on the distinction in *training* methods. Here, object-specific training involves supplying semantic details for each object, leading to isolated training. Conversely, object-agnostic training combines various categories without explicit landmark semantics, leading to combined training.

Experiment Rationale: 3D-LFM leverages permutation equivariance to accurately lift 2D keypoints into 3D structures across diverse categories, outperforming fixed-array methods by adapting flexibly to variable keypoint configurations. It has been evaluated against non-rigid structure-from-motion approaches [11, 18, 24, 25] that require object-specific inputs, showing its ability to handle diverse categories. For a comprehensive benchmark, we utilize the PASCAL3D+ dataset [26], following C3DPO's [18] methodology, to include a variety of object categories.

Performance: We benchmark 3D-LFM against the notable NRSfM method, C3DPO [18], for multi-object 2D to 3D lifting with 3D supervision. C3DPO, similar to other contemporary methods [11, 24, 25, 28] requiring object-specific details, serves as an apt comparison due to its multi-category approach. Initially replicating conditions with object-specific information, 3D-LFM matches C3DPO's performance, as demonstrated in Fig. 3. This stage uses MPJPE to measure 3D lifting accuracy, with C3DPO's training setup including an MN dimensional array for object details where M represents number of objects with N being maximum number of keypoints, and our model is trained separately on each object to avoid providing object-specific information. The 3D-LFM's strength emerges when object-specific data is withheld. While C3DPO shows a decline without such data, 3D-LFM maintains a lower MPJPE across categories, even when trained collectively across categories using only an N dimensional array. These findings (Fig. 3) highlights 3D-LFM's capabilities, outperforming single-category training and demonstrating its potential as a generalized 2D to 3D lifting solution.

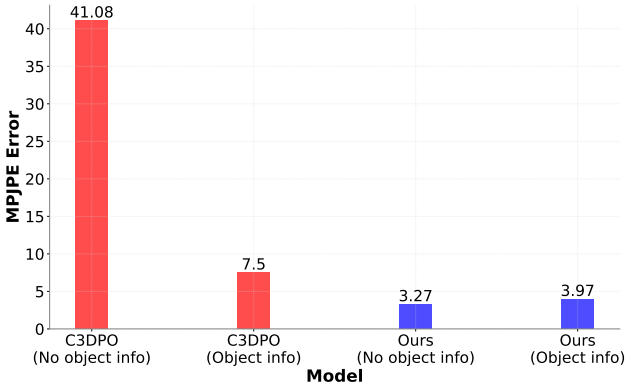


Figure 3. **3D-LFM vs. C3DPO Performance**: MPJPE comparisons using the PASCAL3D+ dataset, this figure demonstrates our model's adaptability in the absence of object-specific information, contrasting with C3DPO's increased error under the same conditions. The analysis confirms 3D-LFM's superiority across diverse object categories, reinforcing its potential for generalized 2D to 3D lifting.

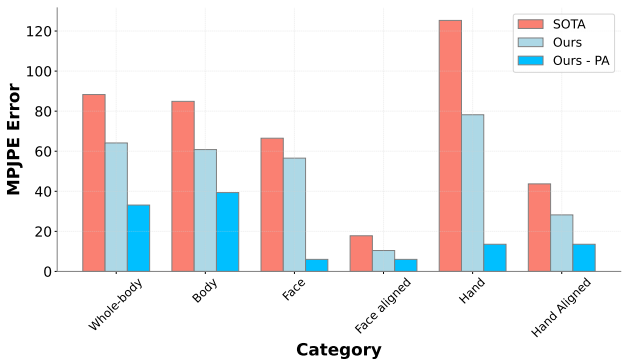


Figure 4. **Performance Comparison on H3WB Benchmark**: This chart contrasts MPJPE errors for whole-body, body, face, aligned face, hand, and aligned hand categories within the H3WB benchmark [32]. Our models, with and without Procrustes Alignment (Ours-PA), outperform current state-of-the-art (SOTA) methods, validating our approach's proficiency in 2D to 3D lifting tasks.

Table 1. **Quantitative performance on H3WB**: Our method demonstrates leading performance across multiple object categories without the need for object-specific designs.

Method	Whole-body	Body	Face/Aligned	Hand/Aligned
SimpleBaseline	125.4	125.7	115.9 / 24.6	140.7 / 42.5
CanonPose w/3D sv.	117.7	117.5	112.0 / 17.9	126.9 / 38.3
Large SimpleBaseline	112.3	112.6	110.6 / 14.6	114.8 / 31.7
Jointformer (extra data)	81.5	78	60.4 / 16.2	117.6 / 38.8
Jointformer	88.3	84.9	66.5 / 17.8	125.3 / 43.7
Ours	**64.13**	**60.83**	**56.55 / 10.44**	**78.21 / 28.22**
Ours – PA	**33.13**	**39.36**	**6.02**	**13.56**

4.2. Benchmark: Object-Specific Models

Next, we benchmark 3D-LFM against leading specialized methods for human body, face, and hands categories. Our model outperforms these specialized methods, showing multi-category learning without the need for category (landmark) semantics. For this study, we evaluate on H3WB dataset [32], a recent benchmark for diverse whole-body pose estimation tasks. This dataset is valuable for its inclusion of multiple object categories and for providing a comparative baseline against methods such as Jointformer [14], SimpleBaseline [16], and CanonPose [23]. Following H3WB's recommended 5-fold cross-validation and submitting the evaluations to benchmark's authors, we report results on the hidden test set. The results shown in Fig. 4 and Table 1 include PA-MPJPE and MPJPE, with test set performance numbers provided directly by the H3WB team, ensuring that our results are verified by an independent third-party.

4.3. Towards foundation model

In this section, we highlight 3D-LFM's role as a foundational model for varied 2D-3D lifting, capable in managing multiple object types and data imbalances. In this subsection, we explore 3D-LFM's scalability for collective dataset training (Sec.4.3.1), its generalization to new categories and rig transfer capabilities (Sec.4.3.2). These studies validate the 3D-LFM's role as a foundation model, capable at leveraging diverse data without requiring specific configurations, thus simplifying the 3D lifting process for varied joint setups.

We start this investigation by showing the capability 3D-LFM in handling 2D-3D lifting for 30+ object categories within the single model, confirming the model's capability to manage imbalanced datasets representative of real-world scenarios as shown in Fig. 1. With a comprehensive range of human, hand, face, inanimate objects, and animal datasets, the 3D-LFM is proven to be scalable, without requiring category-specific adjustments. The subsequent subsections will dissect these attributes further, discussing the 3D-LFM's foundational potential in the 3D lifting domain.

4.3.1 Combined Dataset Training

This study evaluates the 3D-LFM's performance on isolated datasets against its performance on a combined dataset. Initially, the model was trained separately on animal-based supercategory datasets: specifically **OpenMonkey**[1] and **Animals3D**[27]. Subsequently, it was trained on a merged dataset containing a broad spectrum of object categories, including **Human Body-Based** datasets such as AMASS [15]

Table 2. **Quantitative evaluation for OOD scenarios**

Category	**OOD** (mm)	**In-Dist.** (mm)
Cheetah [10]	26.59	10.16
Train [26]	6.88	5.71
Chimpanzee [5]	52.05	42.65

and Human 3.6 [7], **Hand-Based** datasets such as PanOptic Hands [9], **Face-Based** datasets like BP4D+[29], and various **Inanimate Objects** from the PASCAL3D+ dataset[26], along with previously mentioned animal datasets. Isolated training resulted in an average MPJPE of $\mathbf{21.22}$ *mm*, while the combined training method significantly reduced MPJPE to $\mathbf{12.5}$ *mm* on the same animal supercategory validation split. This improvement confirms the potential of 3D-LFM as a pre-training framework and underscores its ability to adapt and generalize from diverse and extensive data collections.

Dataset Selection Rationale: We selected animal-based supercategory datasets to demonstrate combined training's impact on underrepresented categories. We observed greater performance improvements in smaller, unbalanced datasets (as exemplified by PASCAL3D+: from $\mathbf{4.31}$ *mm* to $\mathbf{1.1}$ *mm* and OpenMonkey: from $\mathbf{19.45}$ *mm* to $\mathbf{9.59}$ *mm*) compared to larger datasets with sufficient balance among categories. Consequently, we see minimal gains in more balanced, larger datasets like AMASS (from $\mathbf{1.67}$ *mm* to $\mathbf{1.66}$ *mm*), underscoring the utility of combined training for enhancing performance in underrepresented and long-tail categories.

4.3.2 OOD generalization and rig-transfer:

We evaluate 3D-LFM's generalization to unseen object categories and rig configurations. Its accuracy is highlighted by successful 2D-3D lifting reconstructions of the "Cheetah" from Acinoset [10], which is not included in the typical Animal3D dataset [27], and the "Train" category from PASCAL3D+[26], absent during training. Qualitative reconstructions are shown in Fig. 5, along with the quantitative results in Tab.2 for above categories as well as in-the-wild category like a Chimpanzee from the MBW dataset [5] – which illustrates model's strong OOD generalization and capability to handle in-the-wild data.

Additionally, we show 3D-LFM's capability in transferring rig configurations between datasets, embodying the concept of generic geometry learning. By training on a 17-joint Human3.6M dataset [7] and testing on a 15-joint Panoptic Studio setup [9], our model gives accurate 3D reconstructions despite variations in joint arrangements. This capability is particularly interesting for its efficiency in utilizing data from multiple rigs of the same object, and underscores the model's adaptability, a cornerstone in processing diverse human datasets. It aligns with the broader community's interest in versatile geometry learning, which makes these findings especially compelling. For a more thorough validation, we direct the reader to the ablation section, where qualitative visuals (Fig. 7) and quantitative analysis (Sec. 4.4.3) further highlight 3D-LFM's OOD generalization and rig transfer efficacy.

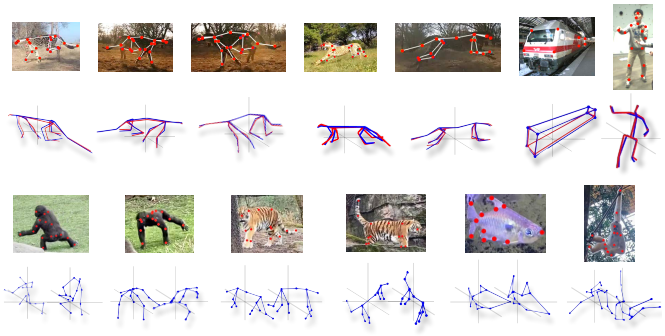


Figure 5. **Generalization to unseen data**: Figure showing 3D-LFM's proficiency in OOD 2D-3D lifting, effectively handling new, unseen categories, and rig generalization from Acinoset [10] PASCAL3D+ [26], and Panoptic studio [9] with varying joint arrangements in top row. The bottom row presents in-the-wild data from the MBW dataset [5], with red dots indicating input keypoints and blue stick figures showing the model's 3D predictions from different angles.

4.4. Ablation

In our ablation studies, we evaluate the 3D-LFM's design elements and their individual contributions to its performance. Detailed experiments on the Human3.6M benchmark [7] and a blend of other datasets including Animal3D [27] and facial datasets [9, 29] were carried out to ablate the role of Procrustean transformation, hybrid attention mechanisms, and tokenized positional encoding (TPE) in enabling the model's scalability and out-of-distribution (OOD) generalization.

4.4.1 Procrustean Transformation

3D-LFM's fusion of the procrustean approach, a first in transformer-based lifting frameworks, concentrates on deformable object components, as outlined in Sec.3.2. By focusing on shape within a standard canonical reference frame and avoiding rigid body transformations, we see faster learning and a decreased MPJPE, as evident by the gap between blue and orange lines in Fig. 6 **(a)** suggests. This fusion is crucial for learning 3D deformations, while utilizing transformers' equivariance. These findings suggest that even for transformers, avoiding rigid transformations' learning aids convergence, most notably with imbalanced datasets.

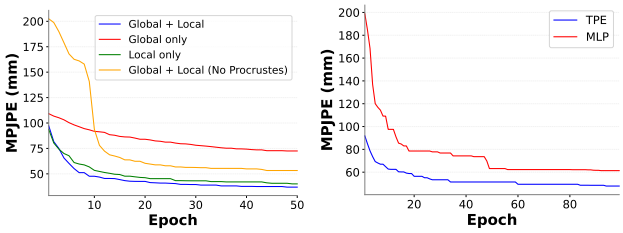


Figure 6. **(a)** Comparing attention strategies in 3D-LFM. The combined local-global approach with procrustean alignment surpasses other configurations in MPJPE reduction over 100 epochs on the Human3.6M validation split. **(b)** rapid convergence and efficiency of the TPE approach compared to the learnable MLP

Table 3. **Impact of TPE on Data Imbalance and Rig Transfer**

Study	Experiment	Model Size	Improvement (%)
Data Imbalance	Underrepr. category (Hippo) [27]	128	3.27
		512	12.28
		1024	22.02
Rig Transfer	17 [7]- to 15 [9]-joint	N/A	12
	15 [9]- to 17 [7]-joint		23.29
	52 [9]- to 83 [29]-joint		52.3

4.4.2 Local-Global vs. Hybrid Attention

In evaluating 3D-LFM's attention strategies, our analysis on the same validation split as above demonstrates the superiority of a hybrid approach combining local (GA) and global (MHSA) attention mechanisms. This integration, particularly when complemented by Procrustean alignment, significantly enhances performance and accelerates convergence, as evidenced in Fig. 6 **(a)**. The distinct advantage of this hybrid system validates our architectural choices, showcasing its efficiency in reducing MPJPE errors and refining model training dynamics.

4.4.3 Tokenized Positional Encoding:

This ablation study assesses the impact of Tokenized Positional Encoding (TPE), which uses analytical Random Fourier Features for encoding positional information. This study examines TPE's influence on model performance in scenarios of data imbalance and rig transfer generalization.
Data imbalance study: When tested on the underrepresented hippo category from the Animal3D dataset [27], TPE based model showed a 3.27% improvement in MPJPE over the baseline MLP with a 128-dimensional model performance as evident in first row of Tab. 3. This improvement grew with the model size. These results highlight TPE's scalability and its faster convergence, especially relevant in imbalanced, OOD scenarios as detailed in Fig. 6 (b). The observed performance boosts suggest that TPE's analytical nature might be more suited to adapting to novel data distributions. Increasing model size amplifies TPE's benefits,

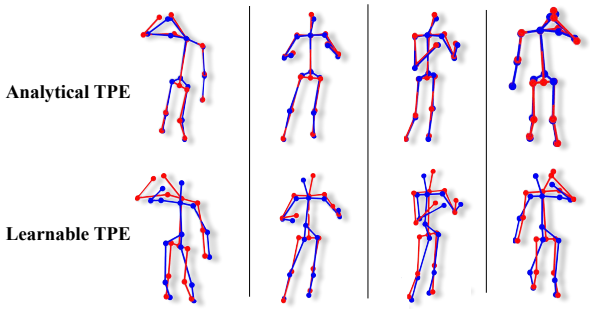


Figure 7. The qualitative improvement in rig transfer using analytical TPE versus learnable MLP projection. This visualization reinforces the necessity of TPE in handling OOD data such as different rigs, unseen during training.

hinting that its fixed analytical approach more adeptly handles OOD intricacies compared to learnable methods like MLPs, which may falter in such situations.
Rig transfer study: Our rig transfer analysis, summarized in Table 3, showcases TPE's adaptability and effectiveness over the MLP baseline across different joint configurations and rig scenarios, with improvements up to 52.3%. These findings, particularly the significant performance boost in complex rig transfers, underscore TPE's robustness in OOD contexts. Figure 7 visually highlights the qualitative differences between TPE and MLP approaches in a rig transfer scenario, where the model trained on a 17-joint [7] configuration is tested on a 15 joint [9] setup.

5. Discussion and Conclusion

The proposed 3D-LFM marks a significant leap in 2D-3D lifting, showcasing scalability and adaptability, addressing data imbalance, and generalizing to new data categories. Its cross-category knowledge transfer requires further investigation and handling of inputs with different perspectives could act as potential limitations. 3D-LFM's efficiency is demonstrated by achieving results comparable to leading methods on [32] benchmark as well as its proficiency in out-of-distribution (OOD) scenarios on limited computational resources. For training duration and computational details, please refer to the supplementary materials. This work establishes a baseline framework for future 3D pose estimation and 3D reconstruction models. In summary, the 3D-LFM creates a universally applicable model for 3D reconstruction from 2D data, paving the way for diverse applications that requires accurate 3D reconstructions from 2D inputs.
Acknowledgement: We extend our gratitude to Ian R. Fasel, Tim Clifford, Javier Movellan, and Matthias Hernandez of Apple for their insightful discussions.

Supplementary Material

I. Training Details

The 3D Lifting Foundation Model (3D-LFM), as detailed in Sec. 4.3.1, was trained across 30 diverse categories on a single `NVIDIA A100 GPU`. This dataset consisted of over 18 million samples, with data heavily imbalanced and mostly dominated by human datasets as shown in Fig. 1. This training highlights the model's practicality, with mixed datasets having imbalance within them.

Model parameters: In the architecture of 3D-LFM, the transformer block consists of four layers, with the hidden dimensions and head counts tailored to the dataset scale. Specifically, for datasets exceeding $10,000$ frames, we used a model dimension of 512 and a head count of 8. Datasets with frame counts ranging from $1,000$ to $10,000$ were assigned model dimensions of 256 with 4 heads. For smaller datasets, consisting 1 to $1,000$ frames, we employed a more compact model with dimensions set at 64 and head counts maintained at 4. This gradation in model complexity ensures a balanced approach, aligning the model capacity with the dataset size, which is particularly critical for achieving computational efficiency and avoiding overfitting on smaller datasets.

Optimizer and scheduler: `GeLU` activations were employed for non-linearity in the feedforward layers. The training process was guided by a `ReduceLROnPlateau` scheduler with a starting learning rate of `0.001` and a patience of 20 epochs. An early stopping mechanism was implemented, halting training if no improvement in MPJPE was noted for 30 epochs, ensuring efficient and optimal performance. This training approach enabled 3D-LFM to surpass leading methods in 3D lifting task proposed by H3WB benchamark [32].

II. Limitations

Perspective-Induced Misinterpretations: The 3D-LFM demonstrates a significant capability in generalizing across object categories. However, it can encounter difficulties when extreme perspective distortions cause 2D inputs to mimic the appearance of different categories. For example, unusual viewing angles can cause a tiger to be misconstrued as a primate, illustrated in Fig. 8 **(c)**, due to the similarity in 2D keypoint configurations. Similarly, depth ambiguities can result in incorrect interpretations of spatial arrangements, such as a monkey's leg extending backward instead of forward (Fig. 8 **(a)**) or misperceiving the orientation of a monkey's head (Fig. 8 **(b)**). These limitations highlight the complexities of single-frame 2D to 3D lifting, where the model's reliance on geometric keypoint arrangements can be deceptive under certain perspectives. Enhanced depth cues and perspective-aware mechanisms are necessary for more accurate single-view 3D reconstructions, pointing towards future directions to integrate additional appearance or visual features based contextual information for resolving these ambiguities.

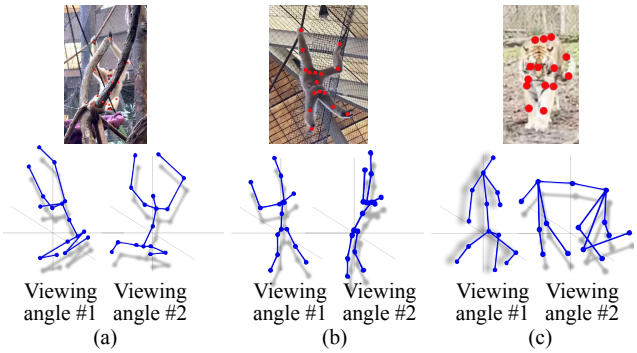


Figure 8. **Challenges in Perspective and Depth Perception**: **(a)** Incorrect leg orientation due to depth ambiguity in monkey capture. **(b)** Misinterpreted head position in a second monkey example. **(c)** A tiger's keypoints distorted by perspective, leading to primate-like 3D predictions."

III. Frequently Asked Questions

Q: How does 3D-LFM handle occlusions?
A: 3D-LFM utilizes a masking strategy alongside Tokenized Positional Encoding (TPE) and joint connectivity to effectively manage occlusions. The model's design allows for accurate 3D predictions and reliable category differentiation even when keypoints are obscured.

Q: Can the 3D-LFM distinguish between different categories under heavy occlusion?
A: Yes, 3D-LFM is designed to distinguish between various categories, such as animals and inanimate objects, even under significant occlusion. This robustness is demonstrated in Fig. 1 and Fig. 5, where the model performs reliably despite the occluded landmarks.

Q: At what point does occlusion begin to affect the model's category identification accuracy?
A: While the model is quite robust to occlusions, there is a threshold beyond which the accuracy begins to diminish. Our ablation study indicates that when more than 60% of landmarks are occluded, the model's ability to accurately identify object categories is compromised due to insufficient data for the TPE and joint connectivity to operate effectively.

Q: Does 3D-LFM use visual features from images for 3D reconstruction?
A: No, 3D-LFM is focused on reconstructing 3D structures from 2D landmarks and does not directly use visual image features. This design choice streamlines the model's training and application to various object categories without the need for visual contextual information.
Q: How does 3D-LFM handle size variations across different object categories?
A: 3D-LFM predicts structures within a canonical frame, where size variations are managed by Procrustean loss. This allows the model to normalize scale differences and ensures consistent 3D predictions across a wide range of object sizes, from large vehicles to smaller animals.
Q: Is 3D-LFM capable of handling sequential inputs, such as videos?
A: While 3D-LFM is primarily designed for single-frame lifting, its architecture does produce stable and coherent outputs over sequences of 2D landmarks. However, it does not explicitly model temporal dynamics, which is an area we aim to explore to improve the model's performance on video data.
Q: What are the potential future enhancements for 3D-LFM?
A: Future enhancements include integrating visual features and temporal dynamics to enhance depth perception and object category differentiation. This will likely improve the robustness and accuracy of the model in more complex, real-world scenarios as discussed in Sec. II of the supplementary material.

IV. More Qualitative Results

Additional qualitative results and project resources can be found on the project's website (https://3dlfm.github.io/).

References

[1] Praneet C Bala, Benjamin R Eisenreich, Seng Bum Michael Yoo, Benjamin Y Hayden, Hyun Soo Park, and Jan Zimmermann. Openmonkeystudio: Automated markerless pose estimation in freely moving macaques. *BioRxiv*, pages 2020–01, 2020. 2, 6

[2] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, pages 690–696. IEEE, 2000. 1

[3] Zheng Chen and Yi Sun. Joint-wise 2d to 3d lifting for hand pose estimation from a single rgb image. *Applied Intelligence*, 53(6):6421–6431, 2023. 2

[4] Mosam Dabhi, Chaoyang Wang, Kunal Saluja, László A Jeni, Ian Fasel, and Simon Lucey. High fidelity 3d reconstructions with limited physical views. In *2021 International Conference on 3D Vision (3DV)*, pages 1301–1311. IEEE, 2021. 2

[5] Mosam Dabhi, Chaoyang Wang, Tim Clifford, László Jeni, Ian Fasel, and Simon Lucey. Mbw: Multi-view bootstrapping in the wild. *Advances in Neural Information Processing Systems*, 35:3039–3051, 2022. 2, 7

[6] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019. 2

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2, 7, 8

[8] Haorui Ji, Hui Deng, Yuchao Dai, and Hongdong Li. Unsupervised 3d pose estimation with non-rigid structure-from-motion modeling. *arXiv preprint arXiv:2308.10705*, 2023. 2, 3

[9] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 2, 7, 8

[10] Daniel Joska, Liam Clark, Naoya Muramatsu, Ricardo Jericevich, Fred Nicolls, Alexander Mathis, Mackenzie W Mathis, and Amir Patel. Acinoset: a 3d pose estimation dataset and baseline models for cheetahs in the wild. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 13901–13908. IEEE, 2021. 2, 7

[11] Chen Kong and Simon Lucey. Deep non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1558–1567, 2019. 1, 2, 3, 5

[12] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81:155–166, 2009. 3

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence 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*, pages 740–755. Springer, 2014. 2

[14] Sebastian Lutz, Richard Blythman, Koustav Ghosal, Matthew Moynihan, Ciaran Simms, and Aljosa Smolic. Jointformer: Single-frame lifting transformer with error prediction and refinement for 3d human pose estimation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1156–1163. IEEE, 2022. 2, 3, 4, 6

[15] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 2, 6

[16] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 2, 3, 6

[17] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020. 2

[18] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7688–7697, 2019. 1, 2, 3, 5

[19] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 3

[20] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[22] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. 4

[23] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13294–13304, 2021. 6

[24] Chaoyang Wang and Simon Lucey. Paul: Procrustean autoencoder for unsupervised lifting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 434–443, 2021. 1, 2, 3, 5

[25] Chaoyang Wang, Chen-Hsuan Lin, and Simon Lucey. Deep nrsfm++: Towards unsupervised 2d-3d lifting in the wild. In *2020 International Conference on 3D Vision (3DV)*, pages 12–22. IEEE, 2020. 1, 2, 4, 5

[26] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014. 2, 5, 7

[27] Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9099–9109, 2023. 2, 6, 7, 8

[28] Haitian Zeng, Xin Yu, Jiaxu Miao, and Yi Yang. Mhr-net: Multiple-hypothesis reconstruction of non-rigid shapes from 2d views. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022. 2, 5

[29] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 2, 7, 8

[30] Jianqiao Zheng, Xueqian Li, Sameera Ramasinghe, and Simon Lucey. Robust point cloud processing through positional embedding. *arXiv preprint arXiv:2309.00339*, 2023. 4

[31] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: Unified pretraining for human motion analysis. *arXiv preprint arXiv:2210.06551*, 2022. 2, 4

[32] Yue Zhu, Nermin Samet, and David Picard. H3wb: Human3. 6m 3d wholebody dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20166–20177, 2023. 2, 5, 6, 8, 9