VIBE: Video Inference for Human Body Pose and Shape Estimation

Muhammed Kocabas[1,2], Nikos Athanasiou[1], Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Max Planck ETH Center for Learning Systems
{mkocabas,nathanasiou,black}@tue.mpg.de

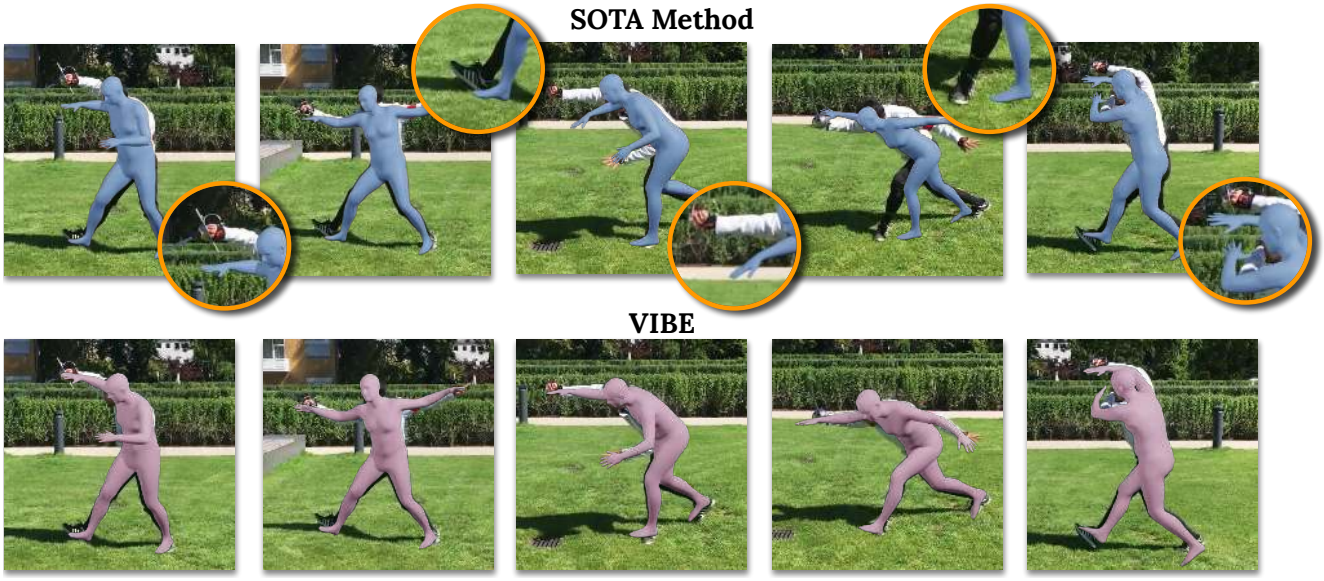

Figure 1: Given challenging in-the-wild videos, a recent state-of-the-art video-pose-estimation approach [30] *(top)*, fails to produce accurate 3D body poses. To address this, we exploit a large-scale motion-capture dataset to train a motion discriminator using an adversarial approach. Our model (VIBE) *(bottom)* is able to produce realistic and accurate pose and shape, outperforming previous work on standard benchmarks.

Abstract

*Human motion is fundamental to understanding behavior. Despite progress on single-image 3D pose and shape estimation, existing video-based state-of-the-art methods fail to produce accurate and natural motion sequences due to a lack of ground-truth 3D motion data for training. To address this problem, we propose "**V**ideo **I**nference for **B**ody Pose and Shape **E**stimation" (VIBE), which makes use of an existing large-scale motion capture dataset (AMASS) together with unpaired, in-the-wild, 2D keypoint annotations. Our key novelty is an adversarial learning framework that leverages AMASS to discriminate between real human motions and those produced by our temporal pose and shape regression networks. We define a novel temporal network architecture with a self-attention mechanism and show that adversarial training, at the sequence level, produces kinematically plausible motion sequences without in-the-wild ground-truth 3D labels. We perform extensive experimentation to analyze the importance of motion and demonstrate the effectiveness of **VIBE** on challenging 3D pose estimation datasets, achieving state-of-the-art performance. Code and pretrained models are available at https://github.com/mkocabas/VIBE*

1. Introduction

Tremendous progress has been made on estimating 3D human pose and shape from a single image [11, 21, 25, 29, 35, 36, 38, 45, 48]. While this is useful for many applications, it is the motion of the body in the world that tells us about human behavior. As noted by Johansson [28] even a

few moving point lights on a human body in motion informs us about behavior. Here we address how to exploit temporal information to more accurately estimate the 3D motion of the body from monocular video. While this problem has received over 30 years of study, we may ask why reliable methods are still not readily available. Our insight is that the previous temporal models of human motion have not captured the complexity and variability of real human motions due to insufficient training data. We address this problem here with a new temporal neural network and training approach, and show that it significantly improves 3D human pose estimation from monocular video.

Existing methods for video pose and shape estimation [30, 53] often fail to produce accurate predictions as illustrated in Fig. 1 (top). A major reason behind this is the lack of in-the-wild ground-truth 3D annotations, which are non-trivial to obtain even for single images. Previous work [30, 53] combines indoor 3D datasets with videos having 2D ground-truth or pseudo-ground-truth keypoint annotations. However, this has several limitations: (1) indoor 3D datasets are limited in the number of subjects, range of motions, and image complexity; (2) the amount of video labeled with ground-truth 2D pose is still insufficient to train deep networks; and (3) pseudo-ground-truth 2D labels are not reliable for modeling 3D human motion.

To address this, we take inspiration from Kanazawa *et al*. [29] who train a single-image pose estimator using only 2D keypoints and an *unpaired* dataset of *static* 3D human shapes and poses using an adversarial training approach. For video sequences, there already exist in-the-wild videos with 2D keypoint annotations. The question is then how to obtain realistic 3D human *motions* in sufficient quality for adversarial training. For that, we leverage the large-scale 3D motion-capture dataset called AMASS [41], which is sufficiently rich to learn a model of how people move. Our approach learns to estimate sequences of 3D body shapes poses from in-the-wild videos such that a discriminator cannot tell the difference between the estimated motions and motions in the AMASS dataset. As in [29], we also use 3D keypoints when available.

The output of our method is a sequence of pose and shape parameters in the SMPL body model format [40], which is consistent with AMASS and the recent literature. Our method learns about the richness of how people appear in images and is grounded by AMASS to produce valid human motions. Specifically, we leverage two sources of unpaired information by training a sequence-based generative adversarial network (GAN) [18]. Here, given the video of a person, we train a temporal model to predict the parameters of the SMPL body model for each frame while a motion discriminator tries to distinguish between real and regressed sequences. By doing so, the regressor is encouraged to output poses that represent plausible motions through minimizing an adversarial training loss while the discriminator acts as weak supervision. The motion discriminator implicitly learns to account for the statics, physics and kinematics of the human body in motion using the ground-truth motion-capture (mocap) data. We call our method **VIBE**, which stands for "**V**ideo **I**nference for **B**ody Pose and Shape **E**stimation."

During training, VIBE takes in-the-wild images as input and predicts SMPL body model parameters using a convolutional neural network (CNN) pretrained for single-image body pose and shape estimation [36] followed by a temporal encoder and body parameter regressor used in [29]. Then, a motion discriminator takes predicted poses along with the poses sampled from the AMASS dataset and outputs a real/fake label for each sequence. We implement both the temporal encoder and motion discriminator using Gated Recurrent Units (GRUs) [14] to capture the sequential nature of human motion. The motion discriminator employs a learned attention mechanism to amplify the contribution of distinctive frames. The whole model is supervised by an adversarial loss along with regression losses to minimize the error between predicted and ground-truth keypoints, pose, and shape parameters.

At test time, given a video, we use the pretrained CNN [36] and our temporal module to predict pose and shape parameters for each frame. The method works for video sequences of arbitrary length. We perform extensive experiments on multiple datasets and outperform all state-of-the-art methods; see Fig. 1 (bottom) for an example of VIBE's output. Importantly, we show that our video-based method always outperforms single-frame methods by a significant margin on the challenging 3D pose estimation benchmarks 3DPW [61] and MPI-INF-3DHP [42]. This clearly demonstrates the benefit of using video in 3D pose estimation.

In summary, the key contributions in this paper are: First, we leverage the AMASS dataset of motions for adversarial training of VIBE. This encourages the regressor to produce realistic and accurate motions. Second, we employ an attention mechanism in the motion discriminator to weight the contribution of different frames and show that this improves our results over baselines. Third, we quantitatively compare different temporal architectures for 3D human motion estimation. Fourth, we achieve state-of-the-art results on major 3D pose estimation benchmarks. Code and pretrained models are available for research purposes at https://github.com/mkocabas/VIBE.

2. Related Work

3D pose and shape from a single image. Parametric 3D human body models [4, 40, 47] are widely used as the output target for human pose estimation because they capture the statistics of human shape and provide a 3D mesh that can be used for many tasks. Early work explores "bot-

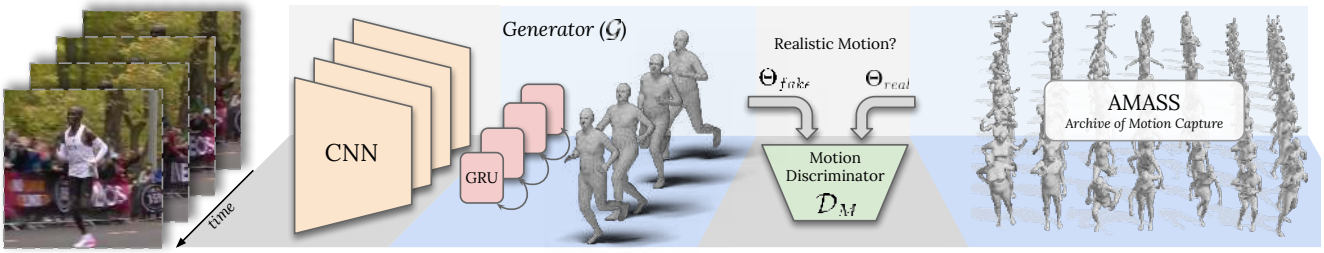

Figure 2: **VIBE architecture.** VIBE estimates SMPL body model parameters for each frame in a video sequence using a temporal generation network, which is trained together with a motion discriminator. The discriminator has access to a large corpus of human motions in SMPL format.

tom up" regression approaches, "top down" optimization approaches, and multi-camera settings using keypoints and silhouettes as input [1, 8, 19, 52]. These approaches are brittle, require manual intervention, or do not generalize well to images in the wild. Bogo *et al*. [11] propose SMPLify, one of the first end-to-end approaches, which fits the SMPL model to the output of a CNN keypoint detector [50]. Lassner *et al*. [38] use silhouettes along with keypoints during fitting. Recently, deep neural networks are trained to directly regress the parameters of the SMPL body model from pixels [21, 29, 45, 48, 55, 57]. Due to the lack of in-the-wild 3D ground-truth labels, these methods use weak supervision signals obtained from a 2D keypoint reprojection loss [29, 55, 57], use body/part segmentation as an intermediate representation [45, 48], or employ a human in the loop [38]. Kolotouros *et al*. [36] combine regression-based and optimization-based methods in a collaborative fashion by using SMPLify in the training loop. At each step of the training, the deep network [29] initializes the SMPLify optimization method that fits the body model to 2D joints, producing an improved fit that is used to supervise the network. Alternatively, several non-parametric body mesh reconstruction methods [37, 51, 59] has been proposed. Varol *et al*. [59] use voxels as the output body representation. Kolotouros *et al*. [37] directly regress vertex locations of a template body mesh using graph convolutional networks [33]. Saito *et al*. [51] predict body shapes using pixel-aligned implicit functions followed by a mesh reconstruction step. Despite capturing the human body from single images, when applied to video, these methods yield jittery, unstable results.

3D pose and shape from video. The capture of human motion from video has a long history. In early work, Hogg *et al*. [23] fit a simplified human body model to images features of a walking person. Early approaches also exploit methods like PCA and GPLVMs to learn motion priors from mocap data [46, 58] but these approaches were limited to simple motions. Many of the recent deep learning methods that estimate human pose from video [15, 24, 43, 49, 44] focus on joint locations only. Several methods [15, 24, 49] use a two-stage approach to "lift" off-the-shelf 2D keypoints into 3D joint locations. In contrast, Mehta *et al*. [43, 44] employ end-to-end methods to directly regress 3D joint locations. Despite impressive performance on indoor datasets like Human3.6M [26], they do not perform well on in-the-wild datasets like 3DPW [61] and MPI-INF-3DHP [42]. Several recent methods recover SMPL pose and shape parameters from video by extending SMPLify over time to compute a consistent body shape and smooth motions [6, 25]. Particularly, Arnab *et al*. [6] show that Internet videos annotated with their version of SMPLify help to improve HMR when used for fine tuning. Kanazawa *et al*. [30] learn human motion kinematics by predicting past and future frames[1]. They also show that Internet videos annotated using a 2D keypoint detector can mitigate the need for the in-the-wild 3D pose labels. Sun *et al*. [53] propose to use a transformer-based temporal model [60] to improve the performance further. They propose an unsupervised adversarial training strategy that learns to order shuffled frames.

GANs for sequence modeling. Generative adversarial networks GANs [5, 18, 27, 39] have had a significant impact on image modeling and synthesis. Recent works have incorporated GANs into recurrent architectures to model sequence-to-sequence tasks like machine translation [54, 62, 63]. Research in motion modelling has shown that combining sequential architectures and adversarial training can be used to predict future motion sequences based on previous ones [9, 20] or to generate human motion sequences [2]. In contrast, we focus on adversarially refining predicted poses conditioned on the sequential input data. Following that direction, we employ a motion discriminator that encodes pose and shape parameters in a latent space using a recurrent architecture and an adversarial objective taking advantage of 3D mocap data [41].

3. Approach

The overall framework of VIBE is summarized in Fig. 2. Given an input video $V = \{I_t\}_{t=1}^{T}$ of length T, of a sin-

[1]Note that they refer to kinematics over time as dynamics.

gle person, we extract the features of each frame I_t using a pretrained CNN. We train a temporal encoder composed of bidirectional Gated Recurrent Units (GRU) that outputs latent variables containing information incorporated from past and future frames. Then, these features are used to regress the parameters of the SMPL body model at each time instance.

SMPL represents the body pose and shape by Θ, which consists of the pose and shape parameters $\theta \in \mathbb{R}^{72}$ and $\beta \in \mathbb{R}^{10}$ respectively. The pose parameters include the global body rotation and the relative rotation of 23 joints in axis-angle format. The shape parameters are the first 10 coefficients of a PCA shape space; here we use the gender-neutral shape model as in previous work [29, 36] Given these parameters, the SMPL model is a differentiable function, $\mathcal{M}(\theta, \beta) \in \mathbb{R}^{6890\times 3}$, that outputs a posed 3D mesh.

Given a video sequence, VIBE computes $\hat{\mathbf{\Theta}} = [(\hat{\theta}_1, \cdots, \hat{\theta}_T), \hat{\beta}]$ where $\hat{\theta}_t$ are the pose parameters at time step t and $\hat{\beta}$ is the single body shape prediction for the sequence. Specifically, for each frame we predict the body shape parameters. Then, we apply average pooling to get a single shape ($\hat{\beta}$) across the whole input sequence. We refer to the model described so far as the temporal generator $\mathcal{G}$. Then, output, $\hat{\mathbf{\Theta}}$, from $\mathcal{G}$ and samples from AMASS, $\mathbf{\Theta}_{real}$, are given to a motion discriminator, $\mathcal{D}_M$, in order to differentiate fake and real examples.

3.1. Temporal Encoder

The intuition behind using a recurrent architecture is that future frames can benefit from past video pose information. This is useful when the pose of a person is ambiguous or the body is partially occluded in a given frame. Here, past information can help resolve and constrain the pose estimate. The temporal encoder acts as a generator that, given a sequence of frames $I_1, \ldots, I_T$, outputs the corresponding pose and shape parameters in each frame. A sequence of T frames is fed to a convolutional network, f, which functions as a feature extractor and outputs a vector $f_i \in \mathbb{R}^{2048}$ for each frame $f(I_1), \ldots, f(I_T)$. These are sent to a Gated Recurrent Unit (GRU) layer [14] that yields a latent feature vector g_i for each frame, $g(f_1), \ldots, g(f_T)$, based on the previous frames. Then, we use g_i as input to T regressors with iterative feedback as in [29]. The regressor is initialized with the mean pose $\bar{\Theta}$ and takes as input the current parameters Θ_k along with the features g_i in each iteration k. Following Kolotouros *et al.* [36], we use a 6D continuous rotation representation [65] instead of axis angles.

Overall, the loss of the proposed temporal encoder is composed of 2D (x), 3D (X), pose (θ) and shape (β) losses when they are available. This is combined with an adversarial $\mathcal{D}_M$ loss. Specifically the total loss of the $\mathcal{G}$ is:

$$L_{\mathcal{G}} = L_{3D} + L_{2D} + L_{SMPL} + L_{adv} \tag{1}$$

where each term is calculated as:

$$L_{3D} = \sum_{t=1}^{T} \|X_t - \hat{X}_t\|_2,$$

$$L_{2D} = \sum_{t=1}^{T} \|x_t - \hat{x}_t\|_2,$$

$$L_{SMPL} = \|\beta - \hat{\beta}\|_2 + \sum_{t=1}^{T} \|\theta_t - \hat{\theta}_t\|_2 \,,$$

where L_{adv} is the adversarial loss explained below.

To compute the 2D keypoint loss, we need the SMPL 3D joint locations $\hat{X}(\Theta) = W\mathcal{M}(\theta, \beta)$, which are computed from the body vertices with a pretrained linear regressor, W. We use a weak-perspective camera model with scale and translation parameters $[s, t], t \in \mathbb{R}^2$. With this we compute the 2D projection of the 3D joints $\hat{X}$, as $\hat{x} \in \mathbb{R}^{j\times 2} = s\Pi(R\hat{X}(\Theta)) + t$, where $R \in \mathbb{R}^3$ is the global rotation matrix and Π represents orthographic projection.

3.2. Motion Discriminator

The body discriminator and the reprojection loss used in [29] enforce the generator to produce *feasible* real world poses that are aligned with 2D joint locations. However, single-image constraints are not sufficient to account for sequences of poses. Multiple inaccurate poses may be recognized as valid when the temporal continuity of movement is ignored. To mitigate this, we employ a motion discriminator, $\mathcal{D}_M$, to tell whether the generated sequence of poses corresponds to a realistic sequence or not. The output, $\hat{\mathbf{\Theta}}$, of the generator is given as input to a multi-layer GRU model f_M depicted in Fig. 3, which estimates a latent code h_i at each time step i where $h_i = f_m(\hat{\Theta}_i)$. In order to aggregate hidden states $[h_i, \cdots, h_T]$ we use self attention [7] elaborated below. Finally, a linear layer predicts a value $\in [0, 1]$ representing the probability that $\hat{\mathbf{\Theta}}$ belongs to the manifold of plausible human motions. The adversarial loss term that is backpropagated to $\mathcal{G}$ is:

$$L_{adv} = \mathbb{E}_{\Theta\sim p_G}[(\mathcal{D}_M(\hat{\mathbf{\Theta}}) - 1)^2] \tag{2}$$

and the objective for $\mathcal{D}_M$ is:

$$L_{\mathcal{D}_M} = \mathbb{E}_{\Theta\sim p_R}[(\mathcal{D}_M(\mathbf{\Theta}) - 1)^2] + \mathbb{E}_{\Theta\sim p_G}[\mathcal{D}_M(\hat{\mathbf{\Theta}})^2] \tag{3}$$

where p_R is a real motion sequence from the AMASS dataset, while p_G is a generated motion sequence. Since $\mathcal{D}_M$ is trained on ground-truth poses, it also learns plausible body pose configurations, hence alleviating the need for a separate single-frame discriminator [29].

Motion Prior (MPoser). In addition to the $\mathcal{D}_M$, we experiment with a motion prior model, which we call MPoser.

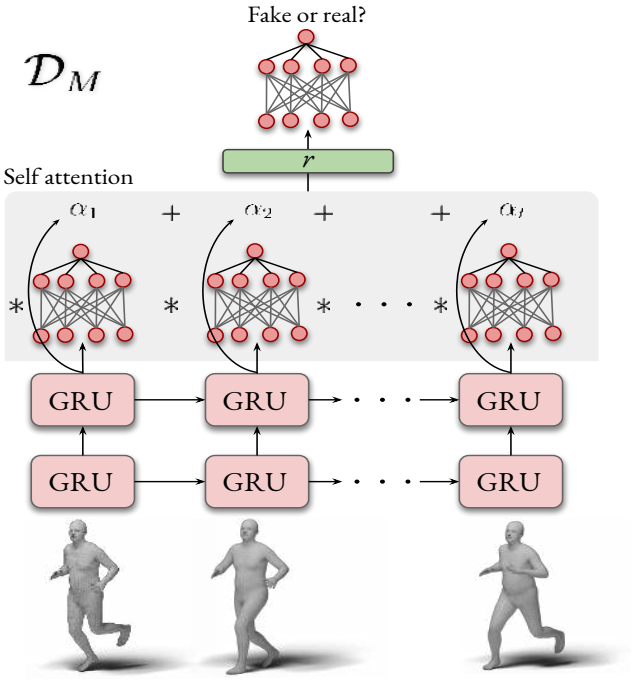

Figure 3: **Motion discriminator architecture** $\mathcal{D}_M$ consists of GRU layers followed by a self attention layer. $\mathcal{D}_M$ outputs a real/fake probability for each input sequence.

It is an extension of the variational body pose prior model VPoser [47] to temporal sequences. We train MPoser as a sequential VAE [32] on the AMASS dataset to learn a latent representation of plausible human motions. Then, we use MPoser as a regularizer to penalize implausible sequences. The MPoser encoder and decoder consist of GRU layers that output a latent vector $z_i \in \mathbb{R}^{32}$ for each time step i. When we employ MPoser, we disable $\mathcal{D}_M$ and add a prior loss $L_{MPoser} = \|z\|_2$ to $L_{\mathcal{G}}$.

Self-Attention Mechanism. Recurrent networks update their hidden states as they process input sequentially. As a result, the final hidden state holds a summary of the information in the sequence. We use a self-attention mechanism [7, 10] to amplify the contribution of the most important frames in the final representation instead of using either the final hidden state h_t or a hard-choice pooling of the hidden state feature space of the whole sequence. By employing an attention mechanism, the representation r of the input sequence $\hat{\mathbf{\Theta}}$ is a learned convex combination of the hidden states. The weights a_i are learned by a linear MLP layer ϕ, and are then normalized using softmax to form a probability distribution. Formally:

$$\phi_i = \phi(h_i), \quad a_i = \frac{e^{\phi_i}}{\sum_{t=1}^{N} e^{\phi_t}}, \quad r = \sum_{i=1}^{N} a_i h_i. \tag{4}$$

We compare our dynamic feature weighting with a static pooling schema. Specifically, the features h_i, representing the hidden state at each frame, are averaged and max pooled. Then, those two representations r_{avg} and r_{max} are concatenated to constitute the final static vector, r, used for the D_m fake/real decision.

3.3. Training Procedure

We use a ResNet-50 network [22] as an image encoder pretrained on single frame pose and shape estimation task [29, 36] that outputs $f_i \in \mathbb{R}^{2048}$. Similar to [30] we precompute each frame's f_i and do not update the ResNet-50. We use $T = 16$ as the sequence length with a mini-batch size of 32, which makes it possible to train our model on a single Nvidia RTX2080ti GPU. Although, we experimented with $T = [8, 16, 32, 64, 128]$, we chose $T = 16$, as it yields the best results. For the temporal encoder, we use a 2-layer GRU with a hidden size of 1024. The SMPL regressor has 2 fully-connected layers with 1024 neurons each, followed by a final layer that outputs $\hat{\mathbf{\Theta}} \in \mathbb{R}^{85}$, containing pose, shape, and camera parameters. The outputs of the generator are given as input to the $\mathcal{D}_M$ as fake samples along with the ground truth motion sequences as real samples. The motion discriminator architecture is identical to the temporal encoder. For self attention, we use 2 MLP layers with 1024 neurons each and $tanh$ activation to learn the attention weights. The final linear layer predicts a single fake/real probability for each sample. We also use the Adam optimizer [31] with a learning rate of 5×10^{-5} and 1×10^{-4} for the $\mathcal{G}$ and $\mathcal{D}_M$, respectively. Finally, each term in the loss function has different weighting coefficients. We refer the reader to Sup. Mat. for further details.

4. Experiments

We first describe the datasets used for training and evaluation. Next, we compare our results with previous frame-based and video-based state-of-the-art approaches. We also conduct ablation experiments to show the effect of our contributions. Finally, we present qualitative results in Fig. 4.

Training. Following previous work [29, 30, 36], we use batches of mixed 2D and 3D datasets. PennAction [64] and PoseTrack [3] are the only ground-truth 2D video datasets we use, while InstaVariety [30] and Kinetics-400 [13] are pseudo ground-truth datasets annotated using a 2D keypoint detector [12, 34]. For 3D annotations, we employ 3D joint labels from MPI-INF-3DHP [42] and Human3.6M [26]. When used, 3DPW and Human3.6M provide SMPL parameters that we use to calculate L_{SMPL}. AMASS [41] is used for adversarial training to obtain real samples of 3D human motion. We also use the 3DPW [61] training set to perform ablation experiments; this demonstrate the strength of our model on in-the-wild data.

	Models	3DPW				MPI-INF-3DHP			H36M	
		PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓	PA-MPJPE ↓	MPJPE ↓	PCK ↑	PA-MPJPE ↓	MPJPE ↓
Frame-based	Kanazawa *et al*. [29]	76.7	130.0	-	37.4	89.8	124.2	72.9	56.8	88
	Omran *et al*. [45]	-	-	-	-	-	-	-	59.9	-
	Pavlakos *et al*. [48]	-	-	-	-	-	-	-	75.9	-
	Kolotouros *et al*. [37]	70.2	-	-	-	-	-	-	50.1	-
	Arnab *et al*. [6]	72.2	-	-	-	-	-	-	54.3	77.8
	Kolotouros *et al*. [36]	59.2	96.9	116.4	29.8	67.5	105.2	76.4	**41.1**	-
Temporal	Kanazawa *et al*. [30]	72.6	116.5	139.3	**15.2**	-	-	-	56.9	-
	Doersch *et al*. [16]	74.7	-	-	-	-	-	-	-	-
	Sun *et al*. [53]	69.5	-	-	-	-	-	-	42.4	**59.1**
	VIBE (direct comp.)	56.5	93.5	113.4	27.1	**63.4**	97.7	**89.0**	41.5	65.9
	VIBE	**51.9**	**82.9**	**99.1**	23.4	64.6	**96.6**	89.3	41.4	65.6

Table 1: **Evaluation of state-of-the-art models on 3DPW, MPI-INF-3DHP, and Human3.6M datasets.** VIBE (direct comp.) is our proposed model trained on video datasets similar to [30, 53], while VIBE is trained with extra data from the 3DPW training set. VIBE outperforms all state-of-the-art models including SPIN [36] on the challenging in-the-wild datasets (3DPW and MPI-INF-3DHP) and obtains comparable result on Human3.6M. "−" shows the results that are not available.

Evaluation. For evaluation, we use 3DPW [61], MPI-INF-3DHP [42], and Human3.6M [26]. We report results with and without the 3DPW training to enable direct comparison with previous work that does not use 3DPW for training. We report Procrustes-aligned mean per joint position error (PA-MPJPE), mean per joint position error (MPJPE), Percentage of Correct Keypoints (PCK) and Per Vertex Error (PVE). We compare VIBE with state-of-the-art single-image and temporal methods. For 3DPW, we report acceleration error (mm/s^2), calculated as the difference in acceleration between the ground-truth and predicted 3D joints.

4.1. Comparison to state-of-the-art results

Table 1 compares VIBE with previous state-of-the-art frame-based and temporal methods. VIBE (direct comp.) corresponds to our model trained using the same datasets as Temporal-HMR [30], while VIBE also uses the 3DPW training set. As standard practice, previous methods do not use 3DPW, however we want to demonstrate that using 3DPW for training improves in-the-wild performance of our model. Our models in Table 1 use pretrained HMR from SPIN [36] as a feature extractor. We observe that our method improves the results of SPIN, which is the previous state-of-the-art. Furthermore, VIBE outperforms all previous frame-based and temporal methods on the challenging in-the-wild 3DPW and MPI-INF-3DHP datasets by a significant amount, while achieving results on-par with SPIN on Human3.6M. Note that, Human3.6M is an indoor dataset with a limited number of subjects and minimal background variation, while 3DPW and MPI-INF-3DHP contain challenging in-the-wild videos.

We observe significant improvements in the MPJPE and PVE metrics since our model encourages temporal pose and shape consistency. These results validate our hypothesis that the exploitation of human motion is important for improving pose and shape estimation from video. In addition to the reconstruction metrics, *e.g*. MPJPE, PA-MPJPE, we also report acceleration error (Table 1). While we achieve smoother results compared with the baseline frame-based methods [29, 36], Temporal-HMR [30] yields even smoother predictions. However, we note that Temporal-HMR applies aggressive smoothing that results in poor accuracy on videos with fast motion or extreme poses. There is a trade-off between accuracy and smoothness. We demonstrate this finding in a qualitative comparison between VIBE and Temporal-HMR in Fig. 5. This figure depicts how Temporal-HMR over-smooths the pose predictions while sacrificing accuracy. Visualizations from alternative viewpoints in Fig. 4 show that our model is able to recover the correct global body rotation, which is a significant problem for previous methods. This is further quantitatively demonstrated by the improvements in the MPJPE and PVE errors. For video results see the GitHub page.

4.2. Ablation Experiments

Table 2 shows the performance of models with and without the motion discriminator, $\mathcal{D}_M$. First, we use the original HMR model proposed by [29] as a feature extractor. Once we add our generator, $\mathcal{G}$, we obtain slightly worse but smoother results than the frame-based model due to lack of sufficient video training data. This effect has also been observed in the Temporal-HMR method [30]. Using $\mathcal{D}_M$ helps to improve the performance of $\mathcal{G}$ while yielding smoother predictions.

When we use the pretrained HMR from [36], we observe a similar boost when using $\mathcal{D}_M$ over using only $\mathcal{G}$. We also experimented with MPoser as a strong baseline against

	3DPW			
	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓
Kanazawa *et al*. [29]	73.6	120.1	142.7	34.3
Baseline (only $\mathcal{G}$)	75.8	126.1	147.5	28.3
$\mathcal{G}$ + $\mathcal{D}_M$	**72.4**	**116.7**	**132.4**	**27.8**
Kolotouros *et al*. [36]	60.1	102.4	129.2	29.2
Baseline (only $\mathcal{G}$)	56.9	90.2	109.5	28.0
$\mathcal{G}$ + MPoser Prior	54.1	87.0	103.9	28.2
$\mathcal{G}$ + $\mathcal{D}_M$ (VIBE)	**51.9**	**82.9**	**99.1**	**23.4**

Table 2: **Ablation experiments with motion discriminator $\mathcal{D}_M$.** We experiment with several models using HMR [29] and SPIN [36] as pretrained feature extractors and add our temporal generator $\mathcal{G}$ along with $\mathcal{D}_M$. $\mathcal{D}_M$ provides consistent improvements over all baselines.

$\mathcal{D}_M$. MPoser acts as a regularizer in the loss function to ensure valid pose sequence predictions. Even though, MPoser performs better than using only $\mathcal{G}$, it is worse than using $\mathcal{D}_M$. One intuitive explanation for this is that, even though AMASS is the largest mocap dataset available, it fails to cover all possible human motions occurring in in-the-wild videos. VAEs, due to over-regularization attributed to the KL divergence term [56], fail to capture real motions that are poorly represented in AMASS. In contrast, GANs do not suffer from this problem [17]. Note that, when trained on AMASS, MPoser gives 4.5mm PVE on a held out test set, while the frame-based VPoser gives 6.0mm PVE error; thus modeling motion matters. Overall, results shown in Table 1 demonstrate that introducing $\mathcal{D}_M$ improves performance in all cases. Although one may think that the motion discriminator might emphasize on motion smoothness over single pose correctness, our experiments with a pose only, motion only, and both modules revealed that the motion discriminator is capable of refining single poses while producing smooth motion.

Dynamic feature aggregation in $\mathcal{D}_M$ significantly improves the final results compared to static pooling ($\mathcal{D}_M$ - concat), as demonstrated in Table 3. The self-attention mechanism enables $\mathcal{D}_M$ to learn how the frames correlate temporally instead of hard-pooling their features. In most of the cases, the use of self attention yields better results. Even with an MLP hidden size of 512, adding one more layer outperforms static aggregation. The attention mechanism is able to produce better results because it can learn a better representation of the motion sequence by weighting features from each individual frame. In contrast, average and max pooling the features produces a rough representation of the sequence without considering each frame in detail. Self-attention involves learning a coefficient for each frame to re-weight its contribution in the final vector (r) producing a more fine-grained output. That validates our intuition that attention is helpful for modeling temporal dependencies in human motion sequences.

Model	PA-MPJPE ↓	MPJPE ↓
$\mathcal{D}_M$ - concat	53.7	85. 9
$\mathcal{D}_M$ - attention [2 layers,512 nodes]	54.2	86.6
$\mathcal{D}_M$ - attention [2 layers,1024 nodes]	**51.9**	**82.9**
$\mathcal{D}_M$ - attention [3 layers,512 nodes]	53.6	85.3
$\mathcal{D}_M$ - attention [3 layers,1024 nodes]	52.4	82.7

Table 3: **Ablation experiments on self-attention.** We experiment with several self-attention configurations and compare our method to a static pooling approach. We report results on the 3DPW dataset with different hidden sizes and numbers of layers of the MLP network.

5. Conclusion

While current 3D human pose methods work well, most are not trained to estimate human motion in video. Such motion is critical for understanding human behavior. Here we explore several novel methods to extend static methods to video: (1) we introduce a recurrent architecture that propagates information over time; (2) we introduce discriminative training of motion sequences using the AMASS dataset; (3) we introduce self-attention in the discriminator so that it learns to focus on the important temporal structure of human motion; (4) we also learn a new motion prior (MPoser) from AMASS and show it also helps training but is less powerful than the discriminator. We carefully evaluate our contributions in ablation studies and show how each choice contributes to our state-of-the-art performance on video benchmark datasets. This provides definitive evidence for the value of training from video.

Future work should explore using video for supervising single-frame methods by fine tuning the HMR features, examine whether dense motion cues (optical flow) could help, use motion to disambiguate the multi-person case, and exploit motion to track through occlusion. In addition, we aim to experiment with other attentional encoding techniques such as transformers to better estimate body kinematics.

Acknowledgements: We thank Joachim Tesch for helping with Blender rendering. We thank all Perceiving Systems department members for their feedback and the fruitful discussions. This research was partially supported by the Max Planck ETH Center for Learning Systems and the Max Planck Graduate Center for Computer and Information Science.

Disclosure: MJB has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI. MJB has financial interests in Amazon and Meshcapade GmbH.

Figure 4: **Qualitative results of VIBE on challenging in-the-wild sequences.** For each video, the top row shows some cropped images, the middle rows show the predicted body mesh from the camera view, and the bottom row shows the predicted mesh from an alternate view point.

Figure 5: **Qualitative comparison between VIBE (*top*) and Temporal-HMR [30] (*bottom*).** This challenging video contains fast motion, extreme poses, and self occlusion. VIBE produces more accurate poses than Temporal HMR.

References

[1] Ankur Agarwal and Bill Triggs. Recovering 3D human pose from monocular images. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2006. 3

[2] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3D human motion modelling. In *International Conference on Computer Vision*, 2019. 3

[3] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 5

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *SIGGRAPH*, 2005. 2

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017. 3

[6] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3, 6

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015. 4, 5

[8] Alexandru Balan and Michael J Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision*, 2008. 3

[9] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: Probabilistic 3D human motion prediction via GAN. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 3

[10] Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of The International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2018. 5

[11] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*, 2016. 1, 3

[12] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 5

[13] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[14] Kyunghyun Cho, Bart van Merrienboer, aglar Glehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014. 2, 4

[15] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, and Arjun Jain. Learning 3D human pose from structure and motion. *European Conference on Computer Vision*, 2018. 3

[16] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3D pose estimation: motion to the rescue. In *Advances in Neural Information Processing*, 2019. 6

[17] Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. In *International Conference on Learning Representations*, 2020. 7

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing*, 2014. 2, 3

[19] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3D structure with a statistical image-based shape model. In *International Conference on Computer Vision*, 2003. 3

[20] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and Jose M. F. Moura. Adversarial geometry-aware human motion prediction. In *European Conference on Computer Vision*, 2018. 3

[21] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 1, 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 2016. 5

[23] David Hogg. Model-based vision: A program to see a walking person. In *Image and Vision Computing*, 1983. 3

[24] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3D human pose estimation. In *European Conference on Computer Vision*, 2018. 3

[25] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *International Conference on 3DVision*, 2017. 1, 3

[26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2014. 3, 5, 6

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[28] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. In *Perception and Psychophysics*, 1973. 1

[29] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 4, 5, 6, 7

[30] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 3, 5, 6, 8

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014. 5

[32] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. 5

[33] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. 3

[34] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. MultiPoseNet: Fast multi-person pose estimation using pose residual network. In *European Conference on Computer Vision (ECCV)*, 2018. 5

[35] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3D human pose using multi-view geometry. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 1

[36] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision*, 2019. 1, 2, 3, 4, 5, 6, 7

[37] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3, 6

[38] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the People: Closing the loop between 3D and 2D human representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 3

[39] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, 2017. 3

[40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015. 2

[41] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, 2019. 2, 3, 5

[42] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *International Conference on 3DVision*, 2017. 2, 3, 5, 6

[43] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *International Conference on 3DVision*, 2018. 3

[44] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D human pose estimation with a single RGB camera. In *SIGGRAPH*, July 2017. 3

[45] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *International Conference on 3DVision*, 2018. 1, 3, 6

[46] Dirk Ormoneit, Hedvig Sidenbladh, Michael J. Black, and Trevor Hastie. Learning and tracking cyclic human motion. In *Advances in Neural Information Processing*. 2001. 3

[47] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 5

[48] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3, 6

[49] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[50] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[51] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. *International Conference on Computer Vision*, 2019. 3

[52] Leonid Sigal, Alexandru Balan, and Michael J Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in Neural Information Processing*, 2008. 3

[53] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *International Conference on Computer Vision*, 2019. 2, 3, 6

[54] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 2014. 3

[55] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3D human shape and pose prediction. In *British Machine Vision Conference*, 2017. 3

[56] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Scholkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018. 7

[57] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing*, 2017. 3

[58] Raquel Urtasun, David J. Fleet, and Pascal Fua. 3D people tracking with Gaussian process dynamical models. In

IEEE Conference on Computer Vision and Pattern Recognition, 2006. 3

[59] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision*, 2018. 3

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing*, 2017. 3

[61] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision*, 2018. 2, 3, 5, 6

[62] Lijun Wu, Yingce Xia, Li Zhao, Fei Tian, Tao Qin, Jianhuang Lai, , and Tie-Yan Liu. Adversarial neural machine translation. In *Proceedings of The Asian Conference on Machine Learning*, 2018. 3

[63] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Improving neural machine translation with conditional sequence generative adversarial nets. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2018. 3

[64] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *International Conference on Computer Vision*, 2013. 5

[65] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4