

Object Agnostic 3D Lifting in Space and Time

Supplementary Material

A. Extra training details

We train our model with the Adam [17] optimizer using $\text{lr} = 10^{-4}$, weight decay of 10^{-6} , $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We set the velocity loss term $\lambda = 5000$. We train for a total of 1000 epochs and use the model with the lowest validation error out of all epochs.

B. Dataset details

B.1. Statistics

An overview of our dataset can be seen in Tab. 7. We showcase the number of joints, animations, and frames for each animal category. Our dataset contains a variety of different categories and rigs to facilitate class-agnostic training and evaluation.

B.2. Examples

We provide an additional **dataset_examples.mp4** video in the supplementary zip file for viewing the 3D skeletons for some animals.

C. Additional results

C.1. Examples

We provide an additional **prediction_examples.mp4** video in the supplementary zip file for viewing some predictions over a series of video sequences. We provide a side-by-side comparison between our approach and 3D-LFM. We also provide examples of out-of-distribution predictions. We believe that these videos provide a more comprehensive comparison than that which can be obtained with 2D images.

C.2. Ablating 2D noise

Alongside Tab. 1 in the main paper, we provide additional experiments on our dataset where we do not apply synthetic noise to the 2D inputs. Tab. 8 shows that the improvements provided by our approach over existing methods is not restricted to situations with noisy 2D poses.

C.3. Human3.6M benchmark

We provide comparisons on the Human3.6M [13] benchmark containing 3.6 million video frames of real humans performing simple tasks in a controlled indoor environment. Following previous lifting works [2, 36], we train on subjects 1, 5, 6, 7, and 8, and hold out subjects 9 and 11 for testing. We use the noisy 2D skeletons provided by [36] that were obtained using Stacked Hourglass Networks [24]. Every fifth frame is used for all experiments

with no significant degradation in performance. We additionally compare with the original MotionBERT [36], denoted with \dagger in Tab. 11, that uses human-specific augmentation during training and human-specific semantic correspondences. We find that, when there is an abundance of data for a single object, leveraging object-specific information is preferred to class-agnostic training. Tab. 11 further demonstrates the benefit of our approach over the current SOTA class-agnostic method [8]. We outperform 3D-LFM on each metric, translating to improved 3D object structure and motion consistency across entire sequences. The class-agnostic MotionBERT outperforms both our approach and 3D-LFM, likely due to the having twice as many parameters and an architecture that was specifically designed for large-scale human data.

C.4. Per-animal multi-category

Here we provide the per-animal results that correspond to Tab. 2 in the main paper. Tab. 12 demonstrates the usefulness of training with all categories at once instead of specializing to a single category. We find that this holds true for our model across all animal categories. The same is seen for MotionBERT and 3D-LFM with exception of the chicken category, where it is sometimes slightly better to do chicken-only training.

C.5. Out-of-distribution categories and rigs

Here we provide tabulated results for our out-of-distribution (OOD) experiments found in Sec. 4.2 in the main paper. We show in Tabs. 9 and 10 that our model does indeed maintain superior OOD performance across all metrics compared to existing methods. In the case of unseen number of joints (Tab. 10), we cannot evaluate MotionBERT as it is unable to handle a number of joints that is more than the maximum seen during training.

C.6. Extreme occlusion

We provide a visual comparison between our approach and 3D-LFM when there is an extreme (60%) occlusion of the object. Fig. 7 demonstrates the robustness of our approach in this scenario. While 3D-LFM fails to properly reconstruct the 3D object structure (Fig. 7a) for even a single frame, our method is capable of maintaining high-fidelity reconstruction (Fig. 7b). We show FA-MPJPE results for intermediate occlusion levels in Fig. 8.

| | Bear | Buck | Bunny | Chicken | Deer | Dog | Elk | Fox | Moose | Puma | Rabbit | Raccoon | Tiger | Total |
|------------|-------|-------|-------|---------|-------|-------|-------|-------|-------|-------|--------|---------|-------|--------|
| Animations | 67 | 42 | 45 | 7 | 56 | 65 | 67 | 37 | 59 | 68 | 45 | 54 | 66 | 678 |
| Frames | 4,464 | 3,168 | 3,072 | 432 | 3,648 | 4,128 | 5,328 | 2,304 | 3,792 | 5,808 | 3,072 | 4,176 | 4,992 | 48,384 |
| Joints | 21 | 27 | 25 | 19 | 29 | 22 | 26 | 26 | 29 | 26 | 25 | 28 | 27 | 330 |

Table 7. **An overview of our synthetic dataset.** We measure the total number of animation sequences and frames for each animal. We also provide the number of joints that constitute each animal.

| Method | Bear | Buck | Bunny | Chicken | Deer | Dog | Elk | Fox | Moose | Puma | Rabbit | Raccoon | Tiger | Avg |
|------------|-------------|--------------|-------------|-------------|-------------|-------------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| MotionBERT | 94.9 | 216.1 | 17.7 | 110.7 | 198.4 | 49.1 | 233.1 | 38.0 | 161.8 | 257.4 | 28.1 | 83.1 | 220.8 | 131.5 |
| 3D-LFM | 49.8 | 151.2 | 23.8 | 106.7 | 151.3 | 53.6 | 149.8 | 21.5 | 277.0 | 157.4 | 40.4 | 62.3 | 162.9 | 108.0 |
| Ours | 28.7 | 125.4 | 15.8 | 80.4 | 59.4 | 31.6 | 122.4 | 13.7 | 98.2 | 93.7 | 17.6 | 45.4 | 93.1 | 63.5 |
| MotionBERT | 90.9 | 205.3 | 17.0 | 98.6 | 189.9 | 45.7 | 211.7 | 36.4 | 143.3 | 237.5 | 26.3 | 80.7 | 210.9 | 122.6 |
| 3D-LFM | 31.3 | 100.4 | 12.7 | 97.6 | 67.4 | 33.8 | 104.4 | 15.3 | 113.5 | 114.3 | 26.4 | 49.6 | 107.1 | 67.1 |
| Ours | 26.3 | 108.9 | 10.6 | 70.5 | 52.8 | 26.2 | 98.6 | 11.8 | 78.6 | 87.6 | 15.1 | 43.9 | 81.6 | 54.8 |
| MotionBERT | 2.6 | 9.1 | 0.9 | 3.9 | 7.1 | 2.0 | 9.4 | 1.2 | 13.0 | 9.1 | 1.8 | 3.4 | 8.5 | 5.5 |
| 3D-LFM | 5.3 | 18.5 | 2.3 | 5.7 | 14.1 | 5.2 | 13.8 | 2.3 | 32.3 | 17.0 | 4.8 | 6.6 | 18.9 | 11.6 |
| Ours | 2.2 | 10.1 | 1.2 | 4.2 | 5.0 | 2.5 | 10.8 | 1.0 | 12.6 | 7.8 | 1.7 | 3.1 | 8.7 | 5.4 |

Table 8. **Quantitative comparison when no artificial noise is applied to the 2D keypoints.** We report, in millimeters, the Sequence-Aligned MPJPE (top), Frame-Aligned MPJPE (middle), and Sequence-Aligned MPVE (bottom). Our approach (Ours) consistently outperforms existing methods across all metrics.

| Method | FA-MPJPE ↓ | SA-MPJPE ↓ | SA-MPVE ↓ |
|------------|-------------|-------------|------------|
| MotionBERT | 132.0 | 142.5 | 6.8 |
| 3D-LFM | 112.0 | 143.4 | 21.9 |
| Ours | 75.7 | 86.4 | 6.4 |

Table 9. **OOD generalization on unseen data.** We perform a 13-fold evaluation to assess each method’s ability to handle unseen animal categories. Our approach demonstrates superior proficiency in handling OOD 2D-3D lifting.

| Method | FA-MPJPE ↓ | SA-MPJPE ↓ | SA-MPVE ↓ |
|--------|--------------|--------------|-------------|
| 3D-LFM | 234.4 | 321.7 | 49.0 |
| Ours | 143.0 | 172.9 | 14.5 |

Table 10. **OOD generalization on unseen category and rig.** Models are trained on rigs with 28 or less joints and evaluated on unseen deer and moose categories with 29 joints.

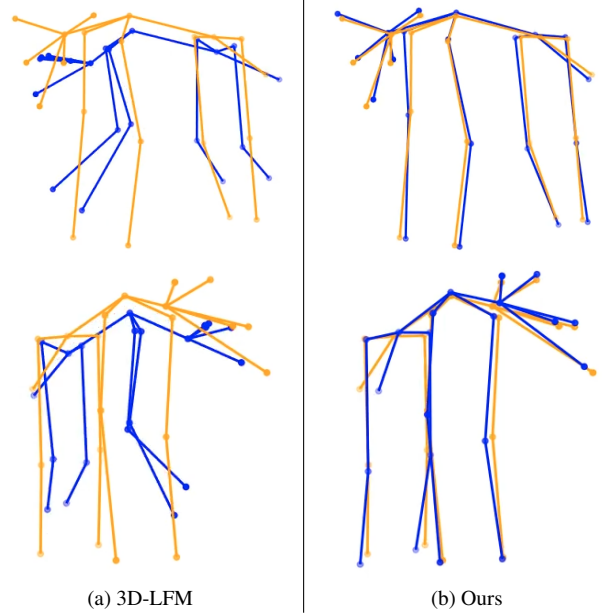


Figure 7. **A comparison of robustness in an extreme case of 60% occlusion.** We showcase the predictions for a bear at 2 different views. Ground truth is orange, prediction is blue.

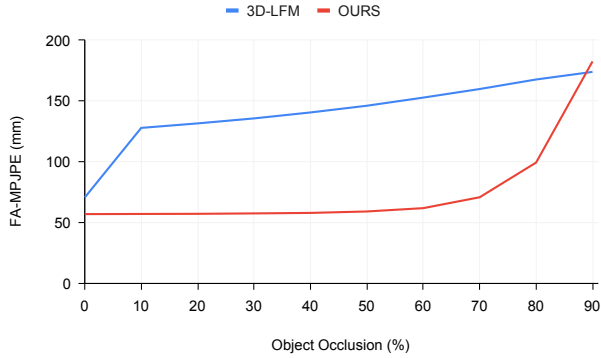


Figure 8. **A comparison of robustness at increasing levels of object occlusion.** Our approach maintains high-fidelity 3D reconstruction, even at extreme levels of occlusion.

| Method | FA-MPJPE↓ | SA-MPJPE↓ | SA-MPVE↓ |
|-------------------------|-----------|-----------|----------|
| MotionBERT [†] | 34.8 | 37.2 | 8.8 |
| MotionBERT | 39.3 | 41.8 | 9.1 |
| 3D-LFM | 48.4 | 63.2 | 29.8 |
| Ours | 42.9 | 49.0 | 11.1 |

Table 11. **Results on the Human3.6M benchmark.** MotionBERT[†] is the original human-specific model. MotionBERT has twice as many parameters and is designed for large-scale human data, while we are designed for small-scale multi-object data.

| Method | MC | Bear | Buck | Bunny | Chicken | Deer | Dog | Elk | Fox | Moose | Puma | Rabbit | Raccoon | Tiger | Avg |
|------------|----|-------------|--------------|-------------|--------------|--------------|-------------|--------------|-------------|--------------|--------------|-------------|-------------|--------------|--------------|
| MotionBERT | - | 142.1 | 315.5 | 36.4 | 110.0 | 297.5 | 108.0 | 293.1 | 70.9 | 381.6 | 288.1 | 94.2 | 107.5 | 353.7 | 199.9 |
| | ✓ | 94.5 | 208.1 | 16.7 | 108.2 | 200.7 | 50.1 | 267.4 | 40.6 | 189.2 | 254.4 | 30.7 | 77.4 | 211.8 | 134.6 |
| 3D-LFM | - | 58.1 | 183.3 | 24.1 | 101.3 | 192.2 | 63.7 | 168.9 | 25.2 | 316.6 | 187.8 | 43.1 | 79.1 | 199.3 | 126.4 |
| | ✓ | 47.6 | 158.2 | 23.2 | 92.3 | 156.8 | 53.9 | 147.8 | 22.2 | 274.7 | 163.4 | 37.8 | 70.0 | 165.4 | 108.7 |
| Ours | - | 70.4 | 205.9 | 22.7 | 94.4 | 152.2 | 78.8 | 193.0 | 38.7 | 271.5 | 158.0 | 47.0 | 67.6 | 269.0 | 128.4 |
| | ✓ | 29.2 | 128.4 | 17.1 | 60.8 | 57.3 | 32.8 | 103.1 | 14.2 | 97.9 | 93.2 | 19.0 | 44.5 | 90.8 | 60.6 |
| MotionBERT | - | 127.5 | 285.0 | 30.6 | 96.4 | 264.4 | 90.9 | 263.0 | 63.8 | 305.9 | 258.4 | 93.6 | 98.4 | 309.6 | 176.0 |
| | ✓ | 90.7 | 198.1 | 16.0 | 99.0 | 195.5 | 45.8 | 246.8 | 39.9 | 170.9 | 235.0 | 28.6 | 74.9 | 203.2 | 126.5 |
| 3D-LFM | - | 38.9 | 138.3 | 13.7 | 89.9 | 115.3 | 45.7 | 123.9 | 21.0 | 177.9 | 152.5 | 27.0 | 67.3 | 148.4 | 89.2 |
| | ✓ | 27.9 | 108.3 | 12.2 | 86.3 | 75.0 | 33.3 | 103.0 | 16.2 | 119.7 | 119.3 | 21.2 | 57.6 | 107.7 | 68.3 |
| Ours | - | 63.1 | 168.5 | 16.9 | 85.0 | 130.1 | 56.1 | 158.3 | 35.1 | 202.1 | 142.5 | 34.8 | 60.0 | 217.2 | 105.4 |
| | ✓ | 26.7 | 107.3 | 11.2 | 54.2 | 50.9 | 27.9 | 86.1 | 12.4 | 81.6 | 85.9 | 15.4 | 42.8 | 79.8 | 52.5 |
| MotionBERT | - | 5.0 | 16.9 | 2.2 | 3.4 | 12.6 | 5.3 | 12.7 | 2.3 | 26.9 | 12.6 | 4.8 | 5.7 | 16.8 | 9.8 |
| | ✓ | 3.2 | 11.0 | 1.1 | 3.9 | 6.9 | 2.6 | 10.4 | 1.3 | 17.8 | 9.8 | 2.1 | 4.0 | 9.8 | 6.5 |
| 3D-LFM | - | 11.3 | 37.3 | 4.0 | 3.6 | 30.9 | 10.3 | 29.7 | 4.8 | 51.4 | 29.7 | 7.1 | 13.2 | 32.6 | 20.5 |
| | ✓ | 7.6 | 29.0 | 3.4 | 8.4 | 26.3 | 8.4 | 26.5 | 3.8 | 43.4 | 27.3 | 6.7 | 12.3 | 30.4 | 18.0 |
| Ours | - | 6.4 | 19.3 | 2.4 | 4.3 | 13.4 | 7.1 | 14.7 | 2.2 | 28.2 | 13.2 | 5.4 | 5.9 | 20.4 | 11.0 |
| | ✓ | 2.5 | 12.1 | 1.3 | 3.4 | 5.9 | 2.9 | 9.3 | 1.2 | 12.5 | 8.9 | 2.0 | 3.6 | 9.1 | 5.7 |

Table 12. **Per-animal comparison of multi-category and single-category training.** We use a ✓ for models trained with multiple categories (MC). We report, in millimeters, the Sequence-Aligned MPJPE (top), Frame-Aligned MPJPE (middle), and Sequence-Aligned MPVE (bottom). Top, middle, and bottom are separated by dual horizontal lines.