CLOAF: CoLlisiOn-Aware Human Flow

Supplementary Material

The followong supplementary material is organized as follows. In Sec. A, we provide more qualitative illustrations on how CLOAF works. In Sec. B, we discuss the technical details of the ablation study, discussed in the main paper. In Sec. C, we provide implementaiton details on the customized motion field and discuss the comparison between such simple field with the one induced by the neural network.

A. Main Experiments

In Sec. 4.5 in the main paper, we discuss different strategies to obtain the initial body posture Θ_0 for the integration. In Fig. A.1, we demonstrate qualitative comparison between initial bodies obtained with different strategies. *Successive Frames* produces a reasonable estimate, since the pose in neighboring frames is very similar. If *Jitter* manages to find an estimate without self-intersections, then it is also decent, since only a few body parameters are changed, hence, most of the non-intersected body parts are not affected at all. The *Keyposes*-produced neighbor is more distant from the given estimate, though, in some circumstances, it can be an only option.

We have *not* found the choice of the starting pose for *Jitter* to be critical. For *Keyposes*, the sensitivity is larger because the key poses have been chosen to be diverse. Hence, usually only one pose among all yields a good starting estimate. It should be noted that increasing the number of key poses helps only marginally, while making the neighbor search more costly.

In video `CLOAF.mp4`, we compare HMR2.0 with CLOAF used as post-processing on one video sequence of 3DPW-test set. As discussed in Sec. 4.2, CLOAF can be used for efficient removal of self-intersections. The inverse operation of Eq. 6 effectively averages the velocities of all points on the limb producing the realistic motion, even when limbs touch each other and points in contact have the same velocities, as shown in the video.

B. Ablations and Technical details

Optimal sampling. In Ablations in Sec. 4.5, we speculate that the linear approximation of the SMPL transformation holds for small displacements $\delta\Theta$. Here we verify this assumption. As in the main paper, we denote $\boldsymbol{f}$ to be the distance between two SMPL bodies Θ_0 and $\Theta_1 = \Theta_0 + \delta\Theta$, computed in the coordinate space:

$$\boldsymbol{f}(\delta\Theta) = \mathbf{X}(\Theta_0 + \delta\Theta) - \mathbf{X}(\Theta_0), \qquad \text{(B.1)}$$

where the $\delta\Theta$ is the random noise of the magnitude $\|\delta\Theta\|$. The forward SMPL transformation $\mathbf{X}(\Theta)$ is computed with Eq. 3, sampling all vertices in the SMPL mesh $S = 6890$, hence, $\boldsymbol{f} \in \mathbb{R}^{3S}$.

We provide Fig. B.2 for an illustration. We measure the distance $\|\boldsymbol{f}(\delta\Theta)\|$ (*left axis*), averaged over all points, with respect to $\|\delta\Theta\|$ that varies from 10^{-9} to 10^{1}. Additionally, we compute the relative error RE following Eq. 12 (*right axis*). Every number is averaged across 10 restarts of $\delta\Theta$. The black line in the background shows the linear trend. The green area marks the range of values of the motion fields that the pretrained network $\boldsymbol{f}_\omega$ motion gives in our experiments. As assumed, for the entire range of displacements, excluding too large values $\|\delta\Theta\| > 10^{-1}$, the linear approximation holds. For larger values of $\|\delta\Theta\|$, the regime becomes non-linear, and approximation is not valid anymore, however, in our experiments, we never observe such large displacements. In the wide range of $\|\delta\Theta\|$ magnitudes, including the target green area, the relative error RE is low ($RE \lesssim 10^{-1}$). Hence, our inverse procedure is accurate. The ablation experiment in Fig. 5 in the main paper is done for $\|\delta\Theta\| = 10^{-2}$ that corresponds to the upper bound of the target regime (green area), as illustrated in Fig. B.2.

Note that all computations with the inverse Jacobian for Fig. B.2 (as for all experiments of the main paper) are done in double precision, since SMPL transformation is very sensitive to single-precision roundings. The numbers for relative error RE (Eq. 12) and timing in Fig. 5 are done by averaging multiple runs; the number of restarts is 10 for both metrics.

To illustrate what $RE \lesssim 10^{-1}$ looks like, we provide Fig. B.3. We sample "ground-truth" vector $\delta\Theta_{\text{GT}}$ once (values in green) and only vary its magnitude to be 10^{-1} (*left*) and 10^{-2} (*right*). Using our inverse procedure, we reconstruct $\hat{\delta\Theta}$ (values in red). The estimate on the left is much less accurate than the one on the right, which is reflected in the relative error RE ($7.0 \cdot 10^{-1}$ and $7.1 \cdot 10^{-2}$, respectively). We see that the values $RE \lesssim 10^{-1}$ can be seen as an indicator of a proper reconstruction.

Linear interpolation. Precise integration in Eq. 6 requires an estimate of $\Theta(t)$. During training, we use the approximation $\tilde{\Theta}$ to skip an inversion step and to stop error accumulation. This could yield a self-intersecting pose. But this only occurs at training time. At inference time, we compute the entire pose sequence reusing the previous estimates. They are not self-intersecting by construction because we start from a non-penetrated body.

It might be assumed that the method works only when linear interpolant poses are not in self-intersection. How-

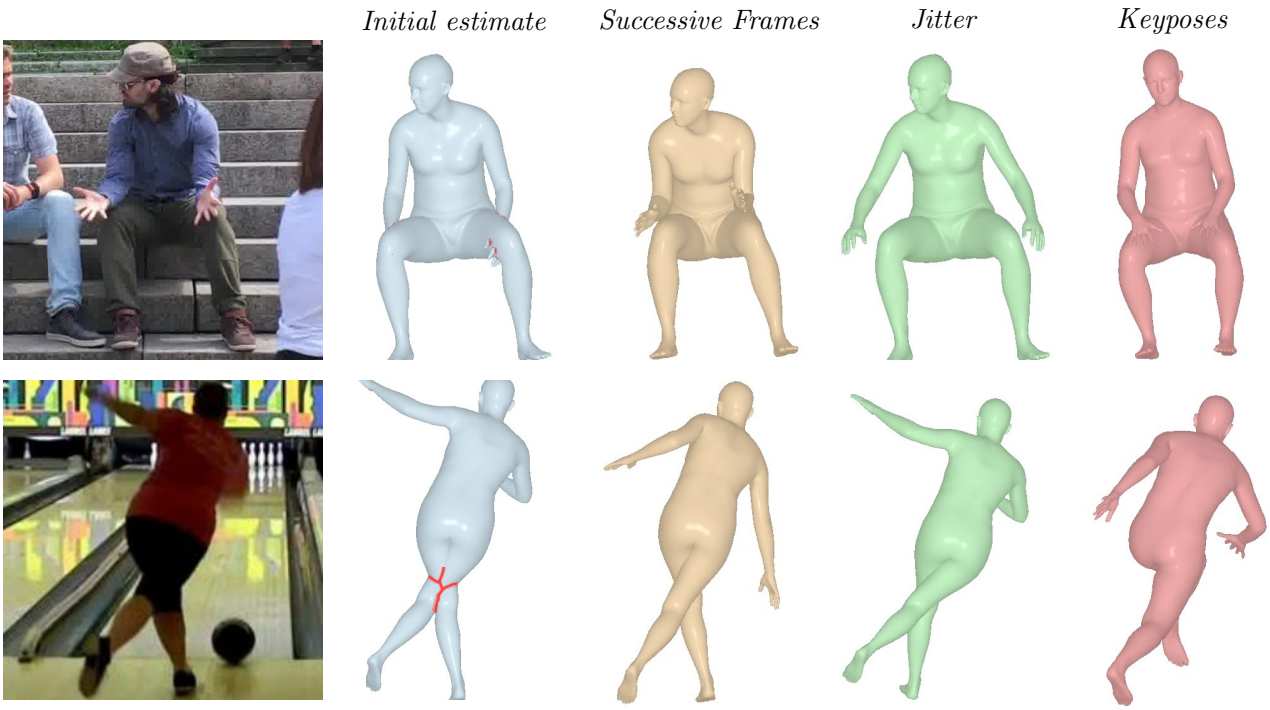

Figure A.1. **Picking the Initial Body Posture.** We demonstrate the examples of the initial body postures Θ_0 to the integration, given the initial estimate that has self-intersections. We compare three strategies for our CLOAF method, Successive Frames, Jitter and Keyposes. None of them has self-penetrations, however, they are different in terms of the distance from the initial estimate.

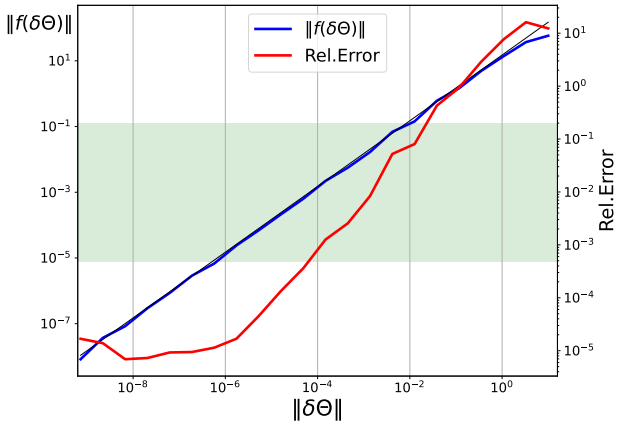

Figure B.2. **SMPL linearity and the Relative Error.** SMPL forward transformation can be seen as linear in the wide range of $\delta\Theta$ deformations. It includes the target area (shown in green) of values that the trained network produces. The Relative Error RE is low in the area as well, which proves our method to be accurate.

ever, the supplementary video `CLOAF.mp4` shows this not to be the case. When the person stands up, the hands move accordingly, even though the linear interpolation is entirely inside the body and should not produce any motion at all.

Technical Details. When dealing with ODEs, efficient integration is a key. To this end, we found that *all* techniques described is Sec. 3 are crucial. Excluding one of these either prevents convergence or makes it too slow. Here we will note a few examples. Without the $\tilde{\Theta}$ approximation, it would take minutes per training sample, instead of less than a second. As for the integration step, it must be small, which is specifically enforced by the solver (Sec. 3) and the optimal sampling (Sec. 4.5). Too long Δt time intervals decrease stability of the training and the network does not converge to a reasonable solution. During training in all our experiments, we sample consecutive poses, hence, $\Delta t_{max} = 1$.

C. Custom Field

In Sec. 4.4, we propose a structure for the simplified motion field that comprises the direction from the selected region towards the target. For the sake of reproducibility, we provide here an exact formulation of such field (only its non-zero component):

$$\boldsymbol{f}(\mathbf{x}) = F\frac{\mathbf{x}_T - \mathbf{x}}{\|\mathbf{x}_T - \mathbf{x}\|_2 + \epsilon}, \tag{C.2}$$

where $\mathbf{x}_T$ is the target point, $\mathbf{x}$ is the point in the selected region, $F = 10^{-3}$ is the magnitude of the field, and $\epsilon = 10^{-6}$

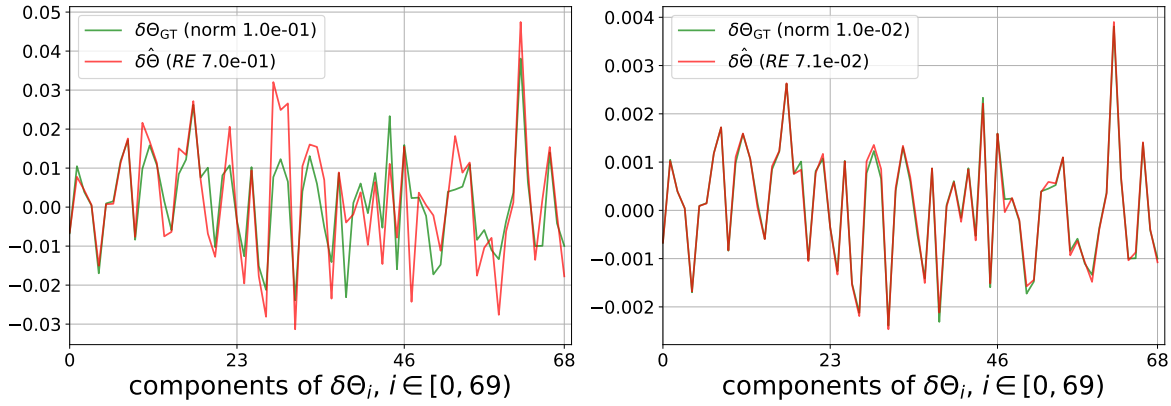

Figure B.3. **Reconstruction of $\delta\Theta$ at different magnitudes.** We randomly sample the vector $\delta\Theta_{\text{GT}}$ and vary only its magnitude to be 10^{-1} and 10^{-2} (green). It brings reconstructions $\delta\hat{\Theta}$ of different quality, $RE = 7.0 \cdot 10^{-1}$ and $RE = 7.1 \cdot 10^{-2}$, respectively (red). The values $RE \lesssim 10^{-1}$ can be seen as an indicator of a proper reconstruction.

is a small regularization for stability. In other words, the field is represented by a vector of a fixed magnitude pointing from the selected region towards the target.

When the non-zero field described by Eq. C.2 is defined, it is blended with the zero field $\mathbf{0}$, as described in Sec. 4.4 in the main paper. The inner and outer regions r_{in} and r_{out} are 10 and 30 *mm*, respectively. The larger values make the field less precise, while the smaller values make it more localized, hence, less smoother, which complexifies the integration.

In the supplementary video `grab_the_box.mp4`, we demonstrate the interaction with objects discussed in Sec. 4.4 and depicted in Fig. 4 in the main paper. The video shows the integration process for two fields, with and without constraints.

Simple Field vs. NN. In the main experiments we exploit a pre-trained neural network to induce the motion field, while later we demonstrate that the simpler field can be used to produce customized motions. The natural question arises: why not to use the simple field for the main experiments without any neural networks at all? The answer is that the neural network is more flexible and can produce more complex motions than the target-driven field, as in Eq. C.2. As discussed in Sec. 3, the network learns to produce an interpolation in the parametric space, while moving in the coordinate space. The integration of the simple field can be seen as an interpolation in the coordinate space, while optimizing a very simple energy function $E = \|\mathbf{x}(\Theta) - \mathbf{x}(\Theta_1)\|$, where Θ_1 is the target pose, with infinitesimal steps $\delta\Theta$. Such a simple field easily gets stuck in local optima, preventing further improvement.

Yet, as shown in Sec 4.4, the simple field approach allows for customization of motion, a task that is not as easily achieved with a neural network.