Behavior-Driven Synthesis of Human Dynamics

Andreas Blattmann* Timo Milbich* Michael Dorkenwald* Björn Ommer
Heidelberg Collaboratory for Image Processing, IWR, Heidelberg University, Germany

Abstract

Generating and representing human behavior are of major importance for various computer vision applications. Commonly, human video synthesis represents behavior as sequences of postures while directly predicting their likely progressions or merely changing the appearance of the depicted persons, thus not being able to exercise control over their actual behavior during the synthesis process. In contrast, controlled behavior synthesis and transfer across individuals requires a deep understanding of body dynamics and calls for a representation of behavior that is independent of appearance and also of specific postures. In this work, we present a model for human behavior synthesis which learns a dedicated representation of human dynamics independent of postures. Using this representation, we are able to change the behavior of a person depicted in an arbitrary posture, or to even directly transfer behavior observed in a given video sequence. To this end, we propose a conditional variational framework which explicitly disentangles posture from behavior. We demonstrate the effectiveness of our approach on this novel task, evaluating capturing, transferring, and sampling fine-grained, diverse behavior, both quantitatively and qualitatively.

1. Introduction

Understanding human appearance, posture and behavior are key problems of computer vision with numerous applications in autonomous driving [37, 39, 22], surveillance [12, 48, 58], medical treatment [6, 52] and beyond. While there has been major progress on representation [57, 46] and - with the advent of deep generative models [33, 23] - synthesis [7, 29] and manipulation [19, 13, 17] of *posture and appearance*, the understanding of representation and synthesis of *behavior* is an open problem.
Human motor behavior is defined by the distinct dynamics of our limbs and the entire body. Take for example a person raising their arm. This is fully determined by the upward movement of the arm. Since the remaining body posture is mostly unaffected, the behavior can be directly performed independently of a particular initial body configuration such as a sitting or standing posture (cf. Fig. 1). Moreover, rather complex behavior like running involves an interplay between certain body limbs, e.g. arms swinging synchronously with the movement of legs, and, thus, is naturally limited to certain postures to start with. To nevertheless enact such behavior from arbitrary starting poses, first a transition to fitting initial body configurations may be required - for instance, a sitting person needs to stand up before being able to walk. Finally, specific body features like size or build do not affect the ability to perform a walking behavior.
While behavior is eventually instantiated as a sequence of individual postures that can be observed in a video, this would be a suboptimal representation: We want the overall behavior to be the same, e.g. raising arm or walking, regardless of the initial posture it starts with. Although we are looking at different realizations it should still be represented as being the same behavior. Consequently, understanding, controlling, and synthesizing behavior calls for separate disentangled representations of the characteristic behavior and of individual (in particular the initial) posture. In contrast, present work on human motion synthesis typically represents behavior directly by means of the observed sequence of postures [3, 38, 61, 55]. Thus, as no explicit understanding and representation of behavior is developed, synthesizing human behavior has been limited to only changing person appearance [55, 9, 54] or forecasting the most likely continuation of the depicted posture sequence [3, 38, 61, 11]. However, controlling such sequences, e.g. to re-enact a novel behavior by an observed person, asks for a posture independent representation which captures only the behavior dynamics to be transferred. Moreover, instantiating the re-enacted behavior requires combining these dynamics with the, potentially significantly different, posture of the target person.
In this paper we propose a conditional variational generative model for controlled human behavior synthesis which only requires a collection of sequences without any class labels provided. Our models learns to understand the characteristic motor dynamics of behavior, which enables us to transfer behavior between videos. We learn a dedicated representation extraction these dynamics from pose sequences while factorizing out posture information. To this end, we propose

*Indicates equal contribution

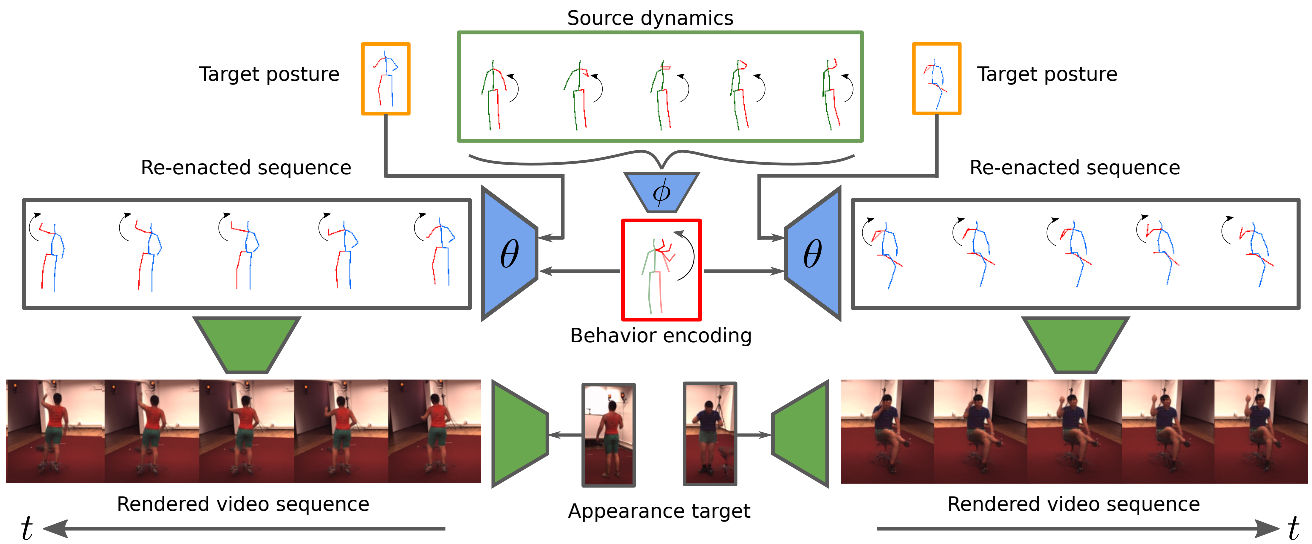


Figure 1. *Our Approach for Behavior Transfer*. Given a source sequence of human dynamics our model infers a behavior encoding which is independent of posture. We can re-enact the behavior by combining it with an unrelated target posture and thus control the synthesis process. The resulting sequence is combined with an appearance to synthesize a video sequence.

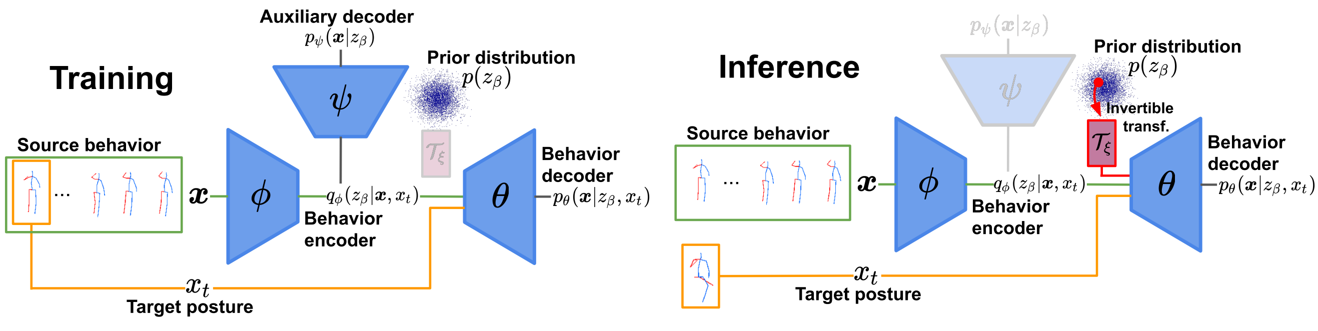
an explicit disentanglement framework for behavior and posture based on an alternating optimization procedure while simultaneously controlling the information flow through our model. In particular, the explicit disentanglement allows our model to re-enact extracted behavior from arbitrary target postures and, if needed, to infer required corresponding transitions itself. Our experiments demonstrate qualitatively and quantitatively that our model meaningfully transfers behavior between sequences and is also able to sample novel and diverse behavior. Quantitative comparison against current approaches for human motion synthesis confirms the competitive performance of our approach.

2. Related Work

Static person rendering. Much work has been proposed to alter certain characteristics of humans depicted in static images like age, gender or body features [30, 32, 34, 50] or synthesizing individual persons in different, unseen poses [19, 18, 36, 13]. The latter task typically requires for explicit disentanglement between certain factors of interest, often depending on paired image data [18, 36, 25]. While these approaches work well factors in static images, our work aims at transferring human behavior and, thus, requires disentanglement of a temporal factor, which is is significantly more complex.

Human video synthesis. Human video synthesis has been addressed in multiple ways. Some approaches synthesize videos directly in the pixel space [51, 2]. Due to the vast complexity of this problem, most approaches are based on mid-level representation of human shape, such as segmentation masks [55, 21] or pose estimates [54, 9, 60]. Chan et al. [9] generate video sequences of dancing persons by first learning correspondences between frames and postures before adding appearance information. A similar sequence-to-sequence translation task is performed in [55, 54, 35]. These works represent behavior directly on instantiated pose sequences, thus lacking the ability to exercise control. Our model understands and explicitly learns a behavior representation which can be used to transfer characteristic behavior dynamics between persons. Another line of research is future human motion prediction based on an initially observed posture sequence [14, 28, 20, 53]. Yuan et al. [62, 61] extend the future motion prediction task using multiple transformations on the latent space to increase the diversity of predicted motions. Chiu et al. [11] propose a hierarchical multi-scale RNN to learn dependencies between individual postures. Martinez et al. [38] use residual RNN architectures to directly model motion velocities. In contrast to our approach, these methods can not control the predicted behavior but only extrapolate the observed posture sequence.

Controlled behavior synthesis. Controlling the behavior to be generated requires a considerable higher degree of understanding than unconditioned prediction or sequence translation. Recent works control mostly only in form of a small, fixed set of predefined actions [21, 24]. Yang et al. [60] condition the synthesis process on action labels. In contrast, we require only a collection of unpaired video sequences and condition the synthesis process on a dedicated representation of behavior independent of posture. DLow [61] splits posture into different sets of keypoints to vary the diversity of predicted future movements for predefined body parts while keeping the others close to the groundtruth future sequence



(a) (b)

Figure 2. *Overview over model training (a) and inference (b).* Each distribution is realized by a deep neural network. During training, the first posture of $\boldsymbol{x}$ serves as conditioning x_t (yellow). Note, that consequently x_t is also part of the encoder input since we do not have multiple training sequences $\boldsymbol{x}$ starting from the same posture x_t available. In inference, i.e. after disentangling posture and behavior, we transfer source behavior (green) to an arbitrary target posture (yellow) or synthesize novel behavior from the prior distribution which is matched to $q_\phi(z_\beta|\boldsymbol{x}, x_t)$ by a learned invertible transformation $\mathcal{T}_\xi$ (red).

and, thus, cannot exercise detailed control.

3D action recognition Action recognition [8, 56] aims at classifying a predefined set of actions from a given video, potentially based on intermediate representations such as 3D keypoints. Although our learned behavior representation is also based on keypoints, we aim at capturing behavior dynamics for detailed synthesis of full behavior. In contrast, action recognition learns discriminative representations which only focus on separating between action classes [49].

3. Approach

Our goal is to control and synthesize videos of human behavior. Since powerful pose estimators [57, 46] are readily available, pose sequences $\boldsymbol{x} = [x_0, ..., x_n], x_i \in \mathbb{R}^{K\times 3}$ are directly used as a basis to represent the behavior observed in video [55, 54, 9], e.g. for changing the depicted person's appearance or predicting likely sequence continuations. While this representation is sufficient to perform the aforementioned tasks, changing the posture sequence to re-enact a different behavior asks for a deeper understanding. Thus, behavior transfer requires separate representations modelling the characteristic motor dynamics of behavior and individual postures. We now present a generative model which extracts and represents human behavior β from a source sequence $\boldsymbol{x}$ independent of the instantiated postures. Given an observed target posture x_t, e.g. in another video, we can then synthesize a re-enacted posture sequence and subsequently translate it to the video domain.

Extracting behavior β from $\boldsymbol{x}$ into a representation $z_\beta \in \mathbb{R}^D$ and subsequently re-combining it with a target pose x_t to instantiate the behavior can be naturally formulated by means of latent variable models such as encoder-decoder frameworks. Such frameworks have been successfully applied for predicting future postures based on $\boldsymbol{x}$, i.e. directly extrapolating the observed posture sequence [61, 53]. However, as we seek to control the behavior to be generated, we require the latent representation z_β to be disentangled from posture information.

Synthesis using conditional generative models Generative models are powerful frameworks which are particularly suited for synthesis tasks. As we not only aim to learn a representation for behavior, but also need to extract it from our input sequences $\boldsymbol{x}$, variational autoencoders (VAE) [33] are a natural choice. Such models approximate the true data distribution $p(\boldsymbol{x}, z_\beta)$ which is assumed to follow the generative process $p(\boldsymbol{x}, z_\beta) = p(\boldsymbol{x}|z_\beta)p(z_\beta)$. To optimize the intractable marginal log-likelihood $\mathbb{E}_{p(\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})]$ of the model distribution $p_\theta(\boldsymbol{x}, z_\beta)$, a variational posterior $q_\phi(z_\beta|\boldsymbol{x})$ is introduced allowing to maximize a lower bound $L(p_\theta, q_\phi) \leq \mathbb{E}_{p(\boldsymbol{x})}[\log p(\boldsymbol{x})]$ [33]. Now, since we want to transfer behavior and condition it on arbitrary target postures, we condition the generative process [47] additionally on x_t, which modifies the lower variational bound and its optimization to

$$\max_{\theta,\phi} \; L(p_\theta, q_\phi) := \mathbb{E}_{q_\phi(z_\beta|\boldsymbol{x}, x_t)}\left[\log p_\theta(\boldsymbol{x}|z_\beta, x_t)\right] - D_{\mathrm{KL}}(q_\phi(z_\beta|\boldsymbol{x}, x_t)||p(z_\beta)) \tag{1}$$

where $p(z_\beta)$ is the prior on the latent representation z_β which is typically modelled as a standard Gaussian distribution $\mathcal{N}(0, I)$. The first term of (1) can be considered to optimize the synthesis quality of the generator $p_\theta(\boldsymbol{x}|z_\beta, x_t)$ while the second part regularizes the encoder $q_\phi(z_\beta|\boldsymbol{x}, x_t)$ to match the Gaussian prior.

Although our generator p_θ has access to both z_β and the conditioning posture x_t, optimizing (1) will in general not encourage our model to learn a factorization of posture information and the behavior representation z_β. Moreover, we have no ground-truth provided for different behaviors starting from the same target posture x_t. Thus, we are only able to train our model by choosing x_t to be the first posture of $\boldsymbol{x}$, which aggravates the need for an explicit disentanglement during the optimization process.

Disentangling posture from behavior While explicit disentanglement between factors of variation has been studied in the domain of static images [40, 25, 36], disentangling complex temporal information, however, is significantly still lacking. Existing works for static images typically resort to supervision by exploiting pairs of data samples sharing one factor while differing in the remaining factors [40, 25], which allows for a natural disentanglement signal. Without having similar supervision available, we need to explicitly disentangle the posture information in $\boldsymbol{x}$ from our latent behavior representation z_β. To this end, we would ideally want to minimize the predictability of the individual postures in $\boldsymbol{x}$ given z_β. However, performing this operation directly on basis of our generator p_θ does not prevent the erasure of body dynamics as well. Instead, we can frame this task using an auxiliary generative model.
Let $\hat{p}_\psi(\boldsymbol{x}|z_\beta)$ be a second generative model aiming at generating $\boldsymbol{x}$ from our behavior representation z_β only, i.e. optimizing the log-likelihood,

$$\max_{\psi} \ \mathbb{E}_{q_\phi(z_\beta|\boldsymbol{x},x_t)} \left[\log \hat{p}_\psi(\boldsymbol{x}|z_\beta)\right] \ . \tag{2}$$

Solving this task requires $\hat{p}_\psi(\boldsymbol{x}|z_\beta)$ to represent posture information which it has to be able to extract from z_β. Exploiting this, we can formulate our disentanglement task as an alternating optimization between our behavior model, i.e. p_θ, q_ϕ, optimizing $L(p_\theta, q_\phi)$ and $\hat{p}_\psi(\boldsymbol{x}|z_\beta)$ optimizing (2), both depending on the posterior $q_\phi(z_\beta|\boldsymbol{x}, x_t)$.[1] To limit the predictability of $\hat{p}_\psi(\boldsymbol{x}|z_\beta)$, we extend (1) resulting in

$$\max_{\theta,\phi} \ L(p_\theta, q_\phi) - \mathbb{E}_{q_\phi(z_\beta|\boldsymbol{x},x_t)} \left[\log \hat{p}_\psi(\boldsymbol{x}|z_\beta)\right] \ . \tag{3}$$

This objective does not explicitly optimize parameters ψ, thus the predictability of $\hat{p}(\psi)$ can only be diminished by removing information about $\boldsymbol{x}$ from z_β. Further, note that p_θ has access to the conditional x_t providing posture information and consequently only requires q_ϕ to provide missing dynamics to generate $\boldsymbol{x}$. As a result, factoring out posture information from z_β is indeed the most viable solution. Moreover, since posture information is excluded from our representation z_β, $p_\theta(\boldsymbol{x}|z_\beta, x_t)$ is required to infer a meaningful continuation of x_t depicting behavior β.
Due to the additional constraint in (3), the already existing pressure to reduce the overall encoded information in z_β imposed by $D_{\mathrm{KL}}(q_\phi(z_\beta|\boldsymbol{x}, x_t)||p(z_\beta))$ is further amplified. This also increases the risk of posterior collapses when using recurrent decoders [5], thus strongly affecting the generative process. Next, we discuss how to alleviate this problem by relaxing the information bottleneck.

[1] However, note that (2) is not optimized over parameters ϕ and consequently does not affect $q_\phi(z_\beta|\boldsymbol{x}, x_t)$.

Relaxing the information bottleneck for improved synthesis The quality of synthesis depends on the expressiveness of $p_\theta(\boldsymbol{x}|z_\beta, x_t)$ which stands in contrast to the regularization of the variational posterior $q_\phi(z_\beta|\boldsymbol{x}, x_t)$ in vanilla variational autoencoding settings [10, 63]. This becomes evident as the regularization $D_{\mathrm{KL}}(q_\phi(z_\beta|\boldsymbol{x}, x_t)||p(z_\beta))$ minimizes an upper bound on the mutual information $I_{q_\phi}(\boldsymbol{x}; z_\beta)$ [45], thus reducing the information captured in z_β. Consequently, a typical solution is to explicitly maximize the mutual information [63, 44]. However, computing reliable estimates of $I_{q_\phi}(\boldsymbol{x}; z_\beta)$ is difficult for complex data [18, 4]. Instead, we resort to a relaxation of the regularization in the original variational problem by only optimizing $D_{\mathrm{KL}}(q_\phi(z_\beta|\boldsymbol{x}, x_t)||p(z_\beta))$ to maintain a certain information budget I_{KL}, i.e. optimizing

$$\begin{aligned} &\max_{\theta,\phi} \ \mathbb{E}_{q_\phi(z_\beta|\boldsymbol{x},x_t)} \left[\log p_\theta(\boldsymbol{x}|z_\beta, x_t)\right] \\ &\qquad \text{s.t.} \quad D_{\mathrm{KL}}(q_\phi(z_\beta|\boldsymbol{x}, x_t)||p(z_\beta)) \leq I_{\mathrm{KL}} \ . \end{aligned} \tag{4}$$

Similar to Peng et al. [43] who constrain discriminator networks, we can optimize (4) using dual gradient decent. Overall, we arrive at our final objective $L(p_\theta, q_\phi)$ by inserting the relaxation constraint into (3) and introducing a scalar coefficient γ_C and the Lagrange multiplier γ_{KL} (which is still optimized via dual gradient decent), i.e.

$$\begin{aligned} L(p_\theta, q_\phi) =& \mathbb{E}_{q_\phi(z_\beta|\boldsymbol{x},x_t)} \left[\log p_\theta(\boldsymbol{x}|z_\beta, x_t)\right] \\ &- \gamma_{\mathrm{KL}} \left(D_{\mathrm{KL}}(q_\phi(z_\beta|\boldsymbol{x}, x_t)||p(z_\beta)) - I_{\mathrm{KL}}\right) \\ &- \gamma_C \ \mathbb{E}_{q_\phi(z_\beta|\boldsymbol{x},x_t)} \left[\log \hat{p}_\psi(\boldsymbol{x}|z_\beta)\right] \ . \end{aligned} \tag{5}$$

Note, that without our explicit disentanglement, relaxing $D_{\mathrm{KL}}(q_\phi(z_\beta|\boldsymbol{x}, x_t)||p(z_\beta))$ would further encourage the entanglement of posture and behavior dynamics in z_β.
Relaxing the regularization $D_{\mathrm{KL}}(q_\phi(z_\beta|\boldsymbol{x}, x_t)||p(z_\beta))$ comes at the cost of a reduced overlap between the variational posterior q_ϕ and prior $p(z_\beta)$ impairing the sampling ability of our model. Next, we correct this missmatch by means of a subsequently learned normalizing flow transformation [32, 15].

Bridging the gap beetween prior and posterior We want to use our model not only to transfer behavior between videos, but also to synthesize novel behavior based on sampling z_β from the prior distribution. Thus, strong deviations of the posterior $q_\phi(z_\beta|\boldsymbol{x}, x_t)$ from $p(z_\beta)$ may reduce the syntheses results due to out-of-distribution samples. Normalizing flows [42, 32] offer an effective way to bridge this sampling gap by learning an explicit, invertible transformation from q_ϕ to $p(z_\beta)$. Such models learn flexible probability

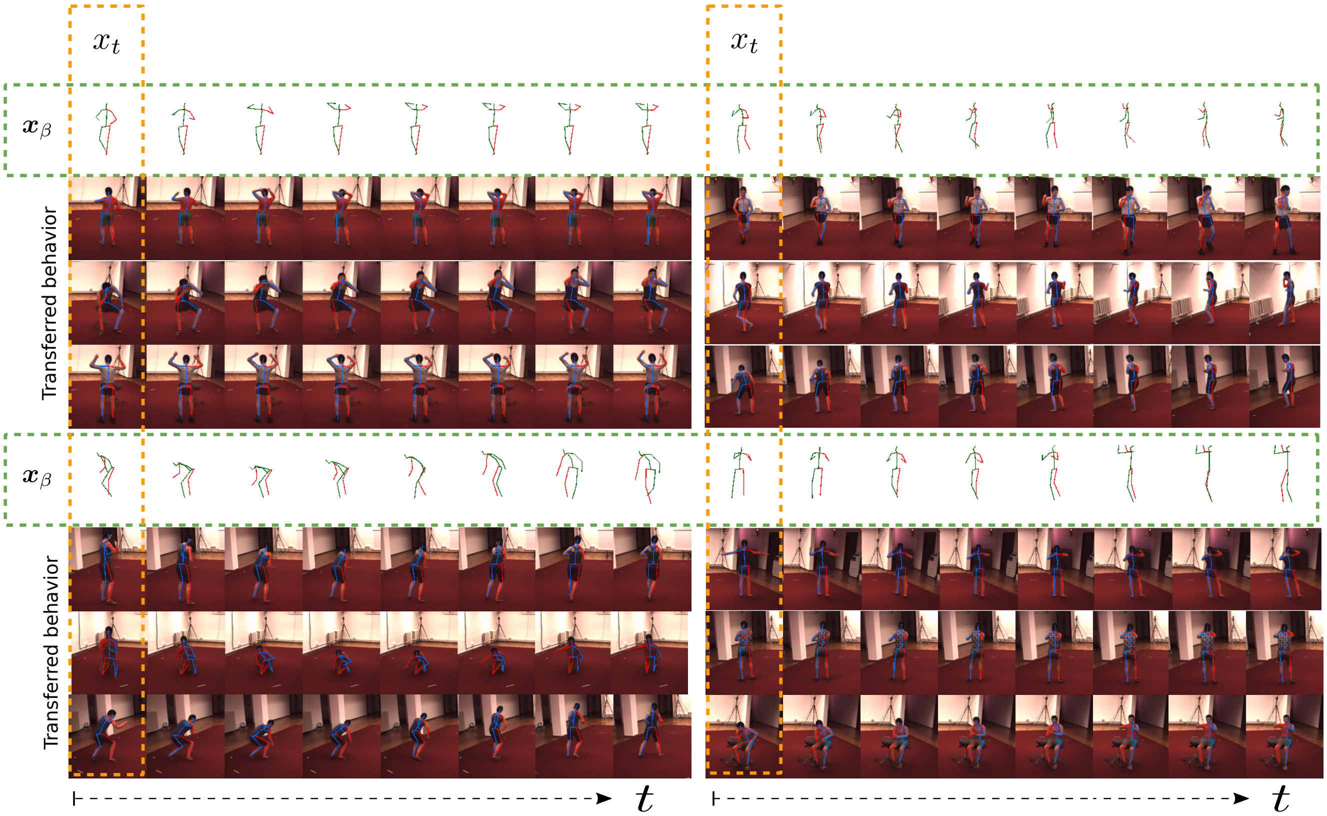


Figure 3. *Behavior Transfer on Human3.6m.* We transfer fine-grained, characteristic body dynamics of an observed behavior x_β to unrelated, significantly different target postures x_t. If required, the target posture is first adjusted by a transition phase before re-enacting the inferred behavior (e.g. top-right example, third row: walking starting from a bent down posture). Note that both transferred postures and images are generated by our models.

distributions $p_u(u)$ over continuous random variables, such as our behavior representation z_β. In particular, normalizing flows learn a bijective mapping $z_\beta \xleftrightarrow{\mathcal{T}_\xi} u$ using the transformation $\mathcal{T}_\xi = h_{\xi_1} \circ h_{\xi_2} \circ \cdots \circ h_{\xi_m}$, a sequence of m invertible functions h_{ξ_j} parametrized by ξ_j. by maximizing the likelihood

$$\mathbb{E}_{q_\phi(z_\beta|\boldsymbol{x},x_t)} \left[\log p_u(\mathcal{T}_\xi(z_\beta)) - \log |\det J_{\mathcal{T}_\xi}(z_\beta)| \right] \quad . \tag{6}$$

Here, $\det J_{\mathcal{T}_\xi}$ is the Jacobian determinant of the invertible transformation. Choosing $p_u(u)$ to follow the same distribution as $p(z_\beta)$ establishes our desired bijective mapping between $q_\phi(z_\beta|\boldsymbol{x}, x_t)$ and $p(z_\beta)$. Sampling novel behavior representations z_β is then performed by $z_\beta = \mathcal{T}_\xi^{-1}(u), u \sim p_u(u)$.

4. Experiments

We now investigate the capabilities of the proposed method to disentangle pose of a sequence from the underlying behavior. The resulting model is evaluated for the tasks of behavior transfer to different start poses and diverse sampling from the behavior representation. Evaluation is performed on the *Human3.6m* dataset [27], a large-scale motion capture dataset which contains 3.6 million video frames of 11 subjects, each of which performs 17 actions. Following previous work [62, 61] we use a 17-joint skeleton of 3D joint locations for training on 5 (S1,S5,S6,S7,S8) and testing on two subjects (S9,S11).

4.1. Architecture and implementation details:

For the task of human behavior transfer, we use sequences of 50 frames as input for our network. The encoder-decoder networks representing $q_\phi(z_\beta|\boldsymbol{x}, x_t)$ and $p_\theta(\boldsymbol{x}|z_\beta, x_t)$ are both implemented as a single-layer LSTM [26] with a hidden dimensionality of 1024. Mean and variance of $q_\phi(z_\beta|\boldsymbol{x}, x_t)$ are realized as linear layers based on the final hidden state of the encoder. For our decoder $p_\theta(\boldsymbol{x}|z_\beta, x_t)$ we initialize the hidden state with the behavior representation z_β. The target posture x_t is the input state of the decoder at the first time step. Subsequently the decoder uses its own predictions from the previous time step as input. For generating the individual postures, we follow [38] and use a single linear layer on top of the LSTM output combined with residual skip connection to the input. The generative model $\hat{p}_\psi$ is im-

Method	T=1		T=10		T=20		T=30		T=40		T=50		*acc.*	d_β
	RE	TDE	RE	TDE	RE	TDE	RE	TDE	RE	TDE	RE	TDE	gt: 0.45	$\mu \pm \sigma$
cAE	0.72	8.30	0.28	1.94	0.26	0.34	0.28	0.23	0.30	0.23	0.33	0.23	0.45	0.92±0.34
cVAE	5.29	9.07	5.28	8.95	5.05	8.81	4.82	8.87	4.55	8.86	4.46	8.80	0.13	0.00 ± 0.00
MT-VAE [59]	1.36	8.90	1.40	8.95	1.39	8.66	1.38	8.45	1.34	8.27	1.37	8.12	0.20	4.44 ±2.05
Ours ($\gamma_C = 0, I_{\text{KL}} = 50$)	1.71	9.01	1.46	7.92	1.22	6.95	1.17	6.15	1.18	5.58	1.30	5.33	0.35	2.82 ± 0.79
Ours ($\gamma_C = 0, I_{\text{KL}} = 100$)	1.24	8.99	0.89	7.09	0.81	5.55	0.78	4.33	0.73	3.48	0.80	3.13	0.39	3.47 ± 0.93
Ours ($\gamma_C = 0, I_{\text{KL}} = 200$)	1.01	8.92	0.67	5.93	0.61	3.74	0.59	2.29	0.58	1.48	0.60	1.30	0.40	4.06 ±1.18
Ours ($I_{\text{KL}} = 50$)	1.96	9.06	1.83	8.74	1.74	8.54	1.67	8.33	1.53	8.12	1.59	7.94	0.38	1.55 ± 0.61
Ours ($I_{\text{KL}} = 100$)	2.01	9.08	1.96	8.78	1.88	8.57	1.76	8.37	1.77	8.15	1.76	8.0	0.38	1.60 ±0.78
Ours ($I_{\text{KL}} = 200$)	1.62	9.06	1.47	8.97	1.56	8.90	1.47	8.77	1.47	8.58	1.38	8.36	0.39	1.58 ±0.71

Table 1. *Evaluation of Behavior Transfer*. We compare different models on the task of behavior transfer using different metrics. For the classifier accuracy, 'gt:0.45' denotes the accuracy on the validation set. For d_β we report mean and standard deviation.

Method	N=10		N=50	
	ASD	FSD	ASD	FSD
cVAE [62]	0.25	0.36	0.16	0.22
DSF [62]	0.38	0.62	0.31	0.42
Ours	**0.63**	**0.88**	**0.45**	**0.58**

(a)

Method	Classifier			
	self	transfer	prior	flow
MT-VAE [59]	**0.49**	0.17	?	?
Ours ($\gamma_C = 0$)	**0.49**	0.14	0.09	0.12
Ours	**0.49**	**0.23**	**0.13**	**0.23**

(b)

Figure 4. *Evaluation of Sampling Capabilities.* (a) Quantitative evaluation of diversity with ASD and FSD [62], numbers are taken from [62] (b) Evaluation of sampling quality with a classifier.

plemented as a three-layer MLP to predict postures $\boldsymbol{x}$ given z_β. We model $p_\theta(\boldsymbol{x}|z_\beta, x_t)$ and $p_\psi(\boldsymbol{x}|z_\beta)$ as Gaussian, thus the expectations in Eq. 5 translate to mean squared errors. We train the network for 50 epochs and set $\gamma_C = 0.1$ and $I_{\text{KL}} = 100$ as discussed in the quantitative evaluation.

Normalizing flow model $\mathcal{T}_\xi$: Our normalizing flow model $\mathcal{T}_\xi$ is implemented as a stacked sequence of 15 invertible neural networks based on an input dimensionality of $D = 1024$. Each consists of 3 blocks of subsequently applied actnorm [32], affine coupling layers [16] and shuffling layers. The affine coupling layers consist of 2 fully connected layers with dimensionality $D = 1024$. We trained the normalizing flow model on a single Titan Xp for 5 epochs with batchsize 64 and ADAM [31] optimizer with learning rate 6.5×10^{-6}.

Model for posture-appearance transfer: In order to be able to synthesize realistic RGB videos of human behavior, we translate our generated postures to RGB images. To this end, we utilize our proposed framework for the task of shape and appearance disentanglement [13]. We train a model to obtain an appearance representation from static images, which is independent of the corresponding posture information. Thus, we can use our method to transfer behavior from a source sequence to a given target posture and generate an animated video sequence by frame-wise synthesizing RGB images. More details on our posture and appearance model and further results can be found in the supplementary material.

4.2. Behavior re-enactment

We now evaluate our proposed model qualitatively and quantitatively for the task of behavior transfer and its abilities to sample and synthesize novel behavior.

Qualitative Evaluation: Figure 3 shows examples of transferred behavior. We show the posture sequence x_β exhibiting a source behavior β (top row) and its transfer to different, unrelated postures x_t. The re-enactments depict both the re-enacted posture sequence and the rendered RGB video frames based on the model for posture and appearance transfer. Since our model captures behavior independent of posture, it successfully transfers only the characteristic body dynamics of β and infers potentially needed transitions itself. As a result, the target posture x_t is naturally animated to perform behavior β independent of diverse target postures, such as standing or sitting. For instance, in the example at the left top, each person accurately raises both hands to its head. Note that, in the last example on the top left, the person does not change its posture since the hands are already up. Moreover, the kneeling person on the bottom left only lowers its torso as its knees are already placed on the ground and cannot be bent further. More examples and also videos can be found in the supplementary material.

Quantitative evaluation: Add descritpion about MT-VAE We now evaluate how well our model transfers behavior β extracted from a source sequence x_β to an initial, unrelated target posture x_t taken from random, different sequences.

Meaningful re-enactment of β should only transfer characteristic body dynamics to a target posture x_t. We compare our model to different baseline models, i.e. conditional autoencoder (cAE), vanilla conditional variational autoencoder (cVAE) and our model with and without our proposed posture disentanglement. Each model uses the same architectures except for deviations due to individual training objectives: The cAE is trained without disentanglement and without variational bottleneck. The cVAE is trained on the training objective Eq. (1). Note, due to lacking previous work on behavior transfer, there is no published competitor to compare to.

A model failing this task would typically generate *(i)* sequences which rather exactly copy full postures of the source sequence $\boldsymbol{x}_\beta$ in contrast to transferring only its characteristic dynamics to x_t; or *(ii)* generating behavior different to β such as some likely future behavior of x_t. To identify *(i)*, we measure the transfer displacement error (TDE), i.e. the displacement error between postures of the re-enactment $\boldsymbol{x}_R$ and source $\boldsymbol{x}_\beta$ at time-steps T. For *(ii)*, since we have no ground-truth available for behavior transfer, we measure the average euclidean distance d_β in the representation space z_β between encodings of $\boldsymbol{x}_\beta$ and $\boldsymbol{x}_R$. Combined, a well transferring model should yield re-enacted sequences $\boldsymbol{x}_R$ which is dissimilar to $\boldsymbol{x}_\beta$, thus not merely copying postures (i.e. large TDE). Given this, both sequences should be similar in representation space (i.e. small d_β) to indicate their similarity in behavior. Moreover, the representations need to be informative to exclude degenerated solutions. For the latter we examine their benefit for action classification on H3.6M (acc.). For each experiment we provide detailed protocols and implementation details in the supplementary.

Tab. 1 evaluates these experiments. We observe that cAE exhibits a strong decline in TDE values as T increases, resulting in TDE values close to 0. Thus, this model accurately copies the posture sequence x_β, instead of inferring behavior β and potentially needed transitions itself (cf. suppl. video material). Consequently, its behavior representation only captures posture information, rather than body dynamics. In contrast, the cVAE model consistently reaches high TDE scores, thus generating posture sequences which are very different from x_β. However, the distance d_β shows that the model suffers from posterior collapse, hence z_β is neglected and only likely continuations following x_t are predicted (see also supplementary material). Relaxing the information bottleneck of cVAE (i.e. our model without disentanglement, Eq.(4)) alleviates the posterior collapse and z_β becomes informative. Looking at different settings for I_{KL}, we see that TDE values slowly decrease with T and ranges between cAE and cVAE, while exhibiting large values of d_β. We attribute this to a distorted latent representation being learned. In contrast, our full model with explicit disentanglement of behavior and posture exhibits large TDE values matching those of the cVAE while at the same time mapping $\boldsymbol{x}_\beta$ and $\boldsymbol{x}_R$ close in z_β. Thus, since postures are very different, closeness in z_β arises from similarity in body dynamics, highlighting actual behavior transfer. The accuracy of the action classifier (*acc.*) confirms that the captured dynamics are informative, almost matching the classifier result when training on ground truth sequences.

To provide an additional, explicit measure for disentanglement of behavior and posture, we adapt an evaluation procedure inspired by works on identifying latent factors of variation [40]. To this end, we train a regression network to predict posture coordinates of x_β from its encoding z_β at different time-steps and report the average regression errors (RE) in Tab. 1. Naturally, the cAE model results in very low errors due to copying, while the cVAE exhibits large REs due to the posterior collapse. Comparing our model with and without disentanglement demonstrates consistently higher prediction errors, indicating that only few posture information is encoded in z_β. Moreover, our analysis shows that our model is robust to the choice of I_{KL}. In the remainder of the experiments we choose $I_{\text{KL}} = 100$.

4.3. Sampling and synthesis of novel behavior

We now evaluate our model on the task of synthesizing novel behavior by sampling behavior representations z_β from the prior distribution $p(z_\beta)$. Following other approaches for human motion synthesis [61, 62, 1] we evaluate the aspect of sampling quality [1] and diversity [61, 62]. To address the first we train a binary classifier to distinguish between 25k

	Pose-Knows	HP-GAN	GMVAE	DSF	DLOW	Ours
APD	6.72	7.24	6.77	9.33	11.74	**12.24**

Table 2. *Evaluation of Sampling Diversity*. Our model outperforms other approaches on human motion synthesis in terms of APD [61]. Numbers are taken from [61].

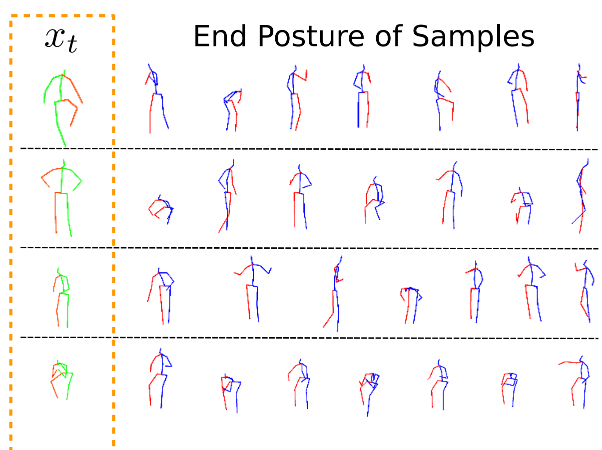
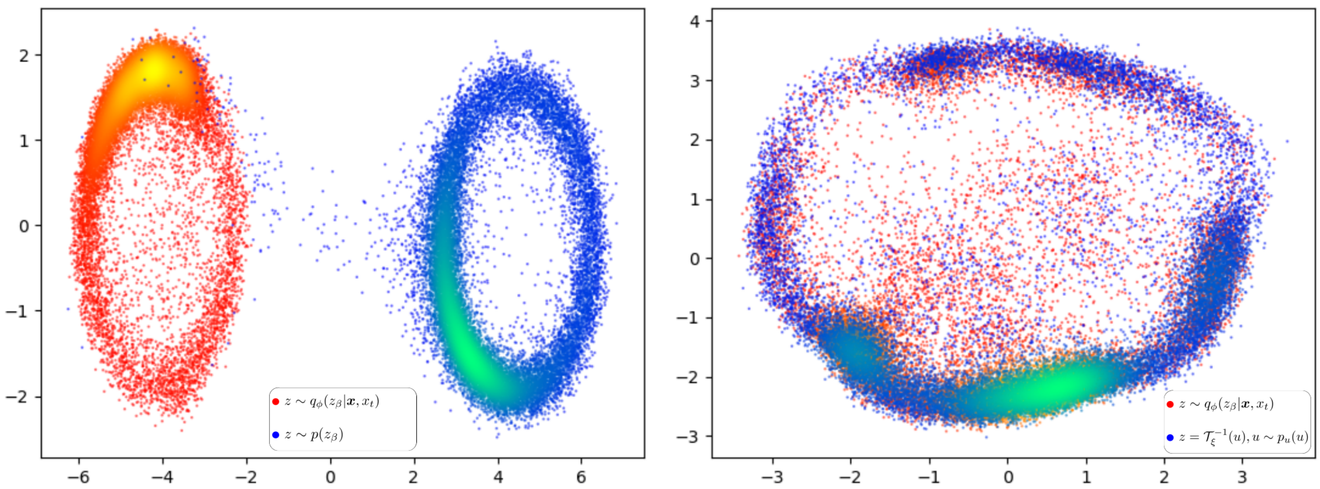


Figure 5. Qualitative visualization of diversity by showing the end poses from our sampled behaviors.



(a) (b)

Figure 6. 2D-UMAP of prior and posterior samples without (a) and with (b) learning the normalizing flow transformation $\mathcal{T}_\xi$ for distribution alignment.

ground-truth and 25k generated sequences. The accuracy of the classifier determines the realism of the evaluated samples and is reported in Fig. 4 (b). The implementation details for the classifier can be found in the supplementary. Posture sequences synthesized using the explicit invertible mapping $\mathcal{T}_\xi$ between prior and posterior $q_\phi(z_\beta|\boldsymbol{x}, x_t)$ are more realistic than directly using prior samples. This is explained by the corrected mismatch between posterior and prior distribution and clearly demonstrated by the visual comparisons in the supplementary material. Moreover, we observe that our explicit disentanglement of posture and behavior significant improves the quality of samples.

For evaluating diversity we follow the evaluation protocol of [61, 62] by using the following metrics: *(i)* Average Pairwise Distance (APD): Average euclidean distance between all pairwise combinations of generated sequences; *(ii)* Average Self Distance (ASD): Average euclidean distance between a generated sequence and its closest neighbor sequence among generations; and *(iii)* Final Self Distance (FSD): Euclidean distance between the last posture of a generated sequence and its closest neighbor's final posture. Note, while APD is measuring the overall variance of the generated sequences, ASD and FSD assess the uniqueness of samples. Fig. 4 (a) compares ASD and FSD scores of our model with the cVAE and the diversity sampler function (DSF) from [62] for sample-set sizes of $K \in \{10, 50\}$ while Tab. 2 we provide APD comparisons with various motion synthesis approaches. For each metric we outperform existing approaches by a significant margin, in particular such approaches [62, 61] which explicitly aim at sampling diversity. Finally, we visually demonstrate the diversity of our samples in Fig 5 by showing the final postures of sampled behaviors.

4.4. Invertible Transformation $\mathcal{T}_\xi$

To highlight the need for learning an explicit mapping between the prior $p(z_\beta)$ and the posterior $q(z_\beta|\boldsymbol{x}, x_t)$, we plot in Fig. 6 2D UMAP [41] visualizations of samples drawn from these distributions without and with using $\mathcal{T}_\xi$. Fig. 6 (a) shows a clear mismatch between both distributions. Fig. 6 (b) demonstrates that applying the transformation $\mathcal{T}_\xi$ helps to align prior and posterior, which is also reflected by the results discussed in the paragraph 'Behavior Sampling'.

5. Discussion

We presented a conditional generative model for controlled synthesis and transfer of human behavior. To this end, we learn a dedicated representation for human behavior disentangled from posture. By extracting the characteristic body dynamics from a video depicting a certain behavior, our model is able to animate persons observed in significantly different postures. A particular challenge arises from animating postures which allow for no direct transfer of behavior dynamics, but require an intermediate transition. Correct inference of such transition is essentially a generalization problem asking for synthesis outside the training distribution. While our model successfully infers such transitions to a certain degree, it fails in cases of complex transitions needed, such as enacting a walking behavior by a person which is lying on the ground. This shows that our introduced problem requires a deep understanding of behavior, thus posing a new challenge for research on human motion synthesis in general.

References

[1] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, Stephen Gould, and Amirhossein Habibian. Learning variations in human motion via mix-and-match perturbation, 2019. 7

[2] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *ECCV*, 2018. 2

[3] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. 1

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. 4

[5] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, Aug. 2016. 4

[6] Biagio Brattoli, Uta Büchler, Michael Dorkenwald, Philipp Reiser, Linard Filli, Fritjof Helmchen, Anna-Sophia Wahl, and Björn Ommer. ubam: Unsupervised behavior analysis and magnification using deep learning. *CoRR*, abs/2012.09237, 2020. 1

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019. 1

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018. 3

[9] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3

[10] Xi Chen, P. Diederik Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *ICLR*, 2017. 4

[11] Hsu-Kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018. 1, 2

[12] Ankur Datta, Mubarak Shah, and N Da Vitoria Lobo. Person-on-person violence detection in video data. In *Object recognition supported by user interaction for service robots*. IEEE, 2002. 1

[13] Rodrigo de Bem, Arnab Ghosh, Thalaiyasingam Ajanthan, Ondrej Miksik, N. Siddharth, and Philip H. S. Torr. Dgpose: Disentangled semi-supervised deep generative models for human body analysis. In *abs/1804.06364*, 2018. 1, 2, 6

[14] Emily L Denton and vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, 2017. 2

[15] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *international conference on learning representations*, 2017. 4

[16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. 2017. 6

[17] Michael Dorkenwald, Uta Büchler, and Björn Ommer. Unsupervised magnification of posture deviations across subjects. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8253–8263. IEEE, 2020. 1

[18] Patrick Esser, Johannes Haux, and Björn Ommer. Unsupervised robust disentangling of latent characteristics for image synthesis. In *Proceedings of the Intl. Conf. on Computer Vision (ICCV)*, 2019. 2, 4

[19] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[20] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. *International Conference on Computer Vision (ICCV)*, 2015. 2

[21] Oran Gafni, Lior Wolf, and Yaniv Taigman. Vid2game: Controllable characters extracted from real-world videos. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[22] Patrick Gebert, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. End-to-end prediction of driver intention using 3d convolutional neural networks. In *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019. 1

[23] 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 Systems 27*, 2014. 1

[24] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[25] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and V. S. R. Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *Proceedings IEEE European Conference on Computer Vision*, 2018. 2, 4

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

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

[28] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[29] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[32] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31*, 2018. 2, 4, 6

[33] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014. 1, 3

[34] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 2016. 2

[35] Wen Liu, Zhixin Piao, Min Jie, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[36] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, 2017. 2, 4

[37] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[38] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, Piscataway, NJ, USA, July 2017. IEEE. 1, 2, 5

[39] Sarfaraz Masood, Abhinav Rai, Aakash Aggarwal, Mohammad Najmud Doja, and Musheer Ahmad. Detecting distraction of drivers using convolutional neural network. *Pattern Recognition Letters*, 2018. 1

[40] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, 2016. 4, 7

[41] Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. 2018. 8

[42] George Papamakarios, Eric T. Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference, 2019. 4

[43] Bin Xue Peng, Angjoo Kanazawa, Samuel Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *international conference on learning representations*, 2019. 4

[44] Mary Phuong, Max Welling, Nate Kushman, Ryota Tomioka, and Sebastian Nowozin. The mutual autoencoder: Controlling information in latent code representations, 2018. 4

[45] Ali Rezaabad and Sriram Vishwanath. Learning representations by maximizing mutual information in variational autoencoder, 2019. 4

[46] Iasonas Kokkinos Riza Alp Gueler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3

[47] Kihyuk Sohn, Xinchen Yan, and Honglak Lee. Learning structured output representation using deep conditional generative models. In *International Conference on Neural Information Processing Systems*, 2015. 3

[48] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[49] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle, 2015. 3

[50] P. Upchurch, J. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Weinberger. Deep feature interpolation for image content changes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[51] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, 2016. 2

[52] Anna-Sophia Wahl, U Büchler, A Brändli, Biagio Brattoli, Simon Musall, Hansjörg Kasper, Benjamin V Ineichen, Fritjof Helmchen, Björn Ommer, and Martin E Schwab. Optogenetically stimulating intact rat corticospinal tract post-stroke restores motor control through regionalized functional circuit formation. *Nature communications*, 2017. 1

[53] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *International Conference on Computer Vision*, 2017. 2, 3

[54] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2, 3

[55] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2, 3

[56] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Compressed video action recognition, 2018. 3

[57] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018. 1, 3

[58] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for

anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015. 1

[59] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. MT-VAE: learning motion transformations to generate multimodal human dynamics. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 276–293. Springer, 2018. 6

[60] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *ECCV*, 2018. 2

[61] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction, 2020. 1, 2, 3, 5, 7, 8

[62] Ye Yuan and Kris M. Kitani. Diverse trajectory forecasting with determinantal point processes. In *8th International Conference on Learning Representations, ICLR 2020*, 2020. 2, 5, 6, 7, 8

[63] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders, 2017. 4