Learning 3D Human Dynamics from Video: Supplementary Material

1. Model architecture

Temporal Encoder Figure 1 visualizes the architecture of our temporal encoder f_{movie}. Each 1D convolution has temporal kernel size 3 and filter size 2048. For group norm, we use 32 groups each with 64 channels. We repeat the residual block 3 times, which gives us a field of view of 13 frames.

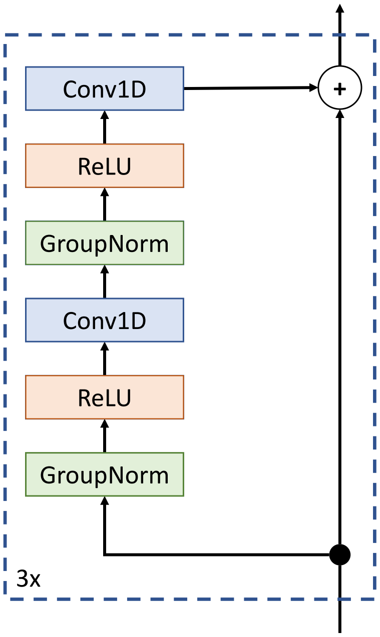

Figure 1: **Architecture of the temporal encoder f_{movie}.**

Hallucinator Our hallucinator consists of two fully-connected layers of 2048 neurons, whose output gets added to the original ϕ as a skip connection.

3D regressors Our f_{3D} regresses the 85D Θ_t vector in an iterative error feedback (IEF) loop [2, 4], where the current estimates are progressively updated by the regressor. Specifically, the regressor takes in the current image feature ϕ_t and current parameter estimate $\Theta_t^{(j)}$, and outputs corrections $\Delta\Theta_t^{(j)}$. The current estimate gets updated by this correction $\Theta_t^{(j+1)} = \Delta\Theta_t^{(j)} + \Theta_t^{(j)}$. This loop is repeated 3 times. We initialize the $\Theta_t^{(0)}$ to be the mean values $\bar{\Theta}$, which we also update as a learned parameter.

The regressor consists of two fully-connected layers, both with 1024 neurons, with a dropout layer in between, followed by a final layer that outputs the 85D outputs. All weights are shared.

The dynamics predictors $f_{\pm\Delta t}$ has a similar form, except it only outputs the 72-D changes in pose θ, and the initial estimate is set to the prediction of the current frame t, *i.e.* $\theta_{t+\Delta t}^{(0)} = \theta_t$. Each $f_{\pm\Delta t}$ learns a separate set of weights.

2. Additional Ablations and Evaluations

In Table 1, we evaluate our method and comparable methods on 2D/3D pose and 3D shape recovery. We provide another ablation of our approach where the constant shape loss (Eq. 1) is not used (Ours – Const). In addition, we include full results from our ablation studies.

Shape Evaluation To measure shape predictions, we report *Posed Mesh Error* (Mesh Pos), which computes the mean Euclidean distance between the predicted and ground truth 3D meshes. Since this metric is affected by the quality of the pose predictions, we also report *Unposed Mesh Error* (Mesh Unp), which computes the same but with a fixed T-pose to evaluate shape independently of pose accuracy. Both metrics are in units of *mm*. Note that accurately capturing the shape of the subject is challenging since only 4 ground truth shapes are available in Human3.6M when training.

3. Failure Modes

While our experiments show promising results, there is still room for improvement.

Smoothing Overall, our method obtains smooth results, but it can struggle in challenging situations, such as person-to-person occlusions or fast motions. Additionally, extreme or rare poses (*e.g.* stretching, ballet) are difficult to capture. Please refer to our supplementary video for examples.

	3DPW						H3.6M			Penn Action
	PCK ↑	MPJPE ↓	PA-MPJPE ↓	Accel Err ↓	Mesh Pos ↓	Mesh Unp ↓	MPJPE ↓	PA-MPJPE ↓	Accel Err ↓	PCK ↑
Martinez *et al*. [5]	-	-	157.0	-	-	-	62.9	47.7	-	-
SMPLify [1]	-	199.2	106.1	-	211.2	61.2	-	82.3	-	-
TP-Net [3]	-	163.7	92.3	-	-	-	**52.1**	**36.3**	-	-
Ours	86.4	127.1	80.1	16.4	144.4	25.8	87.0	58.1	9.3	77.9
Ours + VLOG	91.7	126.7	77.7	15.7	147.4	29.7	85.9	58.3	9.3	78.6
Ours + InstaVariety	**92.9**	**116.5**	**72.6**	**14.3**	**138.6**	26.7	83.7	56.9	9.3	**78.7**
Single-view retrained [4]	84.1	130.0	76.7	37.4	144.9	**24.4**	94.0	59.3	23.9	73.2
Ours – Dynamics	82.6	139.2	78.4	15.2	155.2	24.8	88.6	58.3	**9.1**	71.2
Ours – Const	86.5	128.3	78.2	16.6	145.9	27.5	83.5	57.8	9.3	78.1

Table 1: **Evaluation of baselines, ablations, and our proposed method on 2D and 3D keypoints and 3D mesh.** We compare with three other feed-forward methods that predict 3D joints. None of the models are trained on 3DPW, all of the models are trained on H3.6M, and only our models are trained on Penn Action (TP-Net also uses MPII 2D dataset). We show that training with pseudo-ground truth 2D annotations significantly improves 2D and 3D predictions on the in-the-wild video dataset 3DPW. Single-view is retrained on our data. Ours – Dynamics is trained without the past and future regressors $f_{\pm\Delta t}$. Ours – Const is trained without $L_{\text{const shape}}$.

Dynamics Prediction Clearly, predicting the past and future dynamics from a single image is a challenging problem. Even for us humans, from a single image alone, many motions are ambiguous. Figure 2 visualizes a canonical example of such ambiguity, where it is unclear from the input center image if she is about to raise her arms or lower them. In these cases, our model learns to predict constant pose.

Furthermore, even the pose in a single image can be ambiguous, for example due to motion blur in videos. Figure 3 illustrates a typical example, where the tennis player's arm has disappeared and therefore the model cannot discern whether the person is facing left or right. When the current frame prediction is poor, the resulting dynamics predictions are also not correct, since the dynamics predictions are initialized from the pose of the current frame.

Note that incorporating temporal context resolves many of these static-image ambiguities. Please see our supplementary video for examples.

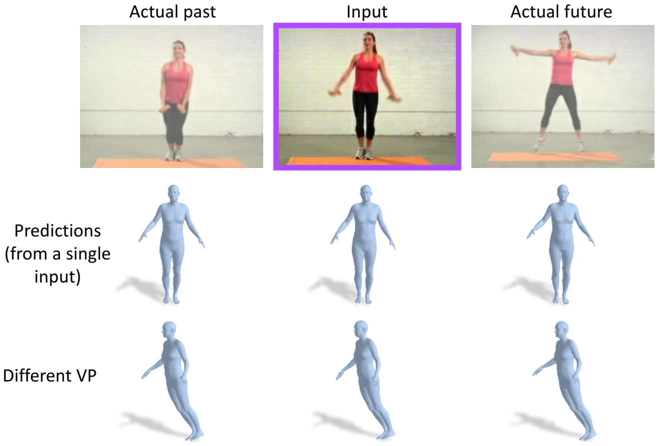

Figure 2: **Ambiguous motion**. Dynamic prediction is difficult from the center image alone, where her arms may reasonably lift or lower in the future.

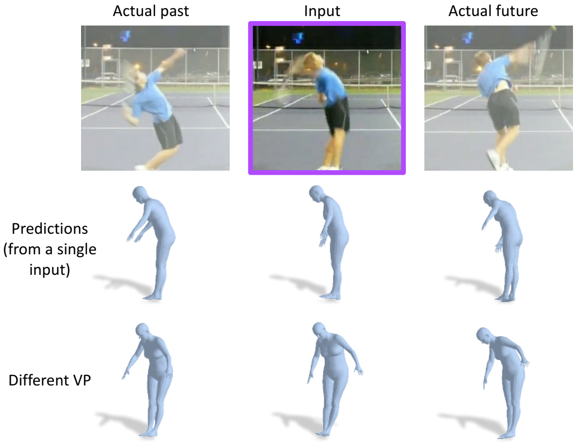

Figure 3: **Ambiguous pose**. The tennis player's pose in the input, center image is difficult to disambiguate between hunched forward verses arched backward due to the motion blur. This makes it challenging for our model to recover accurate dynamics predictions from the single image.

References

[1] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 2

[2] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 1

[3] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, and A. Jain. Structure-aware and temporally coherent 3d human pose estimation. *ECCV*, 2018. 2

[4] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2

[5] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 2