C3PO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion

Supplementary material

The first part of the supplementary material contains discussions regarding the role of the camera translation in the formulation of SFM/NR-SFM (appendix A.1), number of degrees of freedom (appendix A.2) and a proof of lemma 1. Additional information about the architecture of the proposed deep networks is in appendix B. Appendix C provides additional analysis of the robustness of C3DPO to the input noise. Appendix D presents additional qualitative results and appendix E discusses failure modes of our method.

A. Theoretical analysis

This section contains additional information regarding various theoretical aspects of the NR-SFM task.

A.1. Centering

This section summarizes well known results on data centering in orthographic SFM and NR-SFM.

Lemma 2. *Equations* $y_{nk} = \Pi R_n X_k + \Pi T_n$ *hold true for all* $n = 1, \dots, N$ *and* $k = 1, \dots, K$ *if, and only if, equations* $\bar{y}_{nk} = \Pi R_n \bar{X}_k$ *hold true, where*

$$\bar{y}_{nk} = y_{nk} - \frac{1}{K}\sum_{k=1}^{K} y_{nk}, \quad \bar{X}_k = X_k - \frac{1}{K}\sum_{k=1}^{K} X_k.$$

Proof. Average and remove the LHS and RHS of each equation from both sides. □

Lemma 3. *Equation* $y_{nk} = \Pi(R_n \sum_{d=1}^{D} \alpha_{nd} S_{dk} + T_n)$ *holds true for all* $n = 1, \dots, N$ *and* $k = 1, \dots, K$ *if, and only if, equation* $\bar{y}_{nk} = \Pi R_n \left(\sum_{d=1}^{D} \alpha_{nd} \bar{S}_{dk}\right)$ *holds true, where*

$$\bar{y}_{nk} = y_{nk} - \frac{1}{K}\sum_{k=1}^{K} y_{nk}, \quad \bar{S}_{dk} = S_{dk} - \frac{1}{K}\sum_{k=1}^{K} S_{dk}.$$

Proof. Average and remove the LHS and RHS of each equation from both sides. □

A.2. Degrees of freedom and ambiguities

Seen as matrix factorization problems, SFM and NR-SFM have *intrinsic ambiguities*; namely, no matter how many points and views are observed, there is always a space of equivalent solutions that satisfy all the observations. Next, we discuss what are these ambiguities and under which conditions they are minimized.

A.2.1 Structure from motion

The SFM eq. (1) contains $2NK$ constraints and $6N + 3K$ unknowns. However, there is an unsolvable ambiguity: $MX = (MA^{-1})(AX)$ means that, if (M, X) is a solution, so (MA^{-1}, AX) is another, for any invertible matrix $A \in \mathbb{R}^{3\times3}$. If X is full rank and there are at least $N \geq 2$ views, we can show that this is the *only* ambiguity, which has 9 degrees of freedom (DoF). Thus finding a unique solution up to these residual 9 DoF requires $2NK \geq 6N + 3K - 9$. For example, with $N = 2$ views, we require $K \geq 3$ keypoints. Furthermore, the 3D point configuration must not be degenerate, in the sense that X must be full rank.

The ambiguity can be further reduced by considering the fact that the view matrices M are not arbitrary; they are instead the first two rows of rotation matrices. We can exploit this fact by setting $M_1 = I_{2\times3}$ (which also standardize the rotation of the first camera), fixing 6 of the 9 DoF.

A.2.2 Non-rigid structure from motion

The NR-SFM equation contains $2NK$ constraints and $6N + ND + 3DK$ unknowns. The intrinsic ambiguity has at least 9 DoF as in the SFM case. Hence, for a unique solution (up to the intrinsic ambiguity) we must have $2NK \geq 6N + ND + 3DK - 9$. Compared to the SFM case, the number of unknowns grows with the number N of views as $(6 + D)N$ instead of just $6N$, where D is the dimension of the shape basis. Since the number of constraints grows as $(2K)N$, we must have $K \geq 3 + D/2$ keypoints.

Note that once the shape basis S is learned, it is possible to perform 3D reconstruction from a single view by solving (3) for $N = 1$; in this case there are $2K$ equations and $6 + D$ unknowns, which is once more solvable when $K \geq 3 + D/2$.

A.3. Proof of lemma 1

Lemma 4. *The set* $\mathcal{X}_0 \subset \mathbb{R}^{3\times K}$ *has the transversal property if, and only if, there exists a* canonicalization function $\Psi : \mathbb{R}^{3\times K} \to \mathbb{R}^{3\times K}$ *such that, for all rotations* $R \in SO(3)$ *and structures* $X \in \mathcal{X}_0$, $X = \Psi(RX)$.

Proof. Assume first that $\mathcal{X}_0$ has the transversal property. Then the function Ψ is obtained by sending each RX for each $X \in \mathcal{X}_0$ back to X. This definition is well posed: if $RX = \bar{R}\bar{X}$ where both $X, \bar{X} \in \mathcal{X}_0$, then $\bar{X} = (\bar{R})^{-1}RX$ and, due to the transversal property, $X = \bar{X}$.

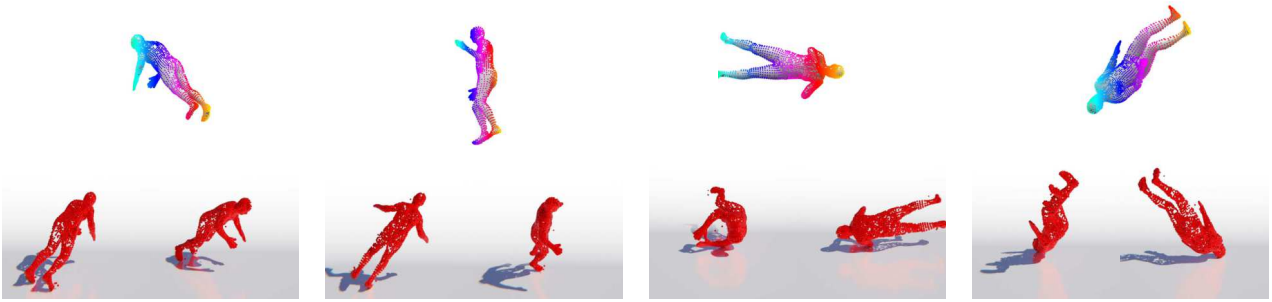

Figure 7: **Qualitative results on S-Up3D** showing input 2D keypoint annotations (top row) and monocular 3D reconstructions of all 6890 vertices of the SMPL model as predicted by C3DPO from two different viewpoints (bottom row).

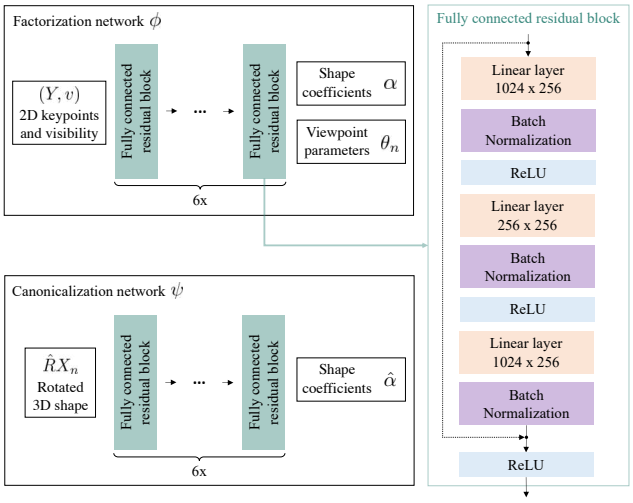

Figure 8: **The architecture of Ψ and Φ.** Both networks share the same trunk (6x fully connected residual layers) and differ in the type of their inputs and outputs.

Assume now that the function Ψ is given and let $X, X' \in \mathcal{X}_0$ such that $X' = RX$ and so $\Phi(X') = \Phi(RX)$. However, by definition, $\Phi(RX) = X$ and $\Phi(X') = \Phi(IX') = X'$, so that $X = X'$. □

B. Architecture of Ψ and Φ

Figure 8 contains a schema of the architecture of Ψ and Φ (both share the same core architecture). It consists of 5 fully connected residual blocks with a kernel size of 1. Empirically, we have observed that using residual blocks, instead of the simpler variant with fully connected layers directly followed by batch normalization and no skip connections, prevents the network from predicting flattened shapes.

C. Analysis of robustness

In order to test the robustness of C3DPO to the noise present in the input 2D keypoints, we devised the following experiment.

We generated several noisy versions of the Synthetic Up3D dataset by adding 2D Gaussian noise (with variance σ) to the 2D input and randomly occluded each 2D input point with probability p_{OCC}. Experiments were ran for different number of input of keypoints (79, 100, 500, 1000) and the evaluation was always conducted on the representative 79 vertices (section 4.1) of S-Up3D-test.

The results of the experiment are depicted in fig. 10. We have observed improved robustness to noise with higher numbers of used keypoints. At the same time, the performance without noise ($\sigma = 0$, $p_{OCC} = 0$) is slightly worse for the setup higher number of keypoints (≥ 500 keypoints). We hypothesize that, when more keypoints are used, the performance deteriorates because the optimizer focuses less on minimizing the reprojection losses of the 79 keypoints that are used for the evaluation.

D. Additional qualitative results

In this section we present additional qualitative results. Figure 7 contains monocular reconstructions of C3DPO trained on the full set of 6890 SMPL vertices of the S-Up3D dataset. Note that we were unable to run [11, 32] on this dataset due to scalability issues of the two algorithms.

E. C3DPO failure modes

The main sources of failures of our method are: (1) Failures of the 2D keypoint detector [19]; (2) Reconstructing "outlier" test 2D poses not seen in training (mainly on Human3.6m); (3) Reconstructing strongly ambiguous 2D poses (in a frontal image of a sitting human, the knee angle cannot be recovered uniquely). The failure mode (1) is depicted in fig. 9.

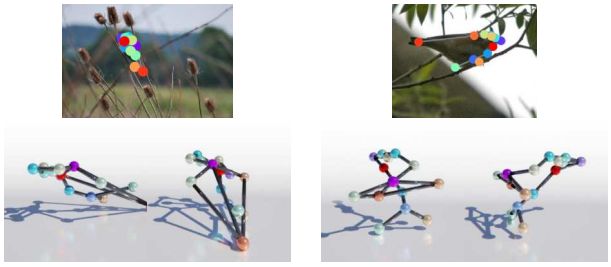

Figure 9: A qualitative example of 2D keypoints lifted by our method. Here, the reconstruction fails due to a failure of the HRNet keypoint detector.

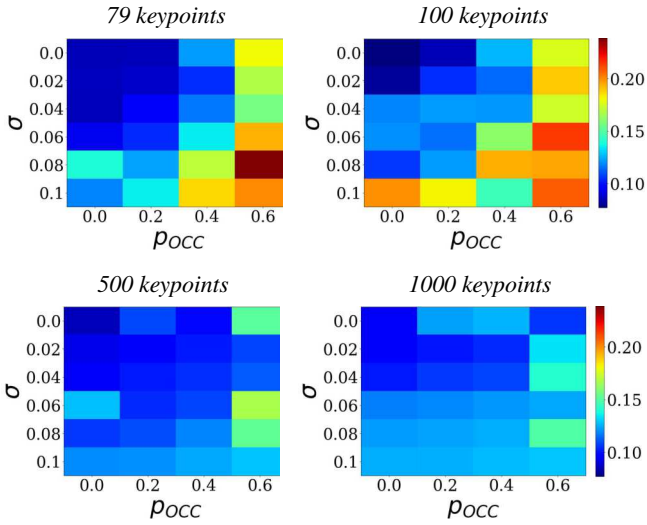

Figure 10: MPJPE on Up3D of C3DPO depending on various levels of Gaussian noise added to 2D inputs (σ-vertical axis) and the probability of occluding an input 2D point (p_{OCC}-horizontal axis) for different numbers of training keypoints (left to right, top to bottom: 79, 100, 500, 1000).