3D-LFM: Lifting Foundation Model

Dr.Sudipto Roy
Associate Professor, AI & DS
Jio University
Ulwe, Maharashtra, India
sudipta1.roy@jioinstitute.edu.in

Bhushan Gunjal
Student, AI & DS
Jio University
Ulwe, Maharashtra, India
Bhushan.G25pgai@jioinstitute.edu.in

Pravin Patrike
Student, AI & DS
Jio University
Ulwe, Maharashtra, India
Pravin.P25pgai@jioinstitute.edu.in

Varsha Rai
Student, AI & DS
Jio University
Ulwe, Maharashtra, India
Varsha.R25pgai@jioinstitute.edu.in

***Abstract*— Reconstructing 3D shapes from 2D landmarks is a key challenge in computer vision. Traditional methods, such as Perspective-n-Point (PnP) techniques, work well for rigid objects but struggle with diverse and complex shapes. Recent deep learning models, like C3DPO and PAUL, have improved this by handling noise, occlusions, and perspective changes. However, these models rely heavily on matching points between 3D datasets, which limits their use to cases with a lot of labeled 3D data. Our model, called 3D Lifting Foundation Model (3D-LFM), uses transformers with permutation equivariance to overcome this limitation. This allows it to handle varying numbers of points, adapt to occlusions, and even work with new object types not seen in training. We achieve top performance on 2D-to-3D lifting benchmarks, marking 3D-LFM as a flexible, first-of-its-kind solution for general 3D reconstruction tasks.**

Keywords— 3D Reconstruction, 2D-to-3D Lifting, Transformer Architecture, Out-of-Distribution (OOD) Generalization, Token Positional Encoding (TPE)

I. INTRODUCTION

This basic problem is that of changing "landmarks" from a view from one angle into 3D structures. It becomes an important issue for those augmented reality and robotics applications aiming at representing shapes, especially ones that are non-rigid, like human and animal forms. The approaches up to now, however have some reliance on particular object models, which limits them regarding scaling for more varied and complexly shaped objects.

Traditional techniques have improved greatly in tasks like human body and hand modeling but struggle to handle scenes that feature complex interactions among several classes of objects or inconsistent body structures, including different skeletal forms. Among the recent deep learning solutions applied to this problem are the C3DPO and PAUL approaches, but the solutions are still founded upon predefined correspondences between 2D landmarks and the 3D structure. This reliance on matched data greatly limits these models, especially when they scale to a wider set of object categories, with each category having its own distinct configuration.

Our model is called the 3D Lifting Foundation Model (3D-LFM), and we introduce a unified approach that does not depend on specific data for any object category. 3D-LFM applies **permutation equivariance** in transformers, automatically finding patterns among 2D landmarks without prior knowledge of the type of object. This allows 3D-LFM to lift 2D points into accurate 3D structures across 30+ categories using a single model, ranging from human figures and animals to common inanimate objects. Importantly, 3D-LFM generalizes well to new categories not seen in training, a property that is called out-of-distribution (OOD) generalization. For example, it has successfully generated 3D representations of cheetahs and other wild animals by training only on the domestic animal categories such as dogs and cats. In addition, it can accommodate different structural configurations, as found in different human skeleton models, which makes it flexible for new data.

In order to handle such a variety of objects, 3D-LFM relies on Procrustean alignment and tokenized positional encoding (TPE) in its graph-based transformer architecture. Procrustean alignment enables the flexible parts of the objects and discards the unwanted rotation, thus saving computation time. TPE is efficient in the model's adaptation to categories with different landmark numbers, hence scalable. Using these methods with skeleton information, 3D-LFM is able to hold accuracy over all different objects while providing hints over which joints relate to each other, which allows it to handle varied object types.

3D-LFM introduces a new foundation for 2D-to-3D lifting because it is one of the earliest models to reach this level of flexibility. The key contributions of our work are

1. A Procrustean transformer which works with non-rigid features in a constant frame are equivariant to permutation.

2. Tokenized positional encoding inside the graph-based transformer has exploited scalability while yielding good results across different datasets.

3. State of the art performance on human, hand, and face benchmarks and good performance on unseen objects and configurations - animals and inanimate objects.

In the following sections we describe the architecture of 3D-LFM and present extensive analysis through experiments, comparing our model with state-of-the-art models. Here, we use terms such as "keypoints," "landmarks," and "joints" somewhat interchangeably to refer to essential points on an object that define its structure and shape.

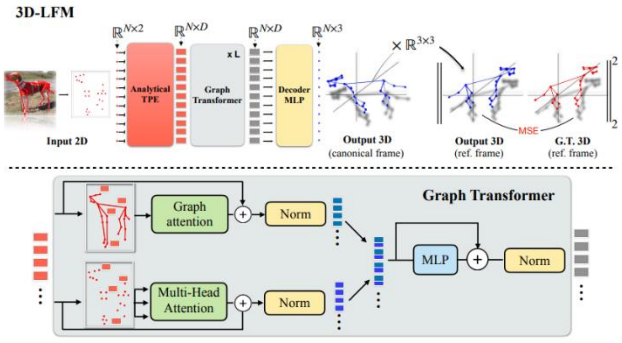

Figure 2. Overview of the 3D Lifting Foundation Model (3D-LFM) architecture: The process begins with the input 2D keypoints undergoing Token Positional Encoding (TPE) before being processed by a series of graph-based transformer layers. The resulting features are then decoded through an MLP into a canonical 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.

II. RELATED WORK

Significantly more sophisticated than the classical PnP-based work [12], the problem of 2D-3D lifting was, initially addressed using a collection of 2D landmarks that received some form of 3D supervision in the guise of known rigid objects in 3D space. Deep learning techniques whose approaches are supported by C3DPO [18], PAUL [24], and Deep NRSfM [11], along with recent transformer-based innovations such as NRSfMFormer [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 particular landmark has the same semantic meaning across all appearances (e.g. chair leg). In practice this is quite limiting at run-time, as one requires a deep understanding of the object category, and rig in order to apply any of these current methods. Further, this greatly hinders the ability of these methods to exploit crossobject and cross-rig datasets, preventing the development of a genuinely generalizable 2D to 3D lifting foundation model – a subject 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 like Jointformer [14] and SimpleBaseline [16] have fine-tuned the single-frame 2D-3D lifting process while generative approaches like MotionCLIP [19] and Human Motion Diffusion Model [20] have laid the ground for Organically object class-bound: The above approaches are even more restrictive than C3PDO, PAUL etc. in that they are organically bound to the object class and are not easily extendable to an arbitrary object class. 3D generative motion-based foundation models..

III. METHODOLOGY

Given a set of 2D keypoints representing the projection of an object's joints in an image, we denote the keypoints matrix as $W \in R^{N\times2}$, 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 $M \in \{0, 1\}^N$, where each element mi of M is defined as

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

The 3D lifting function $f : R^{N\times2} \rightarrow R^{N\times3}$ maps the 2D keypoints to their corresponding 3D structure while compensating for the projection:

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

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

Permutation Equivariance: In order to have scalability and its suitability, to the advantage of adaptation 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 W regardless of order for processes that may have varying joint configurations:

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

where 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 $M \in \{0, 1\}^N$ 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: $W_m = W \odot M$, 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 normalize the zero-centered data to the range [−1, 1] and apply a homography matrix that preserves the aspect ratio to maintain the geometry of the keypoints. To better explain how missing data is dealt with in relation to perspective distortions we refer to Deep NRSFM++[25]. Token Positional Encoding: substitutes classical Correspondence Positional Encoding (CPE) or Joint Embedding used to represent semantic correspondence information (such as [14, 31]) with an does not require explicit correspondence or semantic information. As per-point positional embedding is quite effective, especially random Fourier features [30] on processing OOD data, we compute Token Positional Encoding (TPE) by analytical Random Fourier features (RFF) as follows:

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

where D is the dimensionality of the Fourier feature space, $\omega \in R^{2\times D/2}$ and $b \in R^{D/2}$ 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 TPE(W_c) is then fed into the graph-based transformer network as X^{ℓ} where ℓ indicates the layer number (0 in the above case). This set of features is now ready for processing inside the graphbased 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](L) with global self-attention mechanisms [21](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 \ell , 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, GA(X(ℓ) , A), leverages an adjacency matrix \protect \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 Ul :

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

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 X (L) into the predicted canonical structure Sc via a GeLU-activated MLP:

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

Subsequently, we align Sc with the ground truth Sr , via a Procrustean alignment method that optimizes for the rotation matrix 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 M is a binary mask applied element-wise, and ‖.‖F denotes the Frobenius norm. The optimal 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 Sr, resulting in a scale factor γ, which yields the final predicted shape Sp:

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

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 Sp and the ground truth Sr:

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

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.

IV. RESULTS

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.

4.4.1 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.

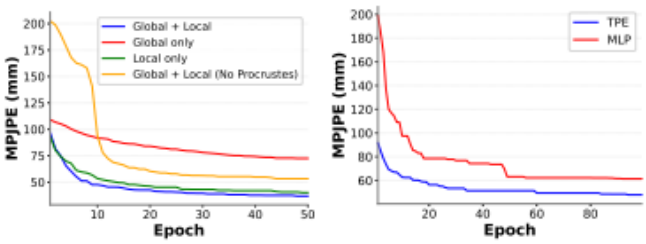

Figure 3. (a) Comparing attention strategies in 3D-LFM. The combined local-global approach with procrustean alignment surpasses other configurations in MPJPE reduction over 100 epochs on the Human3.6M validation split. (b) rapid convergence and efficiency of the TPE approach compared to the learnable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

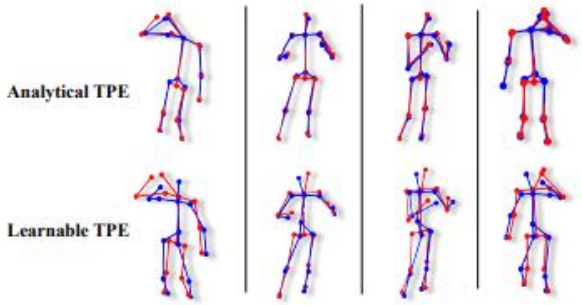

Figure 4. The qualitative improvement in rig transfer using analytical TPE versus learnable MLP projection. This visualization reinforces the necessity of TPE in handling OOD data such as different rigs, unseen during training.

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.

V. DISUSSION AND CONCLUSION

The proposed 3D-LFM model is a great leap in the process of converting 2D images into 3D structures: they are scalable, adaptable, and can handle very varied forms of data, even those with imbalances, generalizing to new categories of objects. The model transfers knowledge across categories very well, although input handling from different viewing perspectives can be optimized even better. 3D-LFM achieves competitive results on standard benchmarks and out-of-distribution scenarios with only a relatively low computational complexity. This model establishes new grounds for 3D pose estimation and reconstruction, becoming a versatile tool for constructing accurate 3D models from 2D inputs in various applications.

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, Laszl ´ o 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] [5] Mosam Dabhi, Chaoyang Wang, Tim Clifford, Laszl ´ o 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-frommotion 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 Dollar, 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 Velickovi ˇ c, Guillem Cucurull, Arantxa Casanova, ´ Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph at- ` tention 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 10474 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