

Learning 3D Human Dynamics from Video

Angjoo Kanazawa*, Jason Y. Zhang*, Panna Felsen*, Jitendra Malik
 University of California, Berkeley

{kanazawa, zhang.j, panna, malik}@eecs.berkeley.edu

Abstract

From an image of a person in action, we can easily guess the 3D motion of the person in the immediate past and future. This is because we have a mental model of 3D human dynamics that we have acquired from observing visual sequences of humans in motion. We present a framework that can similarly learn a representation of 3D dynamics of humans from video via a simple but effective **temporal encoding of image features**. At test time, from video, the learned temporal representation give rise to **smooth 3D mesh** predictions. From a single image, our model can recover the current 3D mesh as well as its 3D past and future motion. Our approach is designed so it can learn from videos with 2D pose annotations in a semi-supervised manner. Though annotated data is always limited, there are millions of videos uploaded daily on the Internet. In this work, we harvest this Internet-scale source of unlabeled data **by training our model on unlabeled video with pseudo-ground truth 2D pose obtained from an off-the-shelf 2D pose detector**. Our experiments show that adding more videos with pseudo-ground truth 2D pose monotonically improves 3D prediction performance. We evaluate our model on the recent challenging dataset of 3D Poses in the Wild and obtain state-of-the-art performance on the 3D prediction task without any fine-tuning. The project website with video can be found at https://akanazawa.github.io/human_dynamics/.

1. Introduction

Consider the image of the baseball player mid-swing in Figure 1. Even though we only see a flat two-dimensional picture, we can infer the player’s 3D pose, as we can easily imagine how his knees bend and arms extend in space. Furthermore, we can also infer his motion in the surrounding moments as he swings the bat through. We can do this because we have a mental model of 3D human dynamics that we have acquired from observing many examples of people in motion.

* equal contribution

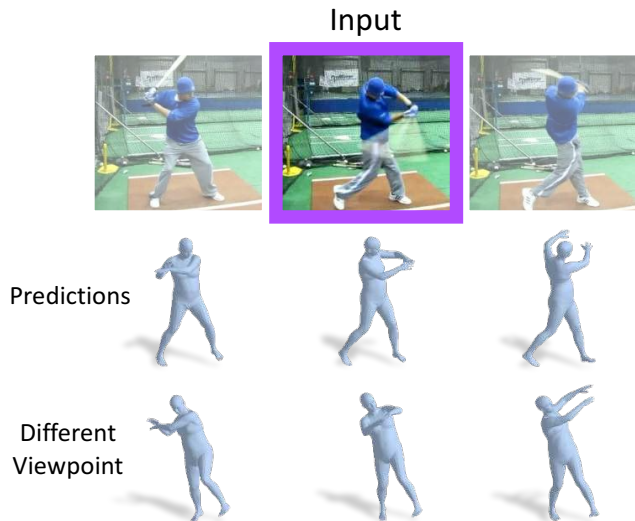


Figure 1: **3D motion prediction from a single image**. We propose a method that, given a single image of a person, predicts the 3D mesh of the person’s body and also hallucinates the future and past motion. Our method can learn from videos with only 2D pose annotations in a **semi-supervised manner**. Note our training set does not have any ground truth 3D pose sequences of batting motion. Our model also produces smooth 3D predictions from video input.

In this work, we present a computational framework that can similarly learn a model of 3D human dynamics from video. **Given a temporal sequence of images, we first extract per-image features, and then train a simple 1D temporal encoder that learns a representation of 3D human dynamics over a temporal context of image features**. We force this representation to capture 3D human dynamics by predicting not only the **current 3D human pose and shape**, but also **changes in pose in the nearby past and future frames**. We transfer the learned 3D dynamics knowledge to static images by learning a hallucinator that can hallucinate the temporal context representation from a single image feature. The hallucinator is trained in a self-supervised manner using the actual output of the temporal encoder. Figure 2 illustrates the overview of our training procedure.

At test time, when the input is a video, the temporal en-

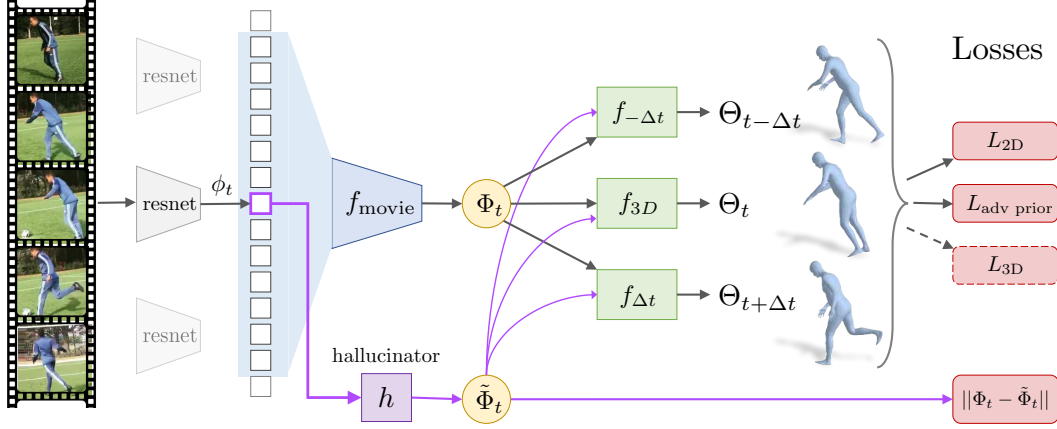


Figure 2: **Overview of the proposed framework.** Given a temporal sequence of images, we first extract per-image features ϕ_t . We train a temporal encoder f_{movie} that learns a representation of **3D human dynamics** Φ_t over the temporal window centered at frame t , illustrated in the blue region. From Φ_t , we predict the 3D human pose and shape Θ_t , as well as the change in pose in the nearby $\pm\Delta t$ frames. The primary loss is 2D reprojection error, **with an adversarial prior to make sure that the recovered poses are valid**. We incorporate 3D losses when 3D annotations are available. We also train a hallucinator h that takes a single image feature ϕ_t and learns to hallucinate its **temporal representation** $\tilde{\Phi}_t$. At test time, the hallucinator can be used to predict dynamics from a single image.

coder can be used to produce **smooth 3D predictions**: having a temporal context reduces **uncertainty and jitter** in the 3D prediction inherent in single-view approaches. The encoder provides the benefit of learned smoothing, which reduces the acceleration error by 56% versus a comparable single-view approach on a recent dataset of 3D humans in the wild. Our approach also obtains state-of-the-art 3D error on this dataset without any fine-tuning. When the input is a single image, the hallucinator can predict the current 3D human mesh as well as the change in 3D pose in nearby future and past frames, as illustrated in Figure 1.

We design our framework so that it can be trained on various types of supervision. A major challenge in 3D human prediction from a video or an image is that **3D supervision is limited in quantity and challenging to obtain at a large scale**. Videos with 3D annotations are often captured in a controlled environment, and **models trained on these videos alone do not generalize to the complexity of the real world**. When 3D ground truth is not available, our model can be trained with 2D pose annotations via the **reprojection loss** [58] and an **adversarial prior** that constrains the 3D human pose to **lie in the manifold of real human poses** [30]. However, the amount of video labeled with ground truth 2D pose is still limited because ground truth annotations are costly to acquire.

While annotated data is always limited, there are millions of videos uploaded daily on the Internet. In this work we harvest this **potentially** unlimited source of unlabeled videos. We curate two large-scale video datasets of humans and train on this data using **pseudo-ground truth** 2D pose obtained from a state-of-the-art 2D pose detector [10]. Excitingly, our experiments indicate that adding more videos with pseudo-ground truth 2D monotonically **improves the**

model performance both in term of 3D pose and 2D reprojection error: 3D pose error reduces by 9% and 2D pose accuracy increases by 8%. Our approach **falls in the category of** omni-supervision [44], a subset of semi-supervised learning where the learner exploits all data along with Internet-scale unlabeled data. We **distill** the knowledge of an accurate 2D pose detector into our 3D predictors through unlabeled video. While omni-supervision has been shown to improve 2D recognition problems, as far as we know, our experiment is the first to show that training on pseudo-ground truth 2D pose labels improves 3D prediction.

In summary, we propose a simple but effective temporal encoder that learns to capture 3D human dynamics. The learned representation allows smooth 3D mesh predictions from video in a feed-forward manner. The learned representation can be transferred to a static image, where from a single image, we can predict the current 3D mesh as well as the change in 3D pose in nearby frames. We further show that our model can leverage an Internet-scale source of unlabeled videos using pseudo-ground truth 2D pose.

2. Related Work

3D pose and shape from a single image. Estimating 3D body pose and shape from a single image is a fundamentally ambiguous task that most methods deal by using some model of human bodies and priors. **Seminal works** in this area [21, 47, 2] rely on **silhouette features** or manual interaction from users [47, 22, 64] to fit the parameters of a statistical body model. A fully automatic method was proposed by Bogo *et al.* [8], which fits the parametric SMPL [35] model to **2D joint locations** detected by an off-the-shelf 2D pose detector [43] with strong priors. Lassner *et*

al. [31] extend the approach to fitting predicted **silhouettes**. [62] explore the multi-person setting. Very recently, multiple approaches integrate the SMPL body model within a deep learning framework [50, 48, 40, 30, 39], where models are trained to directly infer the SMPL parameters. These methods vary in the cues they use to infer the 3D pose and shape: **RGB image** [48, 30], **RGB image and 2D keypoints** [50], **keypoints and silhouettes** [40], or **keypoints and body part segmentations** [39]. Methods that employ **silhouettes** obtain more accurate shapes, but require that the person is **fully visible and unoccluded in the image**. Varol *et al.* explore predicting a voxel representation of human body [51]. In this work we go beyond these approaches by proposing a method that can **predict shape and pose from a single image**, as well as how the **body changes locally in time**.

3D pose and shape from video. While there are more papers that utilize video, most rely on a multi-view setup, which requires significant instrumentation. We focus on videos obtained from a monocular camera. Most approaches take a two-stage approach: first obtaining a single-view 3D reconstruction and then post-processing the result to be smooth via solving a constrained optimization problem [65, 57, 45, 46, 26, 37, 42]. Recent methods obtain accurate shapes and textures of clothing by pre-capturing the actors and making use of silhouettes [49, 59, 23, 4]. While these approaches obtain far more accurate shape, reliance on the **pre-scan** and **silhouettes restricts** these approaches to videos obtained in an interactive and **controlled environments**. Our approach is **complementary** to these two-stage approaches, since all predictions can be post-processed and refined. There are some recent works that output smooth 3D pose and shape: [50] predicts SMPL parameters from two video frames by using **optical flow**, **silhouettes**, and **keypoints** in a self-supervised manner. [3] exploits optical flow to obtain temporally coherent human poses. [29] fits a body model to a sequence of 3D point clouds and 3D joints obtained from multi-view stereo. Several approaches train LSTM models on various inputs such as image features [34], 2D joints [25], or 3D joints [12] to obtain temporally coherent 3D joint outputs. More recently, TP-Net [13] learns a fully convolutional network that smooths the predicted 3D joints. Concurrently to ours, [41] use a fully convolutional network to predict 3D joints from 2D joint sequences. We directly predict the 3D mesh outputs from 2D image sequences and can be trained with images without any ground truth 3D annotation. Furthermore, our temporal encoder predicts the 3D pose changes in nearby frames in addition to the current 3D pose. Our experiments indicate that the prediction losses help the encoder to pay more attention to the dynamics information available in the temporal window.

Learning motion dynamics. There are many methods that predict 2D future outputs from video using **pixels** [16, 15],

flow [54], or **2D pose** [56]. Other methods predict 3D future from 3D inputs [18, 28, 9, 33, 52]. In contrast, our work **predicts future and past 3D pose** from 2D inputs. There are several approaches that predict future from a single image [55, 60, 11, 32, 19], but all approaches predict future in 2D domains, while in this work we propose a framework that predicts 3D motions. Closest to our work is that of Chao *et al.* [11], who forecast 2D pose and then estimate the 3D pose from the predicted 2D pose. In this work, **we predict dynamics directly in the 3D space and learn the 3D dynamics from video**.

3. Approach

Our goal is to learn a representation of 3D human dynamics from video, from which we can 1) obtain **smooth 3D prediction** and 2) **hallucinate 3D motion** from static images. In particular, we develop a framework that can learn 3D human dynamics from unlabeled, everyday videos of people on the Internet. We first define the problem and discuss different tiers of data sources our approach can learn from. We then present our framework that learns to encode **3D human motion dynamics** from videos. Finally, we discuss how to transfer this knowledge to **static images** such that one can hallucinate short-term human dynamics from a static image. Figure 2 illustrates the framework.

3.1. Problem Setup

Our input is a video $V = \{I_t\}_{t=1}^T$ of length T , where each frame is a **bounding-box crop centered around a detected person**. We encode the t th image frame I_t with a visual feature ϕ_t , obtained from a pretrained feature extractor. We train a function f_{movie} that learns a representation Φ_t that encodes the 3D dynamics of a human body given a temporal context of image features centered at frame t . Intuitively, Φ_t is the representation of a “movie strip” of 3D human body in motion at frame t . We also learn a hallucinator $h : \phi_t \mapsto \Phi_t$, whose goal is to hallucinate the movie strip representation from a static image feature ϕ_t .

We ensure that the movie strip representation Φ_t captures the 3D human body dynamics by predicting the 3D mesh of a human body from Φ_t at different time steps. The 3D mesh of a human body in an image is represented by 85 parameters, denoted by $\Theta = \{\beta, \theta, \Pi\}$, which consists of shape, pose, and camera parameters. We use the SMPL body model [35], which is a function $\mathcal{M}(\beta, \theta) \in \mathbb{R}^{N \times 3}$ that outputs the $N = 6890$ vertices of a triangular mesh given the shape β and pose θ . Shape parameters $\beta \in \mathbb{R}^{10}$ define the linear coefficients of a low-dimensional statistical shape model, and pose parameters $\theta \in \mathbb{R}^{72}$ define the **global rotation** of the body and the **3D relative rotations** of the kinematic skeleton of 23 joints in axis-angle representation. Please see [35] for more details. The mesh vertices define 3D locations of k joints $X \in \mathbb{R}^{k \times 3} = \mathcal{W}\mathcal{M}(\beta, \theta)$ via a

pre-trained linear regressor $W \in \mathbb{R}^{k \times N}$. We also solve for the **weak-perspective camera** $\Pi = [s, t_x, t_y]$ that projects the body into the image plane. We denote $x = \Pi(X(\beta, \theta))$ as the projection of the 3D joints.

While this is a well-formed supervised learning task if the ground truth values were available for every video, such 3D supervision is **costly** to obtain and **not available in general**. Acquiring 3D supervision requires extensive instrumentation such as a **motion capture (MoCap) rig**, and these videos captured in a **controlled environment** do not reflect **the complexity of the real world**. While more practical solutions are being introduced [53], 3D supervision is not available for millions of videos that are being uploaded daily on the Internet. In this work, we wish to harness this potentially infinite data source of unlabeled video and propose a framework that **can learn 3D motion from pseudo-ground truth 2D pose predictions obtained from an off-the-shelf 2D pose detector**. Our approach can learn from three tiers of data sources at once: First, we use the MoCap datasets $\{(V_i, \Theta_i, x_i)\}$ with full 3D supervision Θ_i for each video along with ground truth 2D pose annotations for k joints $x_i = \{x_t \in \mathbb{R}^{k \times 2}\}_{t=1}^T$ in each frame. Second, we use datasets of videos in the wild obtained from a monocular camera with human-annotated 2D pose: $\{(V_i, x_i)\}$. Third, we also experiment with videos with *pseudo*-ground truth 2D pose: $\{(V_i, \tilde{x}_i)\}$. See Table 1 for the list of datasets and their details.

3.2. Learning 3D Human Dynamics from Video

A dynamics model of a 3D human body captures how the body changes in 3D over a small change in time. Therefore, we formulate this problem as learning a temporal representation that can simultaneously **predict the current 3D body and pose changes in a short time period**. To do this, we learn a temporal encoder f_{movie} and a 3D regressor f_{3D} that predict the 3D human mesh representation at the current frame, as well as delta 3D regressors $f_{\Delta t}$ that predict how the 3D pose changes in $\pm \Delta t$ time steps.

Temporal Encoder Our temporal encoder consists of several layers of a **1D fully convolutional network** f_{movie} that encodes a temporal window of image features centered at t into a representation Φ_t that encapsulates the 3D dynamics. We use a **fully convolutional model** for its simplicity. Recent literature also suggests that **feed-forward convolutional models empirically out-perform recurrent models while being parallelizable and easier to train with more stable gradients** [7, 38]. Our temporal convolution network has a ResNet [24] based architecture similar to [7, 1].

The output of the temporal convolution network is sent to a 3D regressor $f_{3D} : \Phi_t \mapsto \Theta_t$ **that predicts the 3D human mesh representation at frame t** . We use the same iterative **3D regressor** architecture proposed in [30]. Simply having a

temporal context reduces ambiguity in 3D pose, shape, and viewpoint, resulting in a temporally smooth 3D mesh reconstruction. In order to train these modules from 2D pose annotations, we employ the **reprojection loss** [58] and the **adversarial prior** proposed in [30] to constrain the output pose to lie in the space of possible human poses. The 3D losses are also used when 3D ground truth is available. Specifically, the loss for the current frame consists of the reprojection loss on visible keypoints $L_{2D} = \|v_t(x_t - \hat{x}_t)\|_2^2$, where $v_t \in \mathbb{R}^{k \times 2}$ is the visibility indicator over each keypoint, the 3D loss if available, $L_{3D} = \|\Theta_t - \hat{\Theta}_t\|_2^2$, and **the factorized adversarial prior of [30]**, which trains a discriminator D_k for each joint rotation of the body model $L_{\text{adv prior}} = \sum_k (D_k(\Theta) - 1)^2$. In this work, we regularize the shape predictions using a **shape prior** $L_{\beta \text{ prior}}$ [8]. Together the loss for frame t consists of $L_t = L_{2D} + L_{3D} + L_{\text{adv prior}} + L_{\beta \text{ prior}}$.

Furthermore, each sequence is of the same person, so while the pose and camera may change every frame, the shape remains constant. We express this constraint as a constant shape loss over each sequence:

$$L_{\text{const shape}} = \sum_{t=1}^{T-1} \|\beta_t - \beta_{t+1}\|. \quad (1)$$

Predicting Dynamics We enforce that the learned temporal representation captures the 3D human dynamics by predicting the 3D pose changes in a local time step $\pm \Delta t$. Since we are training with videos, we readily have the 2D and/or 3D targets at nearby frames of t to train the dynamics predictors. Learning to predict 3D changes encourages the network to pay more attention to the temporal cues, and our experiments show that adding this auxiliary loss improves the 3D prediction results. Specifically, given a movie strip representation of the temporal context at frame Φ_t , our goal is to learn a dynamics predictor $f_{\Delta t}$ that predicts the change in 3D parameters of the human body at time $t \pm \Delta t$.

In predicting dynamics, **we only estimate the change in 3D pose parameters θ , as the shape should remain constant and the weak-perspective camera accounts for where the human is in the detected bounding box**. In particular, to improve the robustness of the current pose estimation during training, we augment the image frames with random jitters in scale and translation which emulates the noise in real human detectors. However, such noise should not be modeled by the dynamics predictor.

For this task, we propose a dynamics predictor $f_{\Delta t}$ that outputs the **72D change in 3D pose $\Delta \theta$** . $f_{\Delta t}$ is a function that maps Φ_t and the predicted current pose θ_t to the predicted change in pose $\Delta \theta$ for a specific time step Δt . The delta predictors are trained such that the predicted pose in the new timestep $\theta_{t+\Delta t} = \theta_t + \Delta \theta$ minimizes the reprojection, 3D, and the adversarial prior losses at time frame $t + \Delta t$. We use the shape predicted in the current time t

to obtain the mesh for $t \pm \Delta t$ frames. To compute the re-projection loss without predicted camera, we solve for the **optimal scale s and translation \vec{t} that aligns the orthographically projected 3D joints** $x_{\text{orth}} = X[:, :2]$ with the visible ground truth 2D joints $x_{gt} : \min_{s, \vec{t}} \|(sx_{\text{orth}} + \vec{t}) - x_{gt}\|_2$. A closed form solution exists for this problem, and we use the optimal camera $\Pi^* = [s^*, \vec{t}^*]$ to compute the reprojection error on poses predicted at times $t \pm \Delta t$. Our formulation factors away axes of variation, such as shape and camera, so that the delta predictor focuses on learning the temporal evolution of 3D pose. In summary, the overall objective for the temporal encoder is

$$L_{\text{temporal}} = \sum_t L_t + \sum_{\Delta t} L_{t+\Delta t} + L_{\text{const shape}}. \quad (2)$$

In this work we experiment with two Δt at $\{-5, 5\}$ frames, which amounts to ± 0.2 seconds for a 25 fps video.

3.3. Hallucinating Motion from Static Images

Given the framework for learning a representation for 3D human dynamics, we now describe how to transfer this knowledge to static images. The idea is to learn a hallucinator $h : \phi_t \mapsto \tilde{\Phi}_t$ that **maps a single-frame representation ϕ_t to its “movie strip” representation $\tilde{\Phi}_t$** . One advantage of working with videos is that during training, the target representation Φ_t is readily available for every frame t from the temporal encoder. Thus, the hallucinator can be trained in a weakly-supervised manner, minimizing the difference between the hallucinated movie strip and the actual movie strip obtained from f_{movie} :

$$L_{\text{hal}} = \|\Phi_t - \tilde{\Phi}_t\|_2. \quad (3)$$

Furthermore, we pass the hallucinated movie strip to the f_{3D} regressor to minimize the single-view loss as well as the delta predictors $f_{\Delta t}$. **This ensures that the hallucinated features are not only similar to the actual movie strip but can also predict dynamics.** All predictor weights are shared among the actual and hallucinated representations.

In summary we jointly train the temporal encoder, hallucinator, and the **delta 3D predictors** together with overall objective:

$$L = L_{\text{temporal}} + L_{\text{hal}} + L_t(\tilde{\Phi}_t) + \sum_{\Delta t} L_{t+\Delta t}(\tilde{\Phi}_t). \quad (4)$$

See Figure 2 for the overview of our framework.

4. Learning from Unlabeled Video

Although our approach can be trained on 2D pose annotations, annotated data is always limited – the annotation effort for labeling keypoints in videos is substantial. However, millions of videos are uploaded to the Internet every day. On YouTube alone, 300 hours of video are uploaded every minute [6].

Dataset Name	Total Frames	Total Length (min)	Avg. Length (sec)	Annotation Type		
				GT 3D	GT 2D	In-the-wild
Human3.6M	581k	387	48	✓	✓	
Penn Action	77k	51	3		✓	✓
NBA (Ours)	43k	28	3		✓	✓
VLOG peop. (Ours)	353k	236 (4 hr)	8			✓
InstaVariety (Ours)	2.1M	1459 (1 day)	6			✓

Table 1: **Three tiers of video datasets.** We jointly train on videos with: full ground truth 2D and 3D pose supervision, only ground truth 2D supervision, and pseudo-ground truth 2D supervision. Note the difference in scale for pseudo-ground truth datasets.

Therefore, we curate two Internet-scraped datasets with pseudo-ground truth 2D pose obtained by running OpenPose [10]. An added advantage of OpenPose is that it detects toe points, which are not labeled in any of the video datasets with 2D ground truth. Our first dataset is VLOG-people, a subset of the VLOG lifestyle dataset [17] on which OpenPose fires consistently. To get a more diverse range of human dynamics, we collect another dataset, InstaVariety, from Instagram using 84 hashtags such as *#instruction*, *#swimming*, and *#dancing*. A large proportion of the videos we collected contain only one or two people moving with much of their bodies visible, so OpenPose produced reasonably good quality 2D annotations. For videos that contain multiple people, we form our pseudo-ground truth by linking the per-frame skeletons from OpenPose using the Hungarian algorithm-based tracker from Detect and Track [20]. A clear advantage of unlabeled videos is that they can be easily collected at a significantly larger scale than videos with human-annotated 2D pose. Altogether, our pseudo-ground truth data has over 28 hours of 2D-annotated footage, compared to the 79 minutes of footage in the human-labeled datasets. See Table 1 for the full dataset comparison.

5. Experimental Setup

Architecture: We use Resnet-50 [24] pretrained on single-view 3D human pose and shape prediction [30] as our feature extractor, where $\phi_i \in \mathbb{R}^{2048}$ is **the the average pooled features of the last layer**. Since training on video requires a large amount of memory, we precompute the image features on each frame similarly to [1]. This allow us to train on 20 frames of video with mini-batch size of 8 on a single 1080ti GPU. Our temporal encoder consists of **1D temporal convolutional layers**, where each layer is a **residual block** of two 1D convolutional layers of kernel width of 3 with **group norm**. We use **three** of these layers, producing an effective receptive field size of **13 frames**. The final output of the temporal encoder has the same feature dimension as

ϕ . Our hallucinator contains **two fully-connected layers** of size 2048 with **skip connection**. Please see the supplementary material for more details.

Datasets: **Human3.6M** [27] is the only dataset with ground truth 3D annotations that we train on. It consists of motion capture sequences of actors performing tasks in a controlled lab environment. We follow the standard protocol [30] and train on 4 subjects (S1, S6, S7, S8) and test on 2 subjects (S9, S11) with 1 subject (S5) as the validation set.

For in-the-wild video datasets with 2D ground truth pose annotations, we use the **Penn Action** [63] dataset and our own **NBA** dataset. Penn Action consists of 15 sports actions, with 1257 training videos and 1068 test. We set aside 10% of the test set as validation. The NBA dataset contains videos of basketball players attempting 3-point shots in 16 basketball games. Each sequence contains one set of 2D annotations for a single player. We split the dataset into 562 training videos, 64 validation, and 151 test. Finally, we also experiment with the new pseudo-ground truth 2D datasets (Section 4). See Table 1 for the summary of each dataset. Unless otherwise indicated, all models are trained with Human3.6M, Penn Action, and NBA.

We evaluate our approach on the recent 3D Poses in the Wild dataset (**3DPW**) [53], which contains 61 sequences (25 train, 25 test, 12 val) of indoor and outdoor activities. Portable IMUs provide ground truth 3D annotations on challenging in-the-wild videos. To remain comparable to existing methods, we do not train on 3DPW and only used it as a test set. For evaluations on all datasets, we skip frames that have fewer than 6 visible keypoints.

As our goal is not human detection, we assume a temporal tube of human detections is available. We use ground truth 2D bounding boxes if available, and otherwise use the output of OpenPose to obtain a temporally smooth tube of human detections. **All images are scaled to 224x224 where the humans are roughly scaled to be 150px in height.**

6. Experiments

We first evaluate the efficacy of the learned temporal representation and compare the model to local approaches that only use a single image. We also compare our approaches to state-of-the-art 3D pose methods on 3DPW. We then evaluate the effectiveness of training on pseudo-ground truth 2D poses. Finally, we quantitatively evaluate the dynamics prediction from a static image on Human3.6M. We show qualitative results on video prediction in Figure 3 and static image dynamics prediction in Figure 1 and 4. Please see the supplementary for more ablations, metrics, and discussion of failure modes. In addition, a video with more of our results is available at <https://youtu.be/9fNKSZdsAG8>.

6.1. Local vs Temporal Context

We first evaluate the proposed temporal encoder by comparing with a single-view approach that only sees a local window of one frame. As the baseline for the local window, we use a model similar to [30], re-trained on the same training data for a fair comparison. We also run an ablation by training our model with our temporal encoder but without the dynamics predictions $f_{\Delta t}$.

In order to measure smooth predictions, we propose an **acceleration error**, which measures the average difference between ground truth 3D acceleration and predicted 3D acceleration of each joint in mm/s^2 . This can be computed on 3DPW where ground truth 3D joints are available. On 2D datasets, we simply report the acceleration in mm/s^2 .

We also report other standard metrics. For 3DPW, we report the mean per joint position error (MPJPE) and the MPJPE after Procrustes Alignment (PA-MPJPE). Both are measured in millimeters. On datasets with only 2D ground truth, we report accuracy in 2D pose via percentage of correct keypoints [61] with $\alpha = 0.05$.

We report the results on three datasets in Table 2. Overall, we find that our method produces modest gains in 3D pose estimation, large gains in 2D, and a very significant improvement in acceleration error. The temporal context helps to resolve ambiguities, producing smoother, temporally consistent results. Our ablation study shows that access to temporal context alone is not enough; using the auxiliary dynamics loss is important to force the network to learn the *dynamics* of the human.

Comparison to state-of-the-art approaches. In Table 3, we compare our approach to other state-of-the-art methods. None of the approaches train on 3DPW. Note that Martinez *et al.* [36] performs well on the Human3.6M benchmark but achieves the worst performance on 3DPW, showing that methods trained exclusively on Human3.6M do **not generalize** to in-the-wild images. We also compare our approach to TP-Net, a recently-proposed semi-supervised approach that is trained on Human3.6M and MPII 2D pose in-the-wild dataset [5]. TP-Net also learns a temporal smoothing network supervised on Human3.6M. While this approach is highly competitive on Human3.6M, our approach significantly out-performs TP-Net on in-the-wild video. We only compare feed-forward approaches and not methods that smooth the 3D predictions via post-optimization. Such post-processing methods are complementary to feed-forward approaches and would benefit any of the approaches.

6.2. Training on pseudo-ground truth 2D pose

Here we report results of models trained on the two Internet-scale datasets we collected with pseudo-ground



Figure 3: **Qualitative results of our approach on sequences from Penn Action, NBA, and VLOG.** For each sequence, the top row shows the cropped input images, the middle row shows the predicted mesh, and the bottom row shows a different angle of the predicted mesh. Our method produces smooth, temporally consistent predictions.

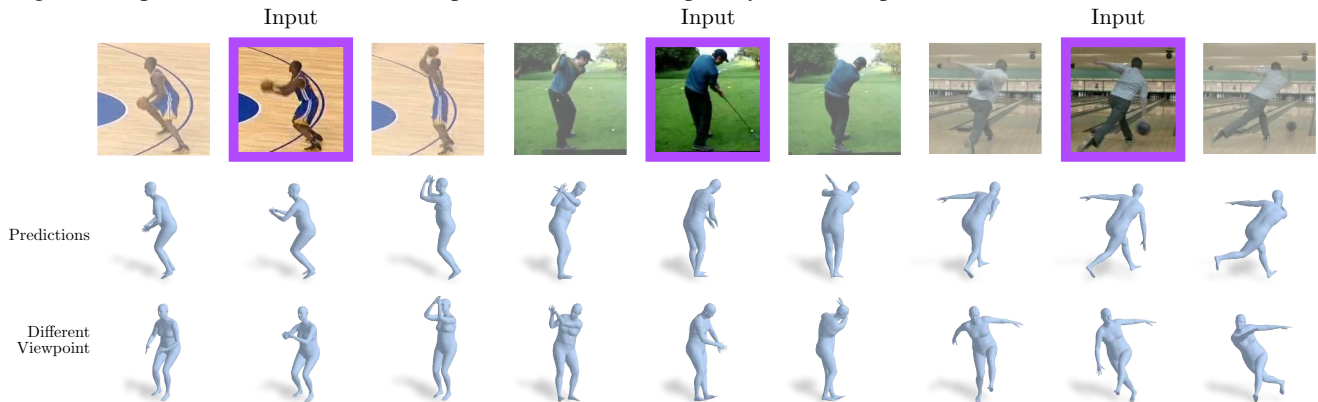


Figure 4: **Predicting 3D dynamics.** In the top row, the boxed image is the single-frame input to the hallucinator while the left and right images are the ground truth past and future respectively. The second and third rows show two views of the predicted meshes for the past, present, and future given the input image.

	3DPW				NBA		Penn Action	
	PCK \uparrow	MPJPE \downarrow	PA-MPJPE \downarrow	Accel Error \downarrow	PCK \uparrow	Accel	PCK \uparrow	Accel
Single-view retrained [30]	84.1	130.0	76.7	37.4	55.9	163.6	73.2	79.9
Context. no dynamics	82.6	139.2	78.4	15.2	64.2	46.6	71.2	29.3
Contextual	86.4	127.1	80.1	16.4	68.4	44.1	77.9	29.7

Table 2: **Local vs temporal context.** Our temporal encoder produces smoother predictions, significantly lowering the acceleration error. We also find that training for dynamic prediction considerably improves 2D keypoint estimation.

	3DPW		H36M
	MPJPE \downarrow	PA-MPJPE \downarrow	PA-MPJPE \downarrow
Martinez <i>et al.</i> [36]	-	157.0	47.7
SMPLify [8]	199.2	106.1	82.3
TP-Net [14]	163.7	92.3	36.3
Ours	127.1	80.1	58.1
Ours + InstaVariety	116.5	72.6	56.9

Table 3: **Comparison to state-of-the-art 3D pose reconstruction approaches.** Our approach achieves state-of-the-art performance on 3DPW. Good performance on Human3.6M does not always translate to good 3D pose prediction on in-the-wild videos.

	3DPW			NBA	Penn
	PCK \uparrow	MPJPE \downarrow	PA-MPJPE \downarrow	PCK \uparrow	PCK \uparrow
Ours	86.4	127.1	80.1	68.4	77.9
Ours + VLOG	91.7	126.7	77.7	68.2	78.6
Ours + InstaVariety	92.9	116.5	72.6	68.1	78.7

Table 4: **Learning from unlabeled video via pseudo ground truth 2D pose.** We collected our own 2D pose datasets by running OpenPose on unlabeled video. Training with these pseudo-ground truth datasets induces significant improvements across the board.

truth 2D pose annotations (See Table 4). We find that the adding more data monotonically improves the model performance both in terms of 3D pose and 2D pose reprojection error. Using the largest dataset, InstaVariety, 3D pose error reduces by 9% and 2D pose accuracy increases by 8% on 3DPW. We see a small improvement or no change on 2D datasets. It is encouraging to see that not just 2D but also 3D pose improves from pseudo-groundtruth 2D pose annotations.

6.3. Predicting dynamics

We quantitatively evaluate our static image to 3D dynamics prediction. Since there are no other methods that predict 3D poses from 2D images, we propose two baselines: a constant baseline that outputs the current frame prediction for both past and future, and an Oracle Nearest Neighbors baseline. We evaluate our method on Human3.6M and compare with both baselines in Table 5.

	Past	Current	Future
	PA-MPJPE \downarrow	PA-MPJPE \downarrow	PA-MPJPE \downarrow
N.N.	71.6	50.9	70.7
Const.	68.6	58.1	69.3
Ours 1	65.0	58.1	65.3
Ours 2	65.7	60.7	66.3

Table 5: **Evaluation of dynamic prediction on Human3.6M.** The Nearest Neighbors baseline uses the pose in the training set with the lowest PA-MPJPE with the ground truth current pose to make past and future predictions. The constant baseline uses the current prediction as the future and past predictions. Ours 1 is the prediction model with Eq. 3, Ours 2 is that without Eq. 3.

Clearly, predicting dynamics from a static image is a challenging task due to inherent ambiguities in pose and the stochasticity of motion. Our approach works well for ballistic motions in which there is no ambiguity in the direction of the motion. **When it's not clear if the person is going up or down our model learns to predict no change.**

7. Discussion

We propose an end-to-end model that learns a model of 3D human dynamics that can 1) obtain smooth 3D prediction from video and 2) hallucinate 3D dynamics on single images at test time. We train a simple but effective temporal encoder from which the current 3D human body as well as how the 3D pose changes can be estimated. Our approach can be trained on videos with 2D pose annotations in a semi-supervised manner, and we show empirically that our model can improve from training on an Internet-scale dataset with pseudo-groundtruth 2D poses. While we show promising results, much more remains to be done in recovering 3D human body from video. Upcoming challenges include dealing with occlusions and interactions between multiple people.

Acknowledgements We thank David Fouhey for providing us with the people subset of VLOG, Rishabh Dabral for providing the source code for TP-Net, Timo von Marcard and Gerard Pons-Moll for help with 3DPW, and Heather Lockwood for her help and support. This work was supported in part by Intel/NSF VEC award IIS-1539099 and BAIR sponsors.

References

- [1] T. Afouras, J. S. Chung, and A. Zisserman. Deep lip reading: A comparison of models and an online application. In *Interspeech*, pages 3514–3518, 2018. 4, 5
- [2] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *TPAMI*, 28(1):44–58, 2006. 2
- [3] T. Alldieck, M. Kassubeck, B. Wandt, B. Rosenhahn, and M. Magnor. Optical flow-based 3d human motion estimation from monocular video. In *GCPR*, pages 347–360. Springer, 2017. 3
- [4] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people model. In *CVPR*, 2018. 3
- [5] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, June 2014. 6
- [6] S. Aslam. Youtube by the numbers. <https://www.omnicoreagency.com/youtube-statistics/>, 2018. Accessed: 2018-05-15. 5
- [7] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 4
- [8] 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, 4, 8
- [9] J. Bütepage, M. J. Black, D. Kragic, and H. Kjellström. Deep representation learning for human motion prediction and classification. In *CVPR*, page 2017. IEEE, 2017. 3
- [10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2, 5
- [11] Y.-W. Chao, J. Yang, B. L. Price, S. Cohen, and J. Deng. Forecasting human dynamics from static images. In *CVPR*, pages 3643–3651, 2017. 3
- [12] H. Coskun, F. Achilles, R. S. DiPietro, N. Navab, and F. Tombari. Long short-term memory kalman filters: Recurrent neural estimators for pose regularization. In *ICCV*, 2017. 3
- [13] R. Dabral, A. Mundhada, U. Kusupati, S. Afaq, and A. Jain. Structure-aware and temporally coherent 3d human pose estimation. *ECCV*, 2018. 3
- [14] R. Dabral, A. Mundhada, U. Kusupati, S. Afaq, A. Sharma, and A. Jain. Learning 3d human pose from structure and motion. In *ECCV*, 2018. 8
- [15] E. L. Denton et al. Unsupervised learning of disentangled representations from video. In *NeurIPS*, pages 4414–4423, 2017. 3
- [16] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *NeurIPS*, pages 64–72, 2016. 3
- [17] D. F. Fouhey, W. Kuo, A. A. Efros, and J. Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018. 5
- [18] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *ICCV*, pages 4346–4354, 2015. 3
- [19] R. Gao, B. Xiong, and K. Grauman. Im2flow: Motion hallucination from static images for action recognition. In *CVPR*, 2018. 3
- [20] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran. Detect-and-Track: Efficient Pose Estimation in Videos. In *CVPR*, 2018. 5
- [21] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV*, page 641. IEEE, 2003. 2
- [22] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *ICCV*, pages 1381–1388. IEEE, 2009. 2
- [23] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. Reticam: Real-time human performance capture from monocular video, 2018. 3
- [24] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer, 2016. 4, 5
- [25] M. R. I. Hossain and J. J. Little. Exploiting temporal information for 3d human pose estimation. In *ECCV*, pages 69–86. Springer, 2018. 3
- [26] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black. Towards accurate marker-less human shape and pose estimation over time. In *International Conference on 3D Vision (3DV)*, pages 421–430, 2017. 3
- [27] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 6
- [28] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, pages 5308–5317, 2016. 3
- [29] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, pages 8320–8329, 2018. 3
- [30] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 3, 4, 5, 6, 8
- [31] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, July 2017. 3
- [32] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Flow-grounded spatial-temporal video prediction from still images. In *ECCV*, 2018. 3
- [33] Z. Li, Y. Zhou, S. Xiao, C. He, Z. Huang, and H. Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *ICLR*, 2018. 3
- [34] M. Lin, L. Lin, X. Liang, K. Wang, and H. Cheng. Recurrent 3d pose sequence machines. In *CVPR*, pages 5543–5552. IEEE, 2017. 3
- [35] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *SIG-GRAPH Asia*, 2015. 2, 3
- [36] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 6, 8

- [37] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. In *SIGGRAPH*, July 2017. 3
- [38] J. Miller and M. Hardt. Stable recurrent models. *ICLR*, 2019. 4
- [39] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*, 2018. 3
- [40] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 3
- [41] D. Pavlo, C. Feichtenhofer, D. Grangier, and M. Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019. 3
- [42] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine. Sfv: Reinforcement learning of physical skills from videos. *SIGGRAPH Asia*, 37(6), Nov. 2018. 3
- [43] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, pages 4929–4937, 2016. 2
- [44] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He. Data distillation: Towards omni-supervised learning. *CVPR*, 2018. 2
- [45] A. Rehan, A. Zaheer, I. Akhter, A. Saeed, M. H. Usmani, B. Mahmood, and S. Khan. Nrsfm using local rigidity. In *WACV*, pages 69–74. IEEE, 2014. 3
- [46] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *ECCV*, pages 509–526. Springer, 2016. 3
- [47] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NeurIPS*, pages 1337–1344, 2008. 2
- [48] J. K. V. Tan, I. Budvytis, and R. Cipolla. Indirect deep structured learning for 3d human shape and pose prediction. In *BMVC*, 2017. 3
- [49] M. Trumble, A. Gilbert, A. Hilton, and J. Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *ECCV*, pages 784–800, 2018. 3
- [50] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *NeurIPS*, pages 5242–5252, 2017. 3
- [51] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 3
- [52] R. Villegas, J. Yang, D. Ceylan, and H. Lee. Neural kinematic networks for unsupervised motion retargeting. In *CVPR*, 2018. 3
- [53] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, sep 2018. 4, 6
- [54] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*, pages 835–851. Springer, 2016. 3
- [55] J. Walker, A. Gupta, and M. Hebert. Dense optical flow prediction from a static image. In *ICCV*, pages 2443–2451, 2015. 3
- [56] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017. 3
- [57] B. Wandt, H. Ackermann, and B. Rosenhahn. 3d reconstruction of human motion from monocular image sequences. *TPAMI*, 38(8):1505–1516, 2016. 3
- [58] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. In *ECCV*, 2016. 2, 4
- [59] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *SIGGRAPH*, 37(2):27:1–27:15, May 2018. 3
- [60] T. Xue, J. Wu, K. Bouman, and B. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NeurIPS*, pages 91–99, 2016. 3
- [61] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *TPAMI*, 35(12):2878–2890, 2013. 6
- [62] A. Zanfir, E. Marinou, and C. Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, pages 2148–2157, 2018. 3
- [63] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *CVPR*, pages 2248–2255, 2013. 6
- [64] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. Parametric reshaping of human bodies in images. In *SIGGRAPH*, page 126. ACM, 2010. 2
- [65] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, pages 4966–4975, 2016. 3