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 benchmark [1].

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. 1 **(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. 1 **(a)**) or misperceiving the orientation of a monkey's head (Fig. 1 **(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.

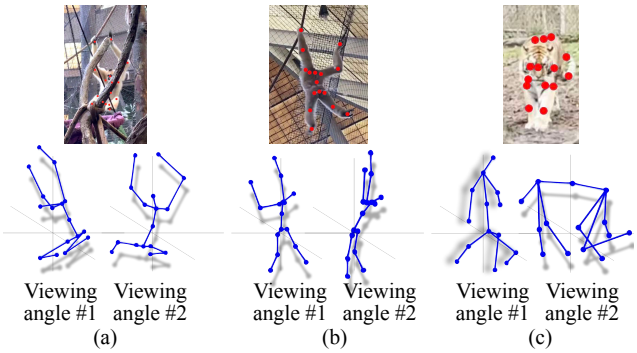

Figure 1. **Challenges in Perspective and Depth Perception**: **(a)** Incorrect leg orientation due to depth ambiguity in monkey capture. **(b)** Misinterpreted head position in a second monkey example. **(c)** A tiger's keypoints distorted by perspective, leading to primate-like 3D predictions."

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] 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. 1