Unsupervised Bi-directional Flow-based Video Generation from one Snapshot

Lu Sheng¹,², Junting Pan², Jiaming Guo³, Jing Shao², Xiaogang Wang⁴, Chen Change Loy⁵
¹School of Software, Beihang University  ²SenseTime Group Limited  
³Institute of Computing Technology, Chinese Academy of Sciences  
⁴CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong  
⁵SenseTime-NTU Joint AI Research Centre, Nanyang Technological University  

lsheng@ee.cuhk.edu.hk

Abstract

Imagining multiple consecutive frames given one single snapshot is challenging, since it is difficult to simultaneously predict diverse motions from a single image and faithfully generate novel frames without visual distortions. In this work, we leverage an unsupervised variational model to learn rich motion patterns in the form of long-term bi-directional flow fields, and apply the predicted flows to generate high-quality video sequences. In contrast to the state-of-the-art approach, our method does not require external flow supervisions for learning. This is achieved through a novel module that performs bi-directional flows prediction from a single image. In addition, with the bi-directional flow consistency check, our method can handle occlusion and warping artifacts in a principle manner. Our method can be trained end-to-end based on arbitrarily sampled natural video clips, and it is able to capture multi-modal motion uncertainty and synthesizes photo-realistic novel sequences. Quantitative and qualitative evaluations over synthetic and real-world datasets demonstrate the effectiveness of the proposed approach over the state-of-the-art methods.

1. Introduction

We wish to address the problem of training a deep generation model for imagining photorealistic videos from just a single image. The problem is a non-trivial task as the model cannot only guess plausible dynamics conditioned on static contents. The problem is thus significantly harder than motion estimation task or video prediction problem in which paired consecutive images are assumed. Under the single-image constraint, choosing a suitable representation learning method becomes critical to the final visual quality and plausibility of the rendered novel sequences.

Existing methods such as MoCoGAN [26], VGAN [28], Visual Dynamics [35], and FRGAN [37] either directly render RGB pixel values or residual images for the modeling of dynamics in novel sequences, but they usually distort appearance patterns and are thus short for preserving visual quality. A recent method proposed by Li et al. [12] learned dense flows to propagate pixels from the reference image directly to the novel sequences, offering a better chance to generate visually plausible videos. However, Li et al. [12] require external and accurate flow supervisions (generated from SpyNet [38] to train the flow generation component, and the inherent warping artifacts were not handled in a principle way. Warping artifacts frequently occur in the generated novel frames, such as the examples in Fig. 1.

In this paper, we adopt the same notion of ‘per-pixel flow propagation’ as in Li et al. [12] but with the following consideration: (1) how to learn robust content-aware flow distributions in an unsupervised manner, without any external flow supervision; (2) how to synthesize photorealistic frames while eliminating warping artifacts such as occlusions and warping duplicates in a principle manner.
To this end, we propose an unsupervised bi-directional flow-based video generation framework solely based on one single image, named as ImagineFlow, which tackles the aforementioned challenges in a unified way. Our model has three appealing properties:

(i) **End-to-End Unsupervised Learning** – It allows end-to-end unsupervised learning to generate novel video from a single snapshot based on per-pixel flow propagation. The unsupervised learning module is new in the literature. It relaxes the need of external flows for supervision. The proposed learning is incredibly convenient and powerful in our task as it does not require laborious ground-truth annotations or preparation and the training data is nearly infinite and rich in motion patterns.

(ii) **Bi-directional Flow Generation** – We formulate a bi-directional flow generator, which outputs forward flows (input image $\rightarrow$ target images), and the backward flows (target frames $\rightarrow$ input image) simultaneously. The bi-directional flows can self-regularize each other according to a well-known cycle consistency criteria, i.e., valid flows always have corresponding inverse flows back to their original locations. The resultant flow distributions are within reasonable flow manifolds even they are learned without explicit flow supervision. In contrast, Li et al. [12] only generate the backward flows, whose reliability is governed by explicit flow supervisions.

(iii) **Occlusion-aware Image Synthesis** – Another merit of the proposed bi-directional flows generation is that the cycle consistency of flows allows us to detect occlusions in a robust and principled manner. Occlusions can be detected in areas where cycle consistency is violated. Our model can leverage the occlusions inferred to help determine between per-pixel flow propagation or pixel hallucination for novel image generation. Specifically, our occlusion-aware image synthesis inputs bilinear warped [38] novel frames overlaid with corresponded visibility (i.e., non-occluded) masks, and then employs a learnable mapping module that projects the warped frames onto the space of natural images. This module generates reasonable contents in the disoccluded area and seamlessly repairs warping duplicates simultaneously.

Ablation study validates the effectiveness of our ImagineFlow model. Extensive experimental evaluations also demonstrate its superior qualitative and quantitative performance over state-of-the-art methods [35, 28, 1, 26, 12].

2. Related Work

**Motion Prediction.** Single image based video generation is closely related to the motion prediction problem. Given an observed image or a short video clip, various methods have been proposed to predict dense future motions by optical flows [21, 32, 30], object trajectories [31], difference or residual images [35], or deep visual representations [27]. While most methods follow a deterministic manner, a few seminal works also try to characterize the uncertainties in the predicted motions, based on probabilistic models such as conditional variational autoencoders [35, 30] or generative adversarial networks [13]. Our work falls into a probabilistic motion modeling framework. We differ to aforementioned studies in that we aim at employing the predicted motion distribution to synthesize structurally coherent novel frames with reasonable motions. This requires new formulation for occlusion reasoning and frame generation.

**Motion-based Video Generation** Video generation can be roughly categorized into two classes according to whether it takes condition or not. A series of unconditioned video generation methods synthesize novel videos from scratch, with different learning techniques such as adversarial learning, motion/content separation, recurrent neural networks [25, 26, 22]. The visual qualities of their outputs are still not satisfactory to generate photorealistic videos. Conditioned video generation appears to be a more promising choice to generate visually plausible videos. A category of approaches predict novel frames from consecutive input frames [9, 25, 20, 16, 17, 3]. Another category, which matches our problem setting, predicts future frames just based on one still image [37, 12, 1, 35]. Some single image based methods characterize motions as feature filters or middle-level transformations [1, 29, 35]. These methods usually fail to preserve appearance patterns. To achieve photorealistic generation quality, Li et al. [12] also apply flows as the medium, but the method requires ground-truth flow supervisions and no explicit occlusion handling is introduced. By contrast, our approach is unsupervised, as it learns to generate bi-directional flows directly through the constrain of cycle consistency. The flows are reliable even without explicit supervisions, and they permit occlusion detection in a principle manner. With the structurally coherent flow fields, our method is more effective in rendering realistic video clips. Flow-based warping artifacts are eliminated by an additional occlusion-aware synthesis module.

**Applications and Regularizations for Flows.** Flows have been adopted for various tasks, such as video enhancement by task-oriented flow [34], video interpolation and extrapolation by voxel flow [15], gaze direction detection [5] and novel view synthesis [38]. Reliable flows often require regularizations to strengthen its structural coherence. Some approaches regularize the structures explicitly by flow supervisions [2, 12] or implicitly by adversarial networks [13]. In this work, inspired by the cross validation in stereo and optical flow estimations, we found that simply with cycle flow consistency, the learned flow space can already be effectively enforced with structural coherence without laborious human labeling or preprocessing.
3. Methodology

3.1. Problem Definition

Our task is to learn a probabilistic distribution \( p(\mathcal{I}_T | \mathcal{I}_0) \) conditioned on a reference frame \( \mathcal{I}_0 \), and then sample novel sequences \( \mathcal{I}_T \) from this conditioned distribution, where \( \mathcal{T} = \{1, \ldots, T\} \) are the frame indices.

In this work, we aim at simultaneously predicting a set of backward flows \( \mathcal{W}_b \) from tentative novel frames \( \mathcal{I}_T \) to the reference frame \( \mathcal{I}_0 \), and corresponding forward flows \( \mathcal{W}_f \) inversely from \( \mathcal{I}_0 \) to \( \mathcal{I}_T \), and then leveraging the predicted bi-directional flow fields to generate novel sequences \( \mathcal{I}_T \).

More specifically, this task is equivalent to learning 1) a bi-directional flow generator \( p_\phi(\mathcal{W}_f, \mathcal{W}_b | z, \mathcal{I}_0) \) that is also conditioned on a motion code \( z \), and 2) an occlusion-aware image synthesis module \( R_\omega(\cdot) \). The random motion code \( z \) may be sampled from a standard Gaussian distribution. \( \phi \) and \( \omega \) are network parameters. The complete model is learned from a training set of sequence pairs \( \{(\mathcal{I}_T^{(n)}, \mathcal{I}_0^{(n)})\}_{n=1,...,N} \) under an unsupervised manner.

3.2. Recap: Image Synthesis by Backward Warping

Given the backward flows \( \mathcal{W}_b \) from the tentative target frames to the reference frame \( \mathcal{I}_0 \), the synthesized frames are usually warped via bilinear sampling from \( \mathcal{I}_0 \) [6]:

\[
\mathcal{I}_{t=0} = \mathcal{F}(x, \mathcal{W}_b^t | \mathcal{I}_0) = \mathcal{I}_0(\mathcal{W}_b^t(x) + x), \forall t \in \mathcal{T}. \tag{1}
\]

Existing flow-based models, such as Li et al. [12], only apply the backward flows to generate novel frames. However, in the framework of unsupervised learning, using backward flows alone is less favored due to three issues:

**Warping Artifacts.** Backward warping operations will produce artifacts due to occlusions. Occlusions result in unfilled holes and generate warping duplicates in the warped images.

**Motion Inconsistency.** Since unsupervised backward flow learning usually applies photometric consistency to learn the flow space, the learned backward flow distributions may not align with the real flow space, especially when the sequences contain plain area or repeated patterns. Reliable flows should be cross-consistent outside the occlusion regions, i.e., an object in the target frames should always be able to predict reliable inverse flows back to the object in the reference frame, as illustrated in Fig. 2, similar to the concerns arisen in optical flow and stereo estimation [4].

**Structure Inconsistency.** Moreover, \( \mathcal{W}_b \) and novel frames (not the reference frame) are co-aligned in their spatial distributions, as shown in Fig. 2. It means that the backward flows do not only have to capture pixel-wise motions but also need to present the spatial structures in the target frames. Unfortunately, naïve unsupervised learning usually tends to generate flows aligned with the condition \( \mathcal{I}_0 \) rather than novel sequences, thus neither the pixel-wise motion nor the spatial alignment can be well discovered.

3.3. Bi-directional Flow Generation

Different from the backward flows, the forward flows \( \mathcal{W}_f \) are otherwise consistent with the spatial structure of \( \mathcal{I}_0 \), as shown in Fig. 2. And ideally \( \mathcal{W}_f \) and \( \mathcal{W}_b \) should be cross-consistent except the occlusion regions. Thus we also learn the forward flows \( \mathcal{W}_f \) as an auxiliary output concurrently with the backward flows \( \mathcal{W}_b \), and exploit these flows to regularize the spatial structures and enforce motion consistency of the backward flows. Moreover, the cross-consistency between the bi-directional flows also give cues for the occlusion detection.

We generate bi-directional flows from a bi-directional flow generator \( p_\phi(\mathcal{W}_f, \mathcal{W}_b | z, \mathcal{I}_0) \), with two parallel output branches, as visualized in Fig. 3. The paired flows are constrained by the cross consistency that valid pixel-wise paths by \( \mathcal{W}_f \) and \( \mathcal{W}_b \) form loop closure and bi-directional visual consistency between a training pair \( \mathcal{I}_0 \) and \( \mathcal{I}_T = \{\mathcal{I}_t\}_{t \in \mathcal{T}} \).

**Occlusion Detection** Occluded regions are usually the regions where the bi-directional flows are inconsistent. We define the visibility mask \( \mathcal{M}_{0:t} \) indicating pixels in \( \mathcal{I}_0 \) that are also visible in \( \mathcal{I}_t \), according to the backward-to-forward flow difference \( \Delta \mathcal{W}_b^{t-\tau}(x) = \mathcal{W}_b^t + \mathcal{F}(x, \mathcal{W}_f^\tau | \mathcal{W}_b^t) \), sim-
3D feature volume, where the green color shows the flow generator. (c) The 3D feature volume, where the valid condition for the forward-to-backward flow difference is given by $\|\nabla W^f_{t^{-b}}(x)\|_1 < \max\{\alpha, \beta(\|W^f_{t^{-b}}(x)\|_1 + \|F(x, W^f_{t^{-b}}|W^b_{t^{-b}})\|_1)\}. (2)$

Similarly, we also obtain the visibility mask $M_{t^{-o}}$ about pixels in $I_t$ that are also visible in $I_0$, based on the same condition for the forward-to-backward flow difference $\Delta W^b_{t^{-f}}$. The hyper-parameters are set as $\alpha = 1.0$, $\beta = 0.1$ in our experiments. Note that $M_{t^{-o}}$ corresponds to the target frame $I_t$ and $M_{0^{-t}}$ with the reference frame $I_0$.

**Cycle-consistent Flow Learning** The bi-directional flows are learned to enforce their internal cycle consistency, where the valid (i.e., in non-occluded regions) forward (or backward) flows pointing from $I_0$ (or $I_t$) to $I_t$ (or $I_0$) are mirrored by the backward (forward) flows from the warped locations in $I_t$ (or $I_0$) to the original locations in $I_0$ (or $I_t$).

We define the cycle consistency objective $L_{cc}$ by $\ell_1$ norm as

$$L_{cc} = \sum_{t \in T} \sum_{x} M_{0^{-t}}(x) \cdot \|W^f_{t^{-b}}(x) + F(x, W^f_{t^{-b}}|W^b_{t^{-b}})\|_1 + M_{t^{-o}}(x) \cdot \|W^b_{t^{-f}}(x) + F(x, W^b_{t^{-f}}|W^f_{t^{-f}})\|_1. (3)$$

The learned flows should also be constrained by the bi-directional photometric consistency in the valid regions as

$$L_{bic} = \sum_{t \in T} \sum_{x} M_{0^{-t}}(x) \cdot \|I_0(x) - F(x, W^f_{t^{-b}}|I_t)\|_1 + M_{t^{-o}}(x) \cdot \|I_t(x) - F(x, W^b_{t^{-f}}|I_0)\|_1. (4)$$

The bi-directional flows in the occlusion regions are otherwise naively guessed with a smoothness prior within their neighborhood, as we only apply the non-occluded flows to generate the warped frames, as Eq. (1), while leaving the occluded regions undefined.

**Compositional Condition Fusion** In addition to cycle consistency, the proposed bi-directional flow generator $p_\theta(V^f_T, V^b_T|z, I_0)$ are required to generate content-aware flows that are semantically corresponded to the content structures in $I_0$. It is achieved by introducing a compositional condition fusion scheme into the main branch of bi-directional flow generator, which looks like the Hourglass structure [19] that gradually adapts the sampled motion features with multi-level content features $\{c_m\}_{m=1}^M$ extracted from the image encoder $E_\theta(I_0)$, as illustrated in Fig. 4.

Our flow generator starts from fusing the sampled random variable $z$ with the top-level content feature vector $c_1$, by treating the content features as a depth-wise convolution kernel. The fused motion features are upsampled by a deconvolution layer consisting of an upsampling operation and a 3D convolution layer. 3D convolution layers are employed throughout the main branch of $p_\theta(V^f_T, V^b_T|z, I_0)$ to learn spatiotemporal features and offer more complex motion patterns in the generated flows.

The subsequent network repeats several stacked network blocks composed by a 2D-to-3D fusion block, an aforementioned deconvolution layer and an additional 3D convolution layer, before split into two branches of forward and backward flow subnets, as shown in Fig. 4(a). The proposed 2D-to-3D fusion blocks fuses a content feature map $c_m$ and a corresponding 3D motion feature volume $V_m$, by at first concatenating $c_m$ along the channel axis of each time slice of $V_m$, and then being convoluted by one 3D convolution layer for a seamless feature fusion. It suggests that the motion features in any time stamp should explicitly share the same content features with each other, similar as [26].

**3.4. Occlusion-aware Image Synthesis**

Backward bilinear warping operation $F(x, W^b_{t^{-f}}|I_0)$ inherently suffers from warping artifacts, thus it would not produce visually plausible novel videos. Li et al. [12] applied an image refinement module to remove any warping artifacts. We argue that explicit occlusion handling would be more effective in removing artifacts and inpainting contents in the occluded regions.

Unlike the frame interpolation studies [15, 7] that require at least two frames to infer occluded regions, our method only needs one reference image $I_0$ to simultaneously infer bi-directional flows $\{V^f_T, V^b_T\}$. These flows infer the visibility masks $\{M_{0^{-t}}, M_{t^{-o}}\}_{t \in T}$, according to Eq. (2). Consequently, we propose an occlusion-aware image synthesis module $R_\omega$ that accepts the visibility mask $M_{t^{-o}}$ and the naively warped frame $I_{t^{-o}}$. It outputs a refined novel frame $I_{t^{-o}}$ with the suppression of warping artifacts.

The network for the image synthesis module is similar as those for the inpainting task [14], which applies multi-level skip connections in an autoencoder. The encoder borrows the same architecture of the VGG-19 up to the ReLU$4$, while the decoder is symmetrical to the encoder with the nearest neighbor upsampling operations to replace the max pooling operations. The skip connections link the ReLU$4$, $k = \{1, 2, 3\}$ in the encoder to corresponding lay-
In the training stage, the proposed synthesis module also refines the naïvely warped frame $I_0 \rightarrow I_t$ from the target frame $I_t$ to the reference frame $I_0$ with the help of the visibility mask $M_{0 \rightarrow t}$. We train this network using a perceptual loss [8] for bi-directional image synthesis:

$$L_{pp} = \sum_{t \in T} \left\| \tilde{I}_{t+0} - I_t \right\|^2 + \lambda \sum_{k=1}^5 \left\| \Phi_k(\tilde{I}_{t+0}) - \Phi_k(I_t) \right\|^2 + \left\| \tilde{I}_{0 \rightarrow t} - I_0 \right\|^2 + \lambda \sum_{k=1}^5 \left\| \Phi_k(\tilde{I}_{0 \rightarrow t}) - \Phi_k(I_0) \right\|^2,$$

(5)

where $\Phi_k(\cdot)$ denote the features of a pretrained VGG-19 network at the layer $\text{ReLU}_k$, and $\lambda$ is to balance the terms.

The proposed occlusion-aware image synthesis module is able to fill unreliable regions with semantically meaningful contents, and hallucinate fine details to overcome the blurring artifacts caused by the bilinear warping operation. This model is learned in an unsupervised fashion and can be incorporated with the aforementioned bi-directional flow generator for an end-to-end system for flow generation and image synthesis.

### Training Phase

In the training phase, the motion encoder encodes a stack of adjacent frames $\{I_T, I_0\}$ as a 3D volume and produces mean and variance vectors to model the posterior $q_\theta(z|I_0, I_T)$. The image encoder $E_\theta(I_0)$ extracts multi-level content features $\{c_m\}_{m=1}^M$. The bi-directional flow generator $p_\phi(W_{I_T}^f, W_{I_0}^b; z, I_0)$ uses the sampled motion variables $z$ from $q_\phi(z|I_0, I_T)$ as the input. At the end of the flow generator, we produce the initial bi-directional warped frames $\{I_{t \leftarrow 0}, I_{0 \rightarrow t}\}_{t \in T}$ and their visibility masks $\{M_{t \leftarrow 0}, M_{0 \rightarrow t}\}_{t \in T}$. They are then inputted into the proposed occlusion-aware image synthesis module $R_\omega(\cdot)$ to obtain the final synthesized frames $\{I_{t \leftarrow 0}, I_{0 \rightarrow t}\}_{t \in T}$.

The objective of our ImagineFlow model (Fig. 5) extends the variational upper bound [10] of the CVAE model by adding the aforementioned losses

$$L_{\phi, \psi, \theta, \omega}(I_0, I_T) = -D_{\text{KL}}[q_\phi(z|I_0, I_T)||N(z|0, 1)] + \frac{1}{S} \sum_{s=1}^S \lambda_{\text{bi-vc}} L_{\text{bi-vc}}(z^{(s)}) + \lambda_{cc} L_{cc}(z^{(s)}) + \lambda_{pp} L_{pp}(z^{(s)}),$$

where the bi-directional photometric consistency $L_{\text{bi-vc}}$, cycle flow consistency $L_{cc}$ and the perceptual loss $L_{pp}$ are monte-carlo integrated to serve as the negative log-likelihood for this generation model. The KL-divergence aims at constraining the discrepancy between the posterior $q_\phi(z|I_0, I_T)$ and the naïve motion prior. In addition to the above objective, we add a small amount of TV-$\ell_1$ norm to enhance the smoothness of the final images and flows. $\lambda_{\text{bi-vc}} = 1.0$, $\lambda_{cc} = 0.05$ and $\lambda_{pp} = 1.0$.

### Test Phase

In the testing phase, we just require a plain motion prior $p(z) = N(z|0, I)$ to replace $q_\phi(z|I_0, I_T)$ for the sampling of motion variable $z$. We just employ the backward flows $W_{I_0}^b$ and their visibility masks $\{M_{t \leftarrow 0}\}_{t \in T}$ for the final sequence generation. The novel sequence $\{I_{t \leftarrow 0}\}_{t \in T}$ is generated by inputting $\{I_{t \leftarrow 0}, M_{t \leftarrow 0}\}_{t \in T}$ into the occlusion-aware image synthesis module.
4. Experiments

4.1. Settings

Datasets. The proposed model is trained and evaluated on three popular video datasets: UCF-101 dataset [24], Moving MNIST dataset [25] and Exercises dataset [35]. UCF-101 contains 13,320 real video clips from 101 action classes with substantial background movements. Moving MNIST dataset is a synthetic dataset constructed by warping the digits in MINST dataset [11] with affine transformations. The Exercises dataset includes around 60k pairs of frames from real workout videos with a static background.

Implementation Details. Our system was implemented in PyTorch. It is end-to-end trained by Adam optimizer, with a small learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is 32 and the train/test images are cropped and resized to $64 \times 64$ for the Moving MNIST dataset and $128 \times 128$ for the UCF-101 and Exercise Datasets.

We use two settings to show the superiority of the proposed method. (1) For long-term video generation centered around a single frame, we set the number of rendering frames to 8. (2) To compare with prior arts on long-term and short-term future frame predictions, we modify our network by predicting the next 4 frames and the next 1 frame, accordingly. Moreover, the baseline model copies the main structure but only preserves the backward flow branch and the highest fusion block, and removes the occlusion-aware image synthesis module.

Evaluation Metrics. We also quantitatively evaluate the models in addition to subjective tests. We sample 100 sequences for one reference frame and report the best PSNR and SSIM [33], named as PSNR@100 and SSIM@100. A good model should synthesize at least one sequence that is similar to the ground-truth. We also apply the recent Frechet Inception Distance (FID) based on I3D model to evaluate the perceptual performance of the generated sequences.

4.2. Ablation Study

The unique advantage of the proposed ImagineFlow model is its capability of learning robust flow distributions and synthesizing visually plausible videos. It includes several pivotal components that contribute to this feature.

Motion Representation. We show that flows are more reliable than difference images and pixel values in synthesizing novel frames. To verify it, we add two more baselines by changing the output of our baseline network to RGB pixel values and difference images (named as Flow, Pixel and Diff, respectively). As shown in Fig 6(a), in the exercise dataset, Diff blurs out the face details, but Pixel is often too blur to capture upper torsos and arms. In contrast, Flow preserves the detailed contents and produces fewer visual artifacts around body parts. In addition, in the UCF-101 dataset (Fig. 6(b)), Pixel fails to capture larger motions and Diff produces severe aliasing around the legs of the athletes. Evaluations in Fig. 6(c) also quantitatively prove that the flow-based baseline outperforms the rest counterparts.

Component-wise Comparison. Firstly, we validate the bi-directional flow generator. Cyc.Consist outputs bi-directional flows and applies cycle consistent flow learning, while Comp. Fusion adds skip connections between the
image encoder and flow decoder. As shown in Fig. 7(a), compared to the baseline model, Cyc. Consist encourages more physically consistent flows around the upper body so that there are much fewer visual distortions in synthesizing arms. But the generated flows are blur and cannot capture with the content well. Comp. Fusion applies multi-level content features to regularize the spatial structure of the predicted flows (e.g., small flows in the background and homogeneous flows aligned with the upper body). But this structure does not generate physically reasonable flows, so the synthesized arms suffer from warping artifacts (e.g., the arms are much thinner). Their combination (i.e., the flow generator in the ImagineFlow model) performs the best and the predicted flows are not only physically reliable but also consistent with the captured contents. Apart from the qualitative results, either PSNR @100 or SSIM@100 reports similar performance gains in Fig. 7(c), which suggest that the proposed components are complementary to each other.

Our flow-based frame synthesis faithfully inpaints unreliable regions in the occlusion-aware warped frames. For instance, the violinist in Fig. 8(a) is moving right, thus the occluded white board should be revealed in the target frame. Our flow generator successfully discovers these unreliable regions (see Fig. 8(b)), and our frame synthesis module completes these regions and guesses the structure of the white board as what is desired. As for comparison, we also show the backward warped target frame in Fig. 8(c). Since it renders novel frames solely based on pixels in the reference image, the occluded background cannot be discovered, even though the underlying flows are estimated accurately.

Motion Diversity. The proposed method shows sufficient diversity in generating novel sequences. In Fig. 9, different samples for one reference frame in the Exercise and Moving MNIST datasets demonstrate diverse motion variations. The motion is illustrated by creating a RGB image where the magenta channels are from the sampled frame and the green channel from the reference frame, as suggested in [35]. For examples, Fig. 9(a-1) shows diversified motions around the legs and Fig. 9(a-2) visualizes different squat actions. The Moving MNIST dataset gives long-term motion patterns for two digits. Sample 1 in Fig. 9(b-1) shows contractive and clockwise motions, while sample 2 in Fig. 9(b-2) depicts an ascending digit pair.

Motion Complexity. The proposed ImagineFlow model can capture complex and long-term motions. In Fig. 10(a), we demonstrate sampled 9-frame sequences given the 5th frames as the reference. The sequence “surfing” has varying tidal waves over time nearby the surfing board, and the athlete is blending over. In the sequence “violinist”, our ImagineFlow model covers occluded regions and preserves content structures with meaningful motions in playing violin. The long-term ImagineFlow model samples more complex motion patterns than the short-term one, as shown in Fig. 10(b) and (c). The example in Fig. 10(b) shows that the long-term prediction is capable of guessing the rotation of the female dancer while the iterative short-term prediction will gradually distort and blur the contents. The example in Fig. 10(b) also finds that the iterative variant fails to preserve the spatial structure of the digits.

4.3. Experimental Comparisons

Our ImagineFlow model is compared with various single image based video synthesis methods such as Visual Dynamics [35], Video GAN [28], Video Imagination [1], MoCoGAN [26] and Li et al. [12], in which Visual Dynamics is designed for the next frame prediction but Video GAN, Video Imagination, MoCoGAN and Li et al. can be applied for longer video generation.

Exercise Dataset: As shown in Fig. 12(a), the proposed model and Visual Dynamics are compared by sampling two future frames given a same reference frame. Our results do not only present distinct motions but also preserve struc-
Figure 10. (a) 9-frame sequence generation on the UCF-101 dataset, where the center frame is the reference image. (b) and (c) compare the long-term (1 frame to 4 frames prediction) and the short-term (1-frame to 1-frame iterative prediction) ImagineFlow models in capturing complex motions and preserving structural coherence in the UCF-101 and Moving MNIST datasets. In (c) the color coded motions are illustrated for better visualization. Best viewed on screen.

Figure 11. Visual comparison with the state-of-the-art methods on the (a) Exercises dataset, (b) Moving MNIST dataset.

Figure 12. Visual comparison with the state-of-the-art methods on the UCF-101 dataset. The start frame is marked by yellow cycle.

Table 1. FID score comparison on ice dancing sequence.
5. Conclusion

In this paper, we propose a novel probabilistic framework that samples video sequences from a single image. Our model improves the CVAE framework with a bi-directional flow generator and a compositional fusion structure, thus is able to learn content-aware and structural coherent flow distributions. The involved flow-based frame synthesis module renders high-quality novel sequences and solves the rendering artifacts inherently in the warping-based operation. We have shown that the proposed model performs well on both on synthetic and real-world videos.

Appendix: Detailed Network Architecture

The complete network consists of (a) a motion encoder $q_{\phi}(z|I_0, I_T)$, (b) a bi-directional flow generator $p_{\theta}(W_{f}^{k}, W_{\theta}^{k})|z, I_0)$, (c) an image encoder $E_{\theta}(I_0)$ and (d) an occlusion-aware synthesis module $R_{omega}(\cdot)$. In the following subsections, we will specify the network architectures for each component.

Note that batch normalization is applied in the entire network. Leaky ReLU with $\lambda = 0.2$ is the activation function used in each convolution layer. But linear activation function is applied in the output layers. Input images are normalized in the range $[0, 1]$. We only illustrate the network for the input size of $128 \times 128$ and sequence length of 8.

Motion Encoder. The motion encoder uses an image volume of size $N \times 9 \times 128 \times 128 \times 3$ as its input. For a sample in one batch, it stacks an image sequence $I_{T}$ and its reference frame $I_0$ along the time dimension. $N$ is the batch size and $|I_{T}| = 8$. The output has two branches, one indicates the means $\mu$ and the other is $\log \sigma$, as the parameters for the Gaussian posterior $q_{\phi}(z|I_0, I_{T})$. This network is shown in Tab. 2.

Image Encoder. The image encoder takes the reference image $I_0$ as the input, and outputs intermediate content features $\{c_{n}\}_{n=0}^{5}$ through six consecutive 2D convolution layers. The highest level of the content features is $c_1$, which is a $N \times 1 \times 1 \times 1 \times 1024$ tensor. The network specification is shown in Tab. 3.

Bi-directional Flow Generator. The bi-directional flow generator starts from sampling the motion variables $z$ either by the reparameterization trick as $z = \mu + \sigma \circ \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$, or the standard Gaussian distribution $z \sim \mathcal{N}(0, I)$.

The sampled motion tensor of size $N \times 1 \times 1 \times 1 \times 1024$ is then inputted into the flow generator. The first fusion is similar as the depthwise convolution with the convolution kernel as the content feature $c_1$. The subsequent fusions are operated as: at first concatenation of the motion features and the content features along the time dimension, and then a 3D convolution layer to fuse across all dimensions. We use the nearest neighboring upsampling with a 3D convolution layer to deconvolute the preceding motion features. The network outputs are flow volumes of size $N \times 8 \times 128 \times 128 \times 4$, i.e., the number of bi-directional flows is 8. The forward and backward flows are split along the last dimension.

The network architecture of the flow generator is shown in Tab. 4, its main branch actually mirrored the structure of the motion encoder.

Occlusion-aware Image Synthesis. The structure of the occlusion-aware frame synthesis is briefly depicted in the main article. It uses the network layers up to conv4_1 of VGG-19 as the encoder. Its decoder mirrors the encoder with nearest neighbor upsampling to replace the max pooling operation. The skip connections link encoding layers conv$k, k = 1, 2, 3$ to their corresponding decoding layers. The input of this network is a stacked tensor about the warped frame and its visibility map, along the channel dimension.

References

[1] B. Chen, W. Wang, and J. Wang. Video imagination from a single image with transformation generation. In ACM MM, pages 358–366. ACM, 2017. 2, 7
[2] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In ICCV, pages 2758–2766, 2015. 2
[3] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In NIPS, pages 64–72, 2016. 2
[4] D. A. Forsyth and J. Ponce. Computer vision: a modern approach. Prentice Hall Professional Technical Reference, 2002. 3
[5] Y. Ganin, D. Kononenko, D. Sutskever, and Y. Lempitsky. Deepwarps: Photorealistic image resynthesis for gaze manipulation. In ECCV, pages 311–326. Springer, 2016. 2
[6] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In NIPS, pages 2017–2025, 2015. 3
[7] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In CVPR, IEEE, 2018. 4
[8] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In ECCV, pages 694–711. Springer, 2016. 5
[9] N. Kalchbrenner, A. v. d. Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. Video pixel networks. arXiv preprint arXiv:1610.00527, 2016. 2
[10] D. P. Kingma and M. Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013. 5
[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998. 6
| name   | type        | num filters | filter size | stride | padding | activation |
|--------|-------------|-------------|-------------|--------|---------|------------|
| conv1  | conv3d      | 64          | 3 × 3 × 3    | 1 × 1  | 1 × 1 × 1| lrelu      |
| pool1  | maxpool3d   | -           | -           | 1 × 2  | -       | -          |
| conv2  | conv3d      | 64          | 3 × 3 × 3    | 1 × 1  | 1 × 1 × 1| lrelu      |
| pool2  | maxpool3d   | -           | -           | 1 × 2  | -       | -          |
| conv3  | conv3d      | 128         | 3 × 3 × 3    | 1 × 1  | 1 × 1 × 1| lrelu      |
| pool3  | maxpool3d   | -           | -           | 1 × 2  | -       | -          |
| conv4  | conv3d      | 256         | 3 × 3 × 3    | 1 × 1  | 1 × 1 × 1| lrelu      |
| pool4  | maxpool3d   | -           | -           | 2 × 2  | -       | -          |
| conv5  | conv3d      | 512         | 3 × 3 × 3    | 1 × 1  | 1 × 1 × 1| lrelu      |
| pool5  | maxpool3d   | -           | -           | 2 × 2  | -       | -          |

| µ      | conv3d      | 1024        | 2 × 4 × 4    | 1 × 1  | 0 × 0 × 0| linear     |
| log σ  | fc          | 1024        | 2 × 4 × 4    | 1 × 1  | 0 × 0 × 0| linear     |

Table 2. Network specification of the motion encoder.

| name   | type        | num filters | filter size | stride | padding | activation |
|--------|-------------|-------------|-------------|--------|---------|------------|
| c5     | conv2d      | 64          | 4 × 4       | 2 × 2  | 1 × 1   | lrelu      |
| c4     | conv2d      | 64          | 4 × 4       | 2 × 2  | 1 × 1   | lrelu      |
| c3     | conv2d      | 128         | 4 × 4       | 2 × 2  | 1 × 1   | lrelu      |
| c2     | conv2d      | 256         | 4 × 4       | 2 × 2  | 1 × 1   | lrelu      |
| c1     | conv2d      | 512         | 4 × 4       | 2 × 2  | 1 × 1   | lrelu      |
| c0     | conv2d      | 1024        | 4 × 4       | 1 × 1  | 0 × 0   | lrelu      |

Table 3. Network specification of the image encoder.

[12] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Flow-grounded spatial-temporal video prediction from still images. In ECCV, 2018. 1, 2, 3, 4, 7, 8
[13] X. Liang, L. Lee, W. Dai, and E. P. Xing. Dual motion GAN for future-flow embedded video prediction. In ICCV, 2017. 2
[14] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In ECCV, 2018. 4
[15] Z. Liu, R. Yeh, X. Tang, Y. Liu, and A. Agarwala. Video frame synthesis using deep voxel flow. In ICCV, volume 2, 2017. 2, 4
[16] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. arXiv preprint arXiv:1701.01821, 2, 2017. 2
[17] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440, 2015. 2
[18] S. Meister, J. Hur, and S. Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In AAAI, New Orleans, Louisiana, Feb. 2018. 4
[19] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In ECCV, pages 483–499. Springer, 2016. 4
[20] V. Patrascu, A. Handa, and R. Cipolla. Spatio-temporal video autoencoder with differentiable memory. arXiv preprint arXiv:1511.06309, 2015. 2
[21] S. L. Pintea, J. C. van Gemert, and A. W. M. Smeulders. Dejavu: Motion prediction in static images. In ECCV. Springer, 2014. 2
[22] M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In IEEE International Conference on Computer Vision (ICCV), volume 2, page 5, 2017. 2
[23] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In NIPS, pages 3483–3491, 2015. 5
[24] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012. 6
[25] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using LSTMs. In ICML, pages 843–852, 2015. 2, 6
[26] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. In CVPR, 2018. 1, 2, 7, 8
[27] C. Vondr̆k, H. Pirsiavash, and A. Torralba. Anticipating visual representations from unlabeled video. In CVPR, pages 98–106, 2016. 2
[28] C. Vondr̆k, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In NIPS, pages 613–621, 2016. 1, 2, 7
[29] C. Vondr̆k and A. Torralba. Generating the future with adversarial transformers. In CVPR, 2017. 2
Table 4. Network specification of the bi-directional flow generator.

| name   | type       | num filters | filter size | stride | padding | activation |
|--------|------------|-------------|-------------|--------|---------|------------|
| $V_0$  | Sampling of $\{\mu, \sigma\}$ or naïve Gaussian prior |             |             |        |         |            |
| fusion0| Depthwise convolution of $V_0$ by the kernel from $c_1$ |             |             |        |         |            |
| $V_1$  | deconv3d   | 512         | $2 \times 4 \times 4$ | $1 \times 1 \times 1$ | $0 \times 0 \times 0$ | lrelu      |
| fusion1| concatenate $V_1$ with $c_1$ along the channel for each time slice |             |             |        |         |            |
| upconv1| conv3d     | 512         | $3 \times 3 \times 3$ | $1 \times 1 \times 1$ | $1 \times 1 \times 1$ | lrelu      |
|        | upsample   | -           |             |        | $2 \times 2 \times 2$ | -          |
| $V_2$  | conv3d     | 256         | $3 \times 3 \times 3$ | $1 \times 1 \times 1$ | $1 \times 1 \times 1$ | lrelu      |
| fusion2| concatenate $V_2$ with $c_1$ along the channel for each time slice |             |             |        |         |            |
| upconv2| conv3d     | 256         | $3 \times 3 \times 3$ | $1 \times 1 \times 1$ | $1 \times 1 \times 1$ | lrelu      |
|        | upsample   | -           |             |        | $2 \times 2 \times 2$ | -          |
| $V_3$  | conv3d     | 128         | $3 \times 3 \times 3$ | $1 \times 1 \times 1$ | $1 \times 1 \times 1$ | lrelu      |
| fusion3| concatenate $V_3$ with $c_1$ along the channel for each time slice |             |             |        |         |            |
| upconv3| conv3d     | 128         | $3 \times 3 \times 3$ | $1 \times 1 \times 1$ | $1 \times 1 \times 1$ | lrelu      |
|        | upsample   | -           |             |        | $2 \times 2 \times 2$ | -          |
| $V_4$  | conv3d     | 64          | $3 \times 3 \times 3$ | $1 \times 1 \times 1$ | $1 \times 1 \times 1$ | lrelu      |
| fusion4| concatenate $V_4$ with $c_1$ along the channel for each time slice |             |             |        |         |            |
| upconv4| conv3d     | 64          | $3 \times 3 \times 3$ | $1 \times 1 \times 1$ | $1 \times 1 \times 1$ | lrelu      |
|        | upsample   | -           |             |        | $2 \times 2 \times 2$ | -          |
| $V_5$  | conv3d     | 64          | $3 \times 3 \times 3$ | $1 \times 1 \times 1$ | $1 \times 1 \times 1$ | lrelu      |
| fusion5| concatenate $V_5$ with $c_1$ along the channel for each time slice |             |             |        |         |            |
| upconv5| conv3d     | 64          | $3 \times 3 \times 3$ | $1 \times 1 \times 1$ | $1 \times 1 \times 1$ | lrelu      |
| output | conv3d     | 4           | $3 \times 3 \times 3$ | $1 \times 1 \times 1$ | $1 \times 1 \times 1$ | linear     |