Understanding Object Dynamics for Interactive Image-to-Video Synthesis

Andreas Blattmann Timo Milbich Michael Dorkenwald Björn Ommer
Interdisciplinary Center for Scientific Computing, HCI
Heidelberg University, Germany

Abstract

What would be the effect of locally poking a static scene? We present an approach that learns naturally-looking global articulations caused by a local manipulation at a pixel level. Training requires only videos of moving objects but no information of the underlying manipulation of the physical scene. Our generative model learns to infer natural object dynamics as a response to user interaction and learns about the interrelations between different object body regions. Given a static image of an object and a local poking of a pixel, the approach then predicts how the object would deform over time. In contrast to existing work on video prediction, we do not synthesize arbitrary realistic videos but enable local interactive control of the deformation. Our model is not restricted to particular object categories and can transfer dynamics onto novel unseen object instances. Extensive experiments on diverse objects demonstrate the effectiveness of our approach compared to common video prediction frameworks. Project page is available at https://bit.ly/3cxfA2L.

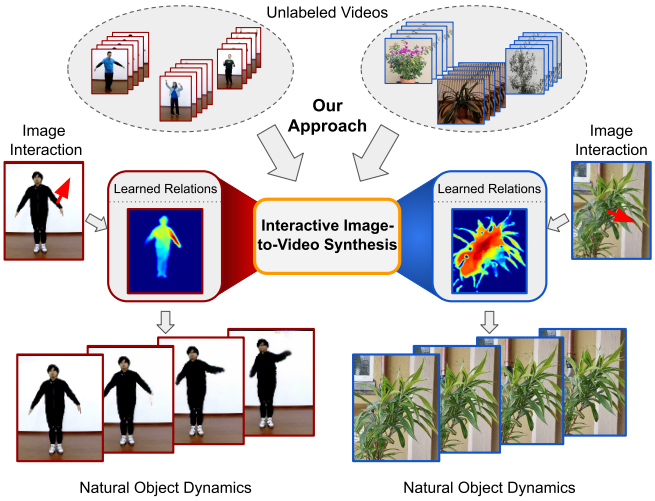

Figure 1. Our approach for interactive image-to-video synthesis learns to understand the relations between the distinct body parts of articulated objects from unlabeled video data, thus enabling synthesis of videos showing natural object dynamics as responses to local interactions.

1. Introduction

From infancy on we learn about the world by manipulating our immediate environment and subsequently observing the resulting diverse reactions to our interaction. Particularly in early years, poking, pulling, and pushing the objects around us is our main source for learning about their integral parts, their interplay, articulation and dynamics. Consider children playing with a plant. They eventually comprehend how subtle touches only affect individual leaves, while increasingly forceful interactions may affect larger and larger constituents, thus finally learning about the entirety of the dynamics related to various kinds of interactions. Moreover, they learn to generalize these dynamics across similar objects, thus becoming able to predict the reaction of a novel object to their manipulation.

Training artificial visual systems to gain a similar understanding of the distinct characteristics of object articulation and its distinct dynamics is a major line of computer vision research. In the realm of still images the interplay between object shape and appearance has been extensively studied, even allowing for controlled, global [43, 41, 21, 16, 15] and local [39, 73] manipulation. Existing work on object dynamics, however, so far is addressed by either extrapolations of observed object motion [67, 27, 11, 33, 67, 45, 5, 61] or only coarse control of predefined attributes such as explicit action labels [71] and imitation of previously observed holistic motion [1, 14]. Directly controlling and, even further, interacting with objects on a local level however, so far, is a novel enterprise. Teaching visual systems to understand the complex dynamics of objects both arising by explicit manipulations of individual parts and to predict and analyze the behavior [12, 3, 4] of the remainder of the object is an exceptionally challenging task. Similarly to a child in the example above, such systems need to know about the natural interrelations of different parts of an object [72]. Moreover, they have to learn how these parts are related by their dynamics to plausibly synthesize temporal

object articulation as a response to our interactions.

In this paper, we present a generative model for *interactive image-to-video synthesis* which learns such a fine-grained understanding of object dynamics and, thus, is able to synthesize video sequences that exhibit natural responses to local user interactions with images on *pixel-level*. Using intuitions from physics [56], we derive a hierarchical recurrent model dedicated to model complex, fine-grained object dynamics. Without making assumptions on objects we learn to interact with and no ground-truth interactions provided, we learn our model from video sequences only.

We evaluate our model on four video datasets comprising the highly-articulated object categories of humans and plants. Our experiments demonstrate the capabilities of our proposed approach to allow for fine-grained user interaction. Further, we prove the plausibility of our generated object dynamics by comparison to state-of-the-art video prediction methods in terms of visual and temporal quality. Figure 1 shows an overview over the capabilities of our model.

2. Related Work

Video Synthesis. Video Synthesis involves a wide range of tasks including video-to-video translation [65], image animation [68, 53, 54], frame interpolation [46, 70, 37, 2, 47], unconditional video generation [63, 57, 9, 44] and video prediction. Given a set of context frames, video prediction methods aim to predict realistic future frames either deterministically [67, 62, 60, 69] or stochastically [17, 11, 62, 33, 67, 51, 45, 5, 61, 32]. A substantial number of methods ground on image warping techniques [64, 38, 19], leading to high-quality short term predictions, while struggling with longer sequence lengths. To avoid such issues, many works use autoregressive models. Due to consistently increasing compute capabilities recent methods aim to achieve this task via directly maximizing likelihood in pixel space, using large scale architectures such as normalizing flows [31, 32] or pixel-level transformers [59, 66]. As such methods introduce excessive computational costs and are slow during inference, most existing methods rely on RNN-based methods, acting autoregressively in the pixel-space [42, 33, 40, 11, 62, 5] or on intermediate representation such as optical flow [35, 34, 49]. However, since they have no means for direct interactions with a depicted object, but instead rely on observing past frames, these methods model dynamics in a holistic manner. By modelling dynamics entirely in the latent space, more recent methods take a step towards a deeper understanding of dynamics [45, 18] and can be used to factorize content from dynamics [18], which are nonetheless modelled holistically. In contrast, our model has to infer plausible motion based on local interactions and, thus, understands dynamics in a more fine-grained way.

Controllable Synthesis of Object Dynamics. Since it requires to understand the interplay between their distinct parts, controlling the dynamics of articulated objects is a highly challenging task. Davis et al. [10] resort to modelling rigid objects as spring-mass systems and animate still image frames by evaluating the resulting motion equations. However, due to these restricting assumptions, their method is only applicable for small deviations around a rest state, thus not able model complex dynamics.

To reduce complexity, existing learning based approaches often focus on modelling human dynamics using low-dimensional, parametric representations such as keypoints [1, 71], thus preventing universal applicability. Moreover, as these approaches are either based on explicit action labels or require motion sequences as input, they cannot be applied to controlling single body parts. When intending to similarly obtain control over object dynamics in the pixel domain, previous methods use ground truth annotations such as holistic motion trajectories [17, 33] for simple object classes without articulation [13]. Hao et al. [23] step towards locally controlling the video generation process by predicting a single next images based on a given input frame and sets of sparse flow vectors. Their proposed approach, however, requires multiple flow vectors for each individual frame of a sequence to be predicted, thus preventing localized, fine-grained control. Avoiding such flaws and indeed allowing for localized control, our approach introduces a latent dynamics model, which is able to model complex, articulated motion based on an interaction at a *single* pixel.

3. Interactive Image-to-Video Synthesis

Given an image frame $x_0 \in \mathbb{R}^{H \times W \times 3}$, our goal is to interact with the depicted objects therein, i.e. we want to initiate a poke $p \in \mathbb{R}^2$ which represents a shift of a single location $l \in \mathbb{N}^2$ within x_0 to its new target location. Moreover, such pokes should also influence the remainder of the image in a natural, physically plausible way.

Inferring the implications of local interactions upon the entire object requires a detailed understanding of its articulation, and, thus of the interrelation between its various parts. Consequently, we require a structured and concise representation of the image x_0 and a given interaction. To this end, we introduce two encoding functions: an object encoder $\mathcal{E}_\sigma$ mapping images x onto a latent object state $\sigma = \mathcal{E}_\sigma(x)$, e.g. describing current object pose and appearance, and the encoder $\mathcal{E}_\phi$ translating the target location defined by p and l to a latent interaction $\phi = \mathcal{E}_\phi(p, l)$ now affecting the initially observed object state $\sigma_0 = \mathcal{E}_\sigma(x_0)$.

Eventually, we want to synthesize a video sequence depicting the response arising from our interaction with the image x_0, represented by means of σ_0 and ϕ. Commonly, such conditional video generation tasks are formulated by

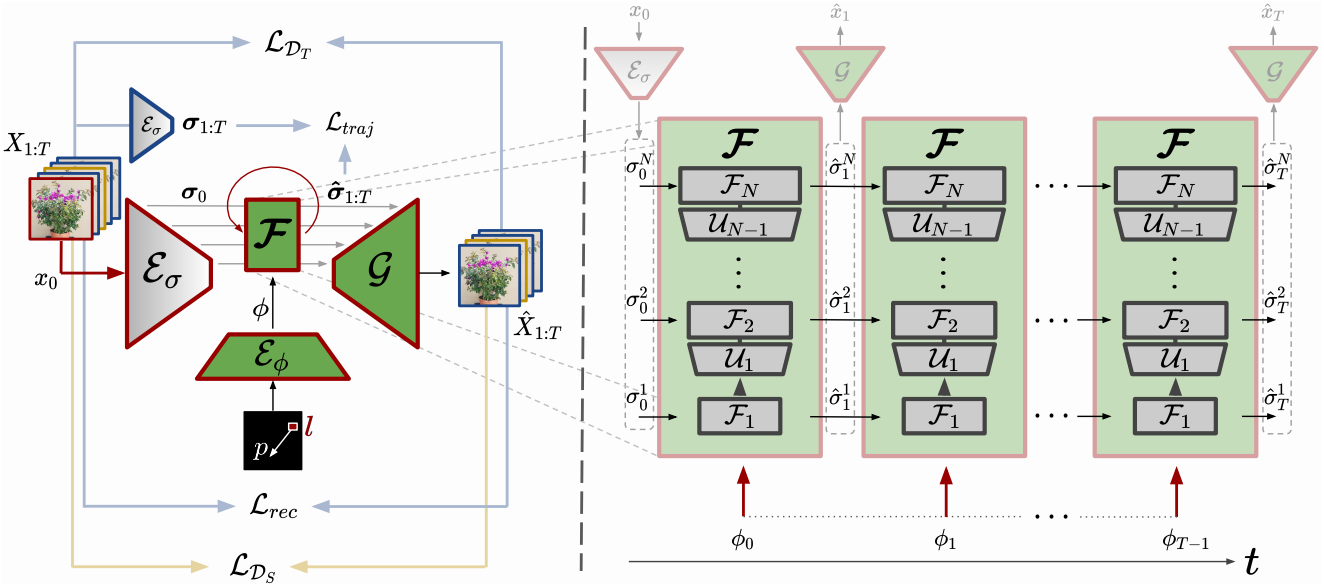

Figure 2. *Left:* Our framework for interactive image-to-video synthesis during training. *Right:* Our proposed hierarchical latent model $\boldsymbol{\mathcal{F}}$ for synthesizing dynamics, consisting of a hierarchy of individual RNNs $\mathcal{F}_n$, each of which operates on a different spatial feature level of the UNet defined by the pretrained encoder $\mathcal{E}_\sigma$ and the decoder $\mathcal{G}$. Given the initial object state $\boldsymbol{\sigma}_0 = [\mathcal{E}_\sigma(x_0)^1, ..., \mathcal{E}_\sigma(x_0)^N]$, $\boldsymbol{\mathcal{F}}$ predicts the next state $\hat{\boldsymbol{\sigma}}_{i+1} = [\hat{\sigma}_{i+1}^1, ..., \hat{\sigma}_{i+1}^N]$ based on its current state $\hat{\sigma}_i$ and the latent interaction $\phi_i = \mathcal{E}_\phi(p, l)$ at the corresponding time step. The decoder $\mathcal{G}$ finally visualizes each predicted object state $\hat{\sigma}_i$ in an image frame $\hat{x}_i$.

means of learning a video generator $\mathcal{G} : (\sigma_0, \phi) \rightarrow X_{1:T} = \{x_1, ..., x_T\}$ [33, 5]. Thus, $\mathcal{G}$ would both model object dynamics and infer their visualization in the RGB space. However, every object class has its distinct, potentially very complex dynamics which - affecting the *entire* object - must be inferred from a localized poke shifting a *single pixel*.Consequently, a model for interactive image-to-video synthesis has to understand the complex implications of the poke for the remaining object parts and, thus, requires to model these dynamics in sufficiently fine-grained and flexible way. Therefore, we introduce a dedicated object dynamics model inferring a trajectory of object states $[\sigma_0, \sigma_1, \dots, \sigma_T]$ representing an object's response to ϕ within an object state space Ω. As a result, $\mathcal{G}$ only needs to generate the individual images $x_i = \mathcal{G}(\sigma_i)$, thus decomposing the overall image-to-synthesis problem.

3.1. A Hierarchical Model for Object Dynamics

In physics, one would typically model the trajectory of object states $\sigma(t)$ as a dynamical system and, thus, describe it as an ordinary differential equation (ODE) [56, 6]

$$\dot{\sigma}(t) = f(\sigma(t), \phi(t)) \, , \, \sigma(0) = \sigma_0 \, , \tag{1}$$

with f the - in our case unknown - evolution function, $\dot{\sigma}$ its first time derivative, and $\phi(t) = \phi \; , \; \forall t \in [0, T]$ the latent external interaction obtained from the poke. Recent work proposes to describe f with fixed model assumptions, such as as an oscillatory system [10]. While this may hold in some cases, it greatly restricts applications to arbitrary, highly-articulated object categories. Avoiding such strong assumptions, the only viable solution is to learn a flexible prediction function $\mathcal{F}$ representing the dynamics in Eq. (1). Consequently, we base $\mathcal{F}$ on recurrent neural network models [24, 8] which can be interpreted as a discrete, first-order approximation[1] to Eq. (1) at time steps $i \in [0, T-1]$

$$\mathcal{F}(\sigma_i, \phi_i) = \sigma_{i+1} = \sigma_i + h \cdot f_a(\sigma_i, \phi_i) \, , \tag{2}$$

with h being the step size between two consecutive predicted states σ_i and σ_{i+1} and f_a an approximation to the derivative at σ_i [6, 48]. However, dynamics of objects can be arbitrarily complex and subtle such as leaves of a tree or plant fluttering in the wind. In such cases, the underlying evolution function f is expected to be similarly complex and, thus, involving only first-order derivatives when modelling Eq. (1) may not be sufficient. Instead, capturing also such high-frequency details actually calls for higher-order terms.

In fact, one can model an N-th order evolution function in terms of N first order ODEs, by introducing a hierarchy $\boldsymbol{\sigma} = [\sigma^1, ..., \sigma^N]$, $\sigma^1 = \sigma$ of state variables, the n-th element of which is proportional to the $(n-1)$-th order of discrete time derivative of the original variable σ [22, 48] Consequently, as those first order ODEs can be well approximated with Eq. (2), we can extend $\mathcal{F}$ to a hierarchy

[1] Order here means order of time derivative. For more information regarding this correspondence between ODEs and RNNs, see [20, 6, 7, 48].

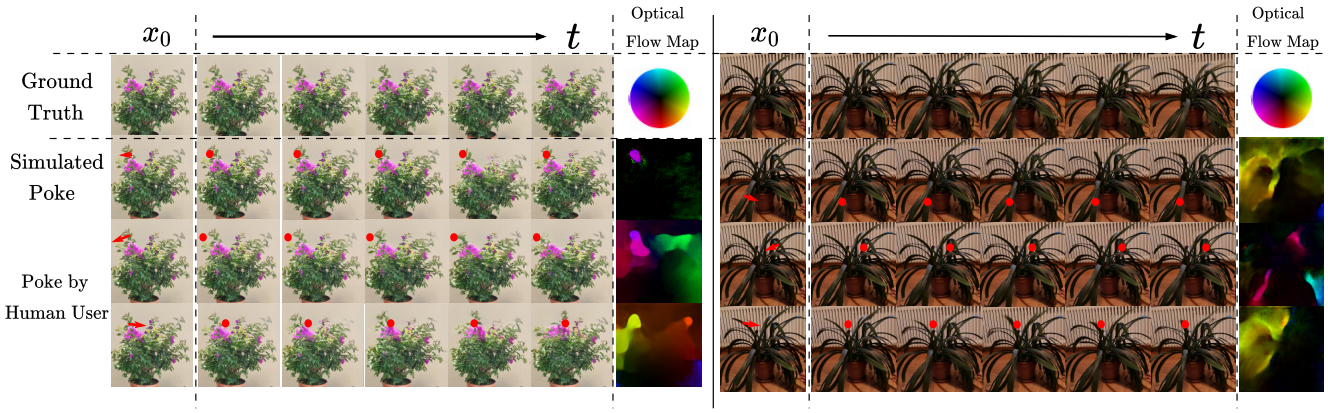

Figure 3. Visualization of the videos generated by our model for two distinct plants within our self-recorded PokingPlants-Dataset: The first row depicts a ground truth sequence from the test set. The second row contains a simulated poke (red arrow) based on this ground truth sequence, using the procedure described in Section 3.2. The last two rows show results of the model to pokes from human users. In the first column, the red arrow indicates the interaction and, thus, also the resulting target location, which is indicated as a red dot in the remaining columns. As the motions within the Poking-Plants dataset are sometimes subtle and not straightforward to detect, we also visualize the visual flow field, which was estimated based on the synthesized videos. We encourage the reader to also view the corresponding video results in the supplemental and on our project page https://bit.ly/3cxfA2L.

of predictors $\boldsymbol{\sigma} = \boldsymbol{\mathcal{F}} = [\mathcal{F}_1, ..., \mathcal{F}_N]$ using a sequence of N RNNs,

$$\sigma_{i+1}^n = \mathcal{F}_n(\sigma_i^n, \sigma_{i+1}^{n-1}), \sigma_0^n = \sigma_0 \tag{3}$$

each operating on the input of its predecessor, except for the lowest level $\mathcal{F}_1$, which predicts the coarsest approximation of the object states based on ϕ_i as

$$\sigma_{i+1}^1 = \mathcal{F}_1(\sigma_i^1, \phi_i)\,,\ \sigma_0^1 = \sigma_0\,. \tag{4}$$

A derivation of our this hierarchy is given in the Appendix C. However, while $\boldsymbol{\mathcal{F}}$ is able to approximate higher-order derivatives up to order N, thus being able to model fine-grained dynamics, we need to make sure that our decoder $\mathcal{G}$ actually captures these details when generating the individual image frames x_i.

Recent work on image synthesis indicates that standard decoder architectures being fed with latent encodings only on the bottleneck-level, are prone to missing out on subtle image details [29, 50], such as those arising from high motion frequencies. Instead, providing a decoder with latent information at each spatial scale has proven to be more powerful [52]. Hence, we model $\mathcal{G}$ to be the decoder of a hierarchical image-to-sequence UNet with the individual predictors $\mathcal{F}_n$ operating on the different spatial feature levels of $\mathcal{G}$. To maintain the hierarchical structure of $\boldsymbol{\mathcal{F}}$, we compensate for the resulting mismatch between the dimensionality of σ_i^{n-1} and σ_i^n by means of upsampling functions $\mathcal{U}_n$. Finally, to fully exploit the power of a UNet structure, we similarly model the encoder $\mathcal{E}_\sigma$ to yield a hierarchical object state $\boldsymbol{\sigma}_0 = [\mathcal{E}_\sigma(x_0)^1, ..., \mathcal{E}_\sigma(x_0)^N]$, which is the initial state to $\boldsymbol{\mathcal{F}}$. Thus, Eq. (3) becomes

$$\sigma_{i+1}^n = \mathcal{F}_n(\sigma_i^n, \mathcal{U}_{n-1}(\sigma_{i+1}^{n-1}))\,, \sigma_0^n = \mathcal{E}_\sigma(x_0)^n \tag{5}$$

with σ_i^n being the predicted object state at feature level $n \in [2, N]$ and time step i. Hence, at each time step we obtain a hierarchy of N latent object states $\boldsymbol{\sigma}_i = [\sigma_i^1, ..., \sigma_i^N]$ on different spatial scales, which are the basis for synthesizing image frames x_i by $\mathcal{G}$. Our full hierarchical predictor $\boldsymbol{\mathcal{F}}$ and its linkage to the overall architecture are shown in the right and left parts of Figure 2.

As our proposed model jointly adds more subtle details in the temporal *and* spatial domain, it accurately captures complex dynamics arising from interactions and simultaneously displays them in the image space.

3.2. Learning Dynamics from Poking

To learn our proposed model for interactive image-to-video synthesis, we would ideally have ground-truth interactions provided, i.e. actual pokes and videos of their immediate impact on the depicted object in x_0. However, gathering such training interactions is tedious and in some cases not possible at all, particularly hindering universal applicability. Consequently, we are limited to cheaply available video sequences and need to infer the supervision for interactions automatically.

While at inference a poke p represents an *intentional* shift of an image pixel in x_0 at location l, at training we only require to observe responses of an object to *some* shifts, as long as the inference task is well and naturally approximated. To this end, we make use of dense optical flow displacement maps [25] $D \in \mathbb{R}^{H \times W \times 2}$ between images x_0 and x_T of our training videos, i.e. their initial and last frames. Simulating training pokes in x_0 then corresponds to sampling pixel displacements $p = (D_{l_1,l_2,1}, D_{l_1,l_2,2})$ from D, with the video sequence $X_{1:T} = \{x_i\}_{i=1}^T$ being

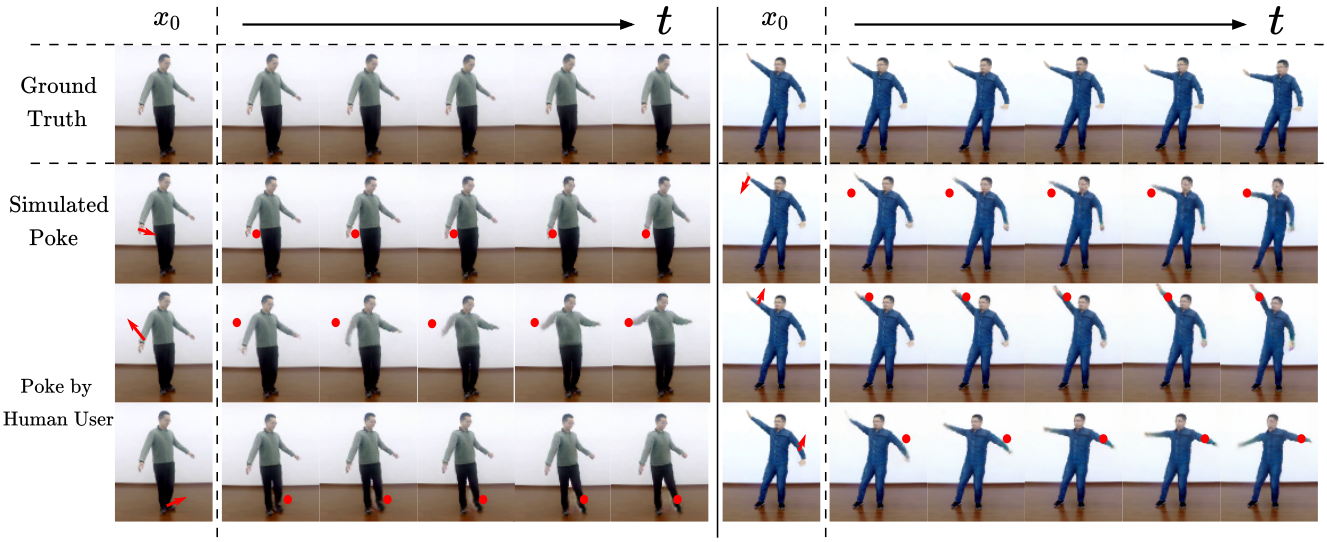

Figure 4. Visualization of the videos generated by our model for two actors within the test set of the iPER [36] dataset: The first row depicts the ground truth sequence. The second row contains a simulated poke (red arrow) based on this ground truth sequence, using the procedure described in Section 3.2. The last two rows show results of the model to pokes from human users. In the first column, the red arrow indicates the interaction and, thus, also the resulting target location, which is indicated as a red dot in the remaining columns. We encourage the reader to also view the accompanying video results in the supplemental and on our project page https://bit.ly/3cxfA2L.

a natural response to p. Using this, we train our interaction conditioned generative model to minimize the mismatch between the individual predicted image frames $\hat{x}_i = \mathcal{G}(\boldsymbol{\mathcal{F}}(\boldsymbol{\sigma}_{i-1}, \phi_{i-1}))$ and those of the video sequence as measured by the perceptual distance [28]

$$\mathcal{L}_{rec} = \sum_{i=1}^{T} \sum_{k=1}^{K} \|\Phi_k(x_i) - \Phi_k(\hat{x}_i)\|_1 \,, \tag{6}$$

where Φ_k denotes the k-th layer of a pre-trained VGG [55] feature extractor.

However, due to the direct dependence of $\mathcal{G}$ on $\boldsymbol{\sigma}_i$, during end-to-end training the state space Ω is continuously changing. Thus, learning object dynamics by means of our hierarchical predictor $\boldsymbol{\mathcal{F}}$ is aggravated. To alleviate this issue, we propose to first learn a fixed object space state Ω by pretraining $\mathcal{E}_\sigma$ and $\mathcal{G}$ on image reconstruction. Training $\boldsymbol{\mathcal{F}}$ on Ω to capture dynamics depicted in $\{x_i\}_{i=0}^{T}$ is then performed by predicting states $\hat{\boldsymbol{\sigma}}_i = \boldsymbol{\mathcal{F}}(\hat{\boldsymbol{\sigma}}_{i-1}, \phi_i)$ approaching the individual states $\boldsymbol{\sigma}_i = [\mathcal{E}_\sigma(x_i)^1, ..., \mathcal{E}_\sigma(x_i)^N]$ of the target trajectory, while simultaneously fine-tuning $\mathcal{G}$ to compensate for prediction inaccuracy, using

$$\mathcal{L}_{traj} = \sum_{i=1}^{T} \sum_{n=1}^{N} \|\mathcal{E}_\sigma(x_i)^n - \hat{\sigma}_i^n\|_2 \,. \tag{7}$$

Finally, to improve the synthesis quality we follow previous work [9, 65] by training discriminators $\mathcal{D}_S$ on frame level and $\mathcal{D}_T$ on the temporal level, resulting in loss functions $\mathcal{L}_{\mathcal{D}_S}$ and $\mathcal{L}_{\mathcal{D}_T}$. Our overall optimization objective then reads

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{traj} \cdot \mathcal{L}_{traj} + \lambda_S \cdot \mathcal{L}_{\mathcal{D}_S} + \lambda_{\mathcal{D}_T} \cdot \mathcal{L}_{D_T} \tag{8}$$

with hyperparameters λ_{traj}, λ_S and λ_T. The overall procedure for learning our network is summarized in Fig. 2.

4. Experiments

In this section we both qualitatively and quantitatively analyze our model for the task of interactive image-to-video synthesis. After providing implementation details, we illustrate qualitative results and analyze the learned object interrelations. Finally, we conduct a quantitative evaluation of both visual quality of our generated videos and the plausibility of the motion dynamics depicted within.

4.1. Implementation Details and Datasets

Foreground-Background-Separation. While our model should learn to capture dynamics associated with objects initiated by interaction, interaction with areas corresponding to the background should be ignored. However, optical flow maps often exhibit spurious motion estimates in background areas, which may distort our model when considered as simulated object pokes for training. To suppress these cases we only consider flow vector exhibiting sufficient motion magnitude as valid pokes, while additionally training our model to ignore background pokes. More details are provided in Appendix E.2. Examples for responses to background interactions can be found in the videos on our project page https://bit.ly/3cxfA2L.

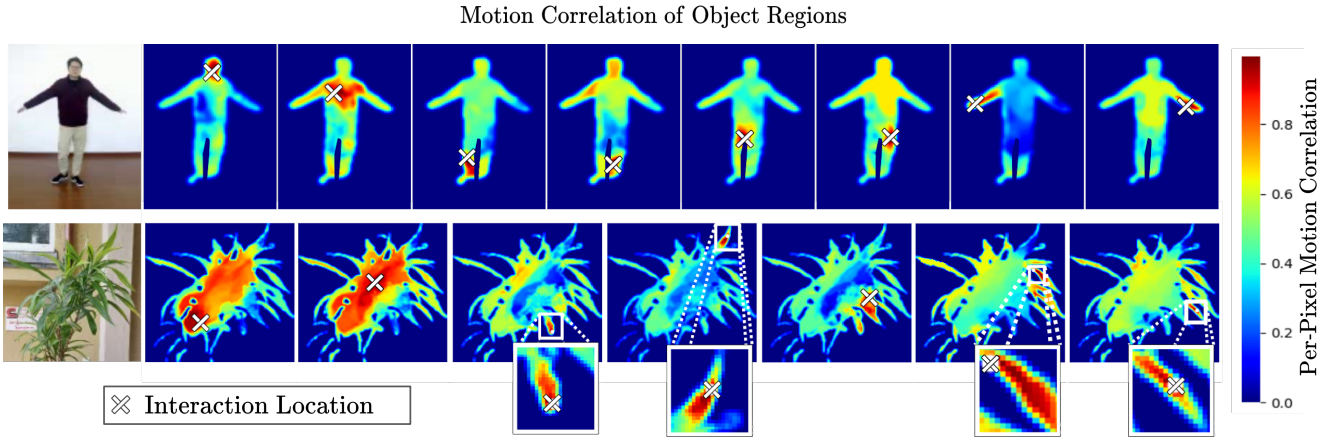

Figure 5. Understanding object structure: By performing 100 random interactions at the same location l within a given image frame x_0 we obtain varying video sequences, from which we compute motion correlations for l with all remaining pixels. By mapping these correlations to the pixel space, we visualize distinct object parts. For a detailed discussion, refer to Section 4.2.

Model Architecture and Training. The individual units $\mathcal{F}_n$ of our hierarchical dynamics model are implemented as Gated Recurrent Units (GRU) [8]. The depth N of our hierarchy of predictors varies among datasets and videos resolutions. For the experiments shown in the main paper, we train our model to generate sequences of 10 frames and spatial size 128×128, if not specified otherwise. For training, we use ADAM [30] optimizer with parameters $\beta = (0.9, 0.99)$ and a learning rate of 10^{-4}. The batch size during training is 10. The weighting factors for our final loss function are chosen as $\lambda_{traj} = 0.1$, $\lambda_{\mathcal{D}_S} = 0.2$ and $\lambda_{\mathcal{D}_T} = 1$. More details regarding architecture, hyperparameters and training can be found in the Appendix E. Additionally, alternate training procedures for different parameterizations of the poke and the latent interaction $\phi(t)$ are presented in Appendix B.

Datasets We evaluate the capabilities of our model to understand object dynamics on four datasets, comprising the highly-articulated object categories of humans and plants.

Poking-Plants (PP) is a self-recorded dataset showing 27 video sequences of 13 different types of pot plants in motion. As the plants within the dataset have large variances in shape and texture, it is particularly challenging to learn a single dynamics model from those data. The dataset consists of overall 43k image frames, of which we use a fifth as test and the remainder as train data. We will release the dataset upon publication.

iPER [36] is a human motion dataset containing 206 video sequences of 30 human actors of varying shape, height gender and clothing, each of which is recorded in 2 videos showing simple and complex movements. We follow the predefined train test split with a train set containing 180k frames and a test set consisting of 49k frames.

Tai-Chi-HD [54] is a video collection of 280 in-the-wild Tai-Chi videos of spatial size 256×256 from youtube which are subdivided into 252 train and 28 test videos, consisting of 211k and 54k image frames. It contains large variances in background and also camera movements, thus serving as an indicator for the real-world applicability of our model. As the motion between subsequent frame is often small, we downsample the datasets by a factor of two.

Dataset	PP	iPER	Tai-Chi	H3.6m
SAVP [33]	136.81	150.39	309.62	160.32
IVRNN [5]	184.54	181.50	202.33	465.55
SRVP [18]	283.90	245.13	238.71	174.65
Ours	**89.67**	**144.92**	**182.28**	**119.89**

Table 1. Comparison with recent work from video prediction. Across datasets, our approach obtains considerably lower FVD-scores.

Human3.6m [26] is large scale human motion dataset, containing video sequences of 7 human actors performing 17 distinct actions. Following previous work on video prediction [67, 45, 18], we centercrop and downsample the videos to 6.25 Hz and use actors S1,S5,S6,S7 and S8 for training and actors S9 and S11 for testing.

4.2. Qualitative Analysis

Interactive Image-to-Video-Synthesis. We now demonstrate the capabilities of our model for the task of interactive image-to-video synthesis. By poking a pixel within a source image x_0, i.e. defining the shift p between the source location l and the desired target location, we require our model to synthesize a video sequence showing the object part around l approaching the defined target location. Moreover, the model should also infer plausible dynamics for the remaining object parts. Given x_0, we visualize the video sequences generated by our model for $i)$ simulated pokes obtained from optical flow estimates (see Sec. 3.2) from test videos and $ii)$ pokes initiated by human users. We compare them to the corresponding ground truth videos starting with x_0. For the object category of plants, we also

Dataset	PP		iPER [36]		TaiChi [54]		Human3.6m [26]	
Method	Ours	Hao et al. [23]	Ours	Hao et al. [23]	Ours	Hao et al. [23]	Ours	Hao et al. [23]
FVD ↓	**174.18**	361.51	**220.34**	235.08	**167.94**	341.79	**129.62**	259.92
LPIPS ↓	**0.10**	0.16	**0.07**	0.11	**0.12**	**0.12**	**0.08**	0.10
PSNR ↑	20.80	**21.28**	**22.67**	21.09	**20.58**	20.41	**22.95**	22.81
SSIM ↑	**0.78**	0.72	**0.89**	0.88	0.75	**0.78**	0.91	**0.93**

Table 2. Comparison on controlled video synthesis. We compare with Hao et al. [23] which we consider the closest previous work to ours.

show optical flow maps between the source frame and the last frame for each generated video, to visualize the overall motion therein.

Figure 3 shows examples for two distinct plants from the PokingPlants dataset of very different appearance and shape. In the left example we interact with the same pixel location varying both poke directions and magnitudes. Our model correctly infers corresponding object dynamics. Note that the interaction location corresponds to a very small part of the plant, initiating both subtle and complex motions. This demonstrates the benefit of our hierarchical dynamics model. As shown by the visualized flow, the model also infers plausible movements for object parts not directly related to the interaction, thus indicating our model to also understand long term relations between distinct object regions. In the right example we observe similar results when interacting with different locations.

As our model does not make any assumptions on the depicted object to manipulate, we are not restricted to particular object classes. To this end we also consider the category of humans which is widely considered for various video synthesis tasks. Figure 4 shows results for two unseen persons from the iPER test set. Again our model infers natural object dynamics for both the interaction target location and the entire object. This highlights that our model is also able to understand highly structured dynamics and natural interrelations of object parts independent of object appearance. Analogous results for the Tai-Chi-HD and Human3.6m datasets are contained in the Appendix A.

Understanding Object Structure. We now analyze the interrelation of object parts learned by our model, i.e. how pixels depicting such parts correlate when interacting with a specific, fixed interaction location.

To this end, we perform 100 random interactions for a given, fixed source frame x_0 at the fixed location $l = (l_1, l_2)$ of different magnitudes and directions. This results in varying videos $\{\hat{X}_k\}_{k=1}^{100}$, exhibiting distinct dynamics of the depicted object. To measure the correlation in motion of all pixels with respect to the interaction location, we first obtain their individual motion for each generated video using optical flow maps between x_0 and the last video frames. Next, for each pixel we compute the L2-distance in each video to the respective interaction poke based on a $[\text{magnitude}, \text{angle}]$ representation, thus obtaining a 100-dimensional difference vector. To measure the correlation for each pixel with l, we now compute the variance over each difference vector to obtain correlation maps.

Figure 5 visualizes such heatmaps for different interaction locations l in the same source frame x_0 for both plants and humans. For humans, we obtain high correlations for the body parts around l, indicating our model to actually understand the human body structure. Considering the plant, we see that poking locations on its trunk (first two columns) intuitively results in highly similar movements of all those pixels close to the trunk. The individual leaves farther away are not that closely correlated, as they might perform oscillations in higher frequencies superposing the generally low-frequency movements of the trunk. When poking these leaves, however, mostly directly neighbouring pixels exhibit high correlation.

4.3. Quantitative Evaluation

To quantitatively demonstrate the ability of our approach to synthesize plausible object motion, we compare with the current state-of-the-art of RNN-based video prediction [33, 5, 18]. For all competing methods we used the provided pretrained models, where available, or trained the models using official code. Moreover, to also evaluate the controllable aspect of our approach, we compare against Hao et al. [23], an approach for controlled video synthesis by predicting independent single frames from sparse sets of flow vectors, which we consider the closest previous work to interactive image-to-video synthesis. We trained their method on our considered datasets using their provided code.

Evaluation Metrics. For evaluation we utilize the following metrics:

Motion Consistency. The Fréchet-Video-Distance (FVD, lower-is-better) is the standard metric to evaluate video predictions tasks, as it is sensitive to the visual quality, as well as plausibility and consistency of motion dynamics of a synthesized video. Moreover it has been shown to correlate well with human judgement [58]. All FVD-scores reported hereafter are obtained from video sequences of length 10.

Prediction accuracy. Our model is trained to understand the global impact of localized pokes on the overall object dynamics and to infer the resulting temporal object articulation. To evaluate this ability, we report distance scores against the ground-truth using three commonly used frame-wise metrics, averaged over time: SSIM (higher-is-better) and PSNR (higher-is-better) directly compare the

predicted and ground truth frames. Since they are based on L2-distances directly in pixel space, they are known disregard high-frequency image details, thus preferring blurry predictions. To compensate, we also report the LPIPS-score (lower-is-better[74]), comparing images based on distance between activations of pretrained deep neural network and, thus exhibit better correlation with human judgement.

Comparison to other methods. We evaluate our approach against SAVP [33], IVRNN [5] and SRVP [18] which are RNN-based state-of-the-art methods in video prediction based on FVD scores. We train these models to predict video sequences of length 10 given 2 context frames corresponding to the evaluation setting. Note, that our method generates sequences consisting of 10 frames based on a *single* frame and a poke. Following common evaluation practice in video prediction [33, 5, 18], we predict sequences are 64×64 for all models, including ours. The results are shown in Tab. 1. Across datasets, our approach results significantly lower FVD-scores demonstrating the effectiveness of our approach to both infer and visualize plausible object dynamics. Particularly on the plants dataset our model performs significantly better than the competitors due to our proposed hierarchic model for capturing fine-grained dynamics.

To quantitatively evaluate our ability for controlled video synthesis, we now compare against Hao et al. [23]. For them to predict videos $[x_0, \dots, x_T]$ of length T, we first sample k flow vectors between x_0 and x_1. We then extend these initial vectors to discrete trajectories of length $T-1$ in the image space by tracking the initial points using consecutive flow vectors between x_i and x_{i+1}. Video prediction is then performed by individually warping x_0 given the shift of its pixels to the intermediate locations at step i in these trajectories. Following their protocol we set $k = 5$. Note, that we only require *a single interactive poke* of arbitrary length for synthesize videos. Tab. 2 compares both approaches. We see that our model performs significantly better by a large margin, especially in FVD measuring the video quality and temporal consistency. This is explained by temporal object dynamics not even being considered in [23], in contrast to our dedicated hierarchical dynamics model. Further, their image warping-based approach typically results in blurry image predictions. Across datasets we outperform their method also in LPIPS scores indicating our proposed method to produce more visually compelling results.

4.4. Ablation Study

Finally, we conduct ablation studies to analyze the effects of the individual components of our proposed method. To limit computational cost, we conduct all ablation experiments using videos of spatial size 64×64. *Ours RNN* indicates our model trained with a common GRU consisting of three individual cells at the latent bottleneck of the UNet instead of our proposed hierarchy of predictors $\boldsymbol{\mathcal{F}}$. For fair evaluation with the baselines we also choose $N = 3$ as depth of the hierarchy. Further, we also evaluate baselines without considering the loss term $\mathcal{L}_{traj}$ which compensates for the prediction inaccuracy of $\boldsymbol{\mathcal{F}}$ (*Ours w/o* $\mathcal{L}_{traj}$) and without pretraining the object encoder $\mathcal{E}_\sigma$ (*Ours single-stage*). Tab. 3 summarizes the ablation results, indicating the the benefits of each individual component across reported metrics and used datasets. We observe the largest impact when using our hierarchical predictor $\boldsymbol{\mathcal{F}}$, thus modelling higher-order terms in our dynamics model across all spatial feature scales. Looking at the remaining two ablations, we also see improvements in FVD-scores. This concludes that operating on a pretrained, consistent object state space Ω and subsequently accounting for the prediction inaccuracies, leads to significantly more stable learning. An additional ablation study on the effects of varying the highest order N of modeled derivative by our method is provided in Appendix D.

Dataset	PP			iPER [36]		
Method	LPIPS ↓	PSNR ↑	FVD ↓	LPIPS ↓	PSNR ↑	FVD ↓
Ours RNN	0.08	20.92	175.42	0.07	23.02	250.64
Ours w/o $\mathcal{L}_{traj}$	0.07	21.53	110.00	0.05	22.82	203.26
Ours (single)	0.07	21.41	115.65	0.06	23.09	220.82
Ours full	**0.06**	**21.81**	**89.67**	**0.05**	**23.11**	**144.92**

Table 3. Ablation study on the PokingPlants and iPER datasets

4.5. Additional Experiments

We also conduct various alternate experiments, including the generalization of our model to unseen types of plants and different interpretations of the poke p, resulting in different capabilities of our model. These experiments are contained in the Appendix B, which also contains many videos showing our results.

5. Conclusion

We propose a generative model for *interactive image-to-video synthesis* which learns to understand the dynamics of articulated objects by capturing the interplay between the distinct object parts, thus allowing to synthesize videos showing natural responses to localized user interactions. The model can be flexibly learned from unlabeled video data without limiting assumptions on object shape. Experiments on a range of datasets prove the plausibility of generated dynamics and indicate the model to produce visually compelling video sequences.

Acknowledgements

This research is funded in part by the German Federal Ministry for Economic Affairs and Energy within the project "KI-Absicherung – Safe AI for automated driving" and by the German Research Foundation (DFG) within project 421703927.

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4):64, 2020. 1, 2

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[3] 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*, 2020. 1

[4] Biagio Brattoli, Uta Büchler, Anna-Sophia Wahl, Martin E. Schwab, and Björn Ommer. LSTM self-supervision for detailed behavior analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3747–3756, 2017. 1

[5] L. Castrejon, N. Ballas, and A. Courville. Improved conditional vrnns for video prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 6, 7, 8

[6] Bo Chang, Minmin Chen, Eldad Haber, and Ed H. Chi. AntisymmetricRNN: A dynamical system view on recurrent neural networks. In *International Conference on Learning Representations*, 2019. 3

[7] Bo Chang, Lili Meng, Eldad Haber, Lars Ruthotto, David Begert, and Elliot Holtham. Reversible architectures for arbitrarily deep residual neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, Apr. 2018. 3

[8] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014. 3, 6

[9] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets, 2020. 2, 5

[10] Abe Davis, Justin G. Chen, and Frédo Durand. Image-space modal bases for plausible manipulation of objects in video. *ACM Trans. Graph.*, 2015. 2, 3

[11] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. volume 80 of *Proceedings of Machine Learning Research*, 2018. 1, 2

[12] Michael Dorkenwald, Uta Büchler, and Björn Ommer. Unsupervised magnification of posture deviations across subjects. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8253–8263, 2020. 1

[13] Frederik Ebert, Chelsea Finn, Alex X. Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, 2017, Proceedings*, volume 78 of *Proceedings of Machine Learning Research*, pages 344–356. PMLR, 2017. 2

[14] Patrick Esser, Johannes Haux, Timo Milbich, and Björn Ommer. Towards learning a realistic rendering of human behavior. In *European Conference on Computer Vision - Workshops*, 2018. 1

[15] 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. 1

[16] Patrick Esser, E. Sutter, and B. Ommer. A variational u-net for conditional appearance and shape generation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 1

[17] Chelsea Finn, Ian J. Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 64–72, 2016. 2

[18] Jean-Yves Franceschi, Edouard Delasalles, Mickael Chen, Sylvain Lamprier, and P. Gallinari. Stochastic latent residual video prediction. *ArXiv*, abs/2002.09219, 2020. 2, 6, 7, 8

[19] Hang Gao, Huazhe Xu, Qi-Zhi Cai, Ruth Wang, Fisher Yu, and Trevor Darrell. Disentangling propagation and generation for video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[20] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, dec 2017. 3

[21] Naama Hadad, Lior Wolf, and Moni Shahar. A two-step disentanglement method. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018. 1

[22] J.K. Hale. *Ordinary Differential Equations*. Dover Books on Mathematics Series. Dover Publications, 2009. 3

[23] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *CVPR*, 2018. 2, 7, 8

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 3

[25] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 4

[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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 6, 7

[27] B. Jin, Y. Hu, Q. Tang, J. Niu, Z. Shi, Y. Han, and X. Li. Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4553–4562, 2020. 1

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 5

[29] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

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

[31] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2018. 2

[32] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A conditional flow-based model for stochastic video generation. In *International Conference on Learning Representations*, 2020. 2

[33] Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *CoRR*, 2018. 1, 2, 3, 6, 7, 8

[34] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[35] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P. Xing. Dual motion GAN for future-flow embedded video prediction. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1762–1770. IEEE Computer Society, 2017. 2

[36] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 5, 6, 7, 8

[37] Yu-Lun Liu, Yi-Tung Liao, Yen-Yu Lin, and Yung-Yu Chuang. Deep video frame interpolation using cyclic frame generation. In *Proceedings of the 33rd Conference on Artificial Intelligence (AAAI)*, 2019. 2

[38] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala. Video frame synthesis using deep voxel flow. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[39] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Björn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[40] C. Lu, M. Hirsch, and B. Schölkopf. Flexible spatio-temporal networks for video prediction. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, pages 2137–2145. IEEE, 2017. 2

[41] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[42] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. 2016. 4th International Conference on Learning Representations, ICLR 2016. 2

[43] 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. 1

[44] Timo Milbich, Miguel Bautista, Ekaterina Sutter, and Björn Ommer. Unsupervised video understanding by reconciliation of posture similarities. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[45] Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin P Murphy, and Honglak Lee. Unsupervised learning of object structure and dynamics from videos. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 1, 2, 6

[46] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[47] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[48] Murphy Yuezhen Niu, Lior Horesh, and Isaac Chuang. Recurrent neural networks in the eye of differential equations, 2019. 3

[49] Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Video generation from single semantic label map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[50] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[51] Fitsum Reda, Guilin Liu, Kevin Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. *SDC-Net: Video Prediction Using Spatially-Displaced Convolution: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*. 2018. 2

[52] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 4

[53] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[54] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019. 2, 6, 7

[55] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 5

[56] Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Westview Press, 2000. 2, 3

[57] S. Tulyakov, Ming-Yu Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1526–1535, 2018. 2

[58] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, 2018. 7

[59] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, 2016. 2

[60] Sjoerd van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *International Conference on Learning Representations*, 2018. 2

[61] Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V. Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 2019. 1, 2

[62] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 2

[63] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016. 2

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

[65] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018. 2, 5

[66] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *International Conference on Learning Representations*, 2020. 2

[67] Nevan Wichers, Ruben Villegas, Dumitru Erhan, and Honglak Lee. Hierarchical long-term video prediction without supervision. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. 1, 2, 6

[68] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[69] Yue Wu, Rongrong Gao, Jaesik Park, and Qifeng Chen. Future video synthesis with object motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[70] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, pages 1106–1125, 2019. 2

[71] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 2

[72] Pradeep Yarlagadda and Björn Ommer. From meaningful contours to discriminative object shape. In *Computer Vision – ECCV 2012*, pages 766–779, 2012. 1

[73] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[74] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 8