Dual Motion GAN for Future-Flow Embedded Video Prediction

Xiaodan Liang, Lisa Lee
Carnegie Mellon University
{xiaodan1,lslee}@cs.cmu.edu

Wei Dai, Eric P. Xing
Petuum Inc.
{wei.dai,eric.xing}@petuum.com

Abstract

Future frame prediction in videos is a promising avenue for unsupervised video representation learning. Video frames are naturally generated by the inherent pixel flows from preceding frames based on the appearance and motion dynamics in the video. However, existing methods focus on directly hallucinating pixel values, resulting in blurry predictions. In this paper, we develop a dual motion Generative Adversarial Net (GAN) architecture, which learns to explicitly enforce future-frame predictions to be consistent with the pixel-wise flows in the video through a dual-learning mechanism. The primal future-frame prediction and dual future-flow prediction form a closed loop, generating informative feedback signals to each other for better video prediction. To make both synthesized future frames and flows indistinguishable from reality, a dual adversarial training method is proposed to ensure that the future-flow prediction is able to help infer realistic future-frames, while the future-frame prediction in turn leads to realistic optical flows. Our dual motion GAN also handles natural motion uncertainty in different pixel locations with a new probabilistic motion encoder, which is based on variational autoencoders. Extensive experiments demonstrate that the proposed dual motion GAN significantly outperforms state-of-the-art approaches on synthesizing new video frames and predicting future flows. Our model generalizes well across diverse visual scenes and shows superiority in unsupervised video representation learning.

1. Introduction

Despite the great progress of deep learning architectures for supervised learning, unsupervised video representation learning for general and scalable visual tasks remains a largely unsolved yet critical research problem. Recently, predicting future frames [22, 20, 28] in a video sequence has surged as a promising direction for unsupervised learning of video data. The key insight behind this is that the model is forced to learn powerful video representations in preceding frames in order to predict how the appearances and motions of video frames will change over time.

Video frame prediction itself is a challenging task due to the complex appearance and motion dynamics of natural scenes. Intuitively, in order to predict realistic pixel values in future frames, the model must be capable of capturing pixel-wise appearance and motion changes so as to let pixel values in previous frames flow into new frames. However, most existing state-of-the-art approaches [20, 28, 18, 16, 26, 37] use generative neural networks to directly synthesize RGB pixel values of future video frames and do not explicitly model the inherent pixel-wise motion trajectories, leading to blurry predictions. Although several recent attempts [23, 16, 26] have tried to alleviate this issue by designing a motion field layer that copies pixels from previous frames, the predictions suffer from notable artifacts due to imprecise intermediate flows.

In this work, we develop a dual motion Generative Adversarial Network (GAN) architecture that learns to explicitly make the synthesized pixel values in future frames coherent with pixel-wise motion trajectories using a dual adversarial learning mechanism. Specifically, it simultaneously resolves the primal future-frame prediction and dual future-flow prediction based on a shared probabilistic motion encoder. Inspired by the success of GANs [6, 13], we establish a dual adversarial training mechanism between two future-frame and future-flow generators, and two frame and flow discriminators, to make the predictions indistinguishable from real data. The underlying dual-learning mechanism bridges the communication between future pixel hallucination and flow prediction by mutually reviewing each other. Our dual motion GAN consists of three fully-differentiable modules as follows.

- A probabilistic motion encoder captures motion uncertainty that may appear in different locations, and produces latent motion representations for preceding frames which are then fed as inputs to two generators.
- The future-frame generator then predicts future frames, which are assessed from two aspects: frame fidelity by a frame discriminator, and flow fidelity by passing the estimated flows between the preceding frame and the predicted frame into a flow discriminator.

- The future-flow generator in turn predicts future flows, which are also assessed from two aspects: flow fidelity by a flow discriminator, and frame fidelity by passing the extrapolated future frame (computed by a nested flow-warping layer) into a frame discriminator.

By learning over symmetric feedback signals from two dual adversarial discriminators, the future-frame generator and future-flow generator mutually benefit from each other's complementary targets, leading to better video prediction. Our dual motion GAN outperforms all existing approaches on synthesizing next frames and long-term future frames of natural scenes after training on car-mounted camera videos from the KITTI dataset [5] and consumer videos from the UCF-101 dataset [27]. We also demonstrate its generalization capability by testing on a different car-cam Caltech dataset [3] and a collection of raw dash-cam videos from YouTube. In addition, we demonstrate the critical design choices of each module through extensive ablation studies. Further experiments on flow estimation, flow prediction, and action classification show our model's superiority on unsupervised video representation learning.

2. Related Work

The proposed dual motion GAN attempts to jointly resolve the future-frame and future-flow prediction problems with a unified architecture. Our review thus focuses on two lines of literature most relevant to our problem.

Motion Prediction. Various methods have been investigated to predict a future motion field [15, 33, 32, 17, 19] and visual representation [28, 30] given an image or a video sequence. Optical flow is the most commonly explored motion field, though large and fast motions can pose problems. Beyond deterministic motion prediction [15, 17], a more recent work [32] proposed using a variational autoencoder as a probabilistic prediction framework to handle intrinsic motion ambiguities. In contrast to prior works that only aim to generate optical flows, our dual motion GAN treats future-flow prediction as a dual task of future-frame prediction, and reviews the flow predictions by a frame discriminator using an dual adversarial learning mechanism. In addition, we introduce a novel probabilistic motion encoder that captures pixel-wise motion uncertainties to model long-term motion dynamics for boosting flow prediction.

Video Frame Prediction. Many experiments with synthesizing video frames have been conducted recently [20, 18, 16, 26, 37, 23]. A line of research [20, 18, 37, 31] focuses on developing advanced networks to directly generate pixel values. However, they often produce blurry predictions since it is hard to model the complex pixel-level distributions of natural images. Several approaches [23, 16, 26] alleviate this blurring problem by resorting to motion field prediction for copying pixels from previous frames. More recently, Sedaghat et al. [26] explores a hybrid multi-task framework to jointly optimize optical flow estimation and frame prediction. However, these models still suffer from notable artifacts due to imprecise intermediate flows and unrealistic frame predictions. In contrast, our dual motion GAN learns to mutually optimize the future-frame generator and future-flow generator. It effectively alleviates the problem of frame deviations accumulating over time by incorporating pixel-wise motion trajectory prediction. Moreover, the dual frame and flow discriminators help drag the distributions of generated frames and flows closer to the real data distribution. Our model is thus able to produce sharp future frames and reasonable future flows simultaneously for a wide range of videos.

3. Dual Motion GAN

We propose the dual motion GAN, a fully differentiable network architecture for video prediction that jointly solves the primal future-frame prediction and dual future-flow prediction. The dual motion GAN architecture is shown in Figure 1. Formally, our dual motion GAN takes a video sequence $\mathbf{v} = \{I_1, \dots, I_t\}$ as input to predict the next frame $\hat{I}_{t+1}$ by fusing the future-frame prediction $\ddot{I}_{t+1}$ and future-flow based prediction $\bar{I}_{t+1}$. We adopt the simple 1×1 convolution filters for the fusing operation. The dual motion generator (shown in Figure 2) consists of five components: a probabilistic motion encoder E, future-frame generator G_I, future-flow generator G_F, flow estimator $Q_{I \to F}$ and flow-warping layer $Q_{F \to I}$. The dual motion discriminator (shown in Figure 3) consists of a frame discriminator D_I and a flow discriminator D_F. More specifically, the probabilistic motion encoder E first maps previous frames to a latent code z. The future-frame generator G_I and future-flow estimator G_F then decode z to predict the future frame $\ddot{I}_{t+1}$ and future flow $\bar{F}_{t+1}$, respectively. The fidelity of $\ddot{I}_{t+1}$ is judged by how well $\ddot{I}_{t+1}$ fools the frame discriminator D_I, and how well the flow $\bar{F}_{t+1}$ between I_t and $\ddot{I}_{t+1}$, estimated using the flow estimator $Q_{I \to F}$, fools the flow discriminator D_F. Similarly, the quality of future-flow prediction is judged by how well $\bar{F}_{t+1}$ fools the flow discriminator D_F, and how well the warped frame $\bar{I}_{t+1}$, generated by warping I_t with $\bar{F}_{t+1}$ using the flow-warping layer $Q_{F \to I}$, fools the frame discriminator D_I.

3.1. Adversarial Dual Objective

In this section, we formally derive the training objective of our dual motion GAN.

VAE: The encoder-generator triplet $\{E, G_I, G_F\}$ constitutes a variational autoencoder (VAE). The probabilistic motion encoder E first maps a video sequence $\mathbf{v}$ into a code z in the latent space $\mathcal{Z}$, and G_I, G_F then decode a randomly perturbed version of z to predict future frames and flows, respectively. Following [12], we assume the components in the latent space $\mathcal{Z}$ are conditionally independent and Gaus-

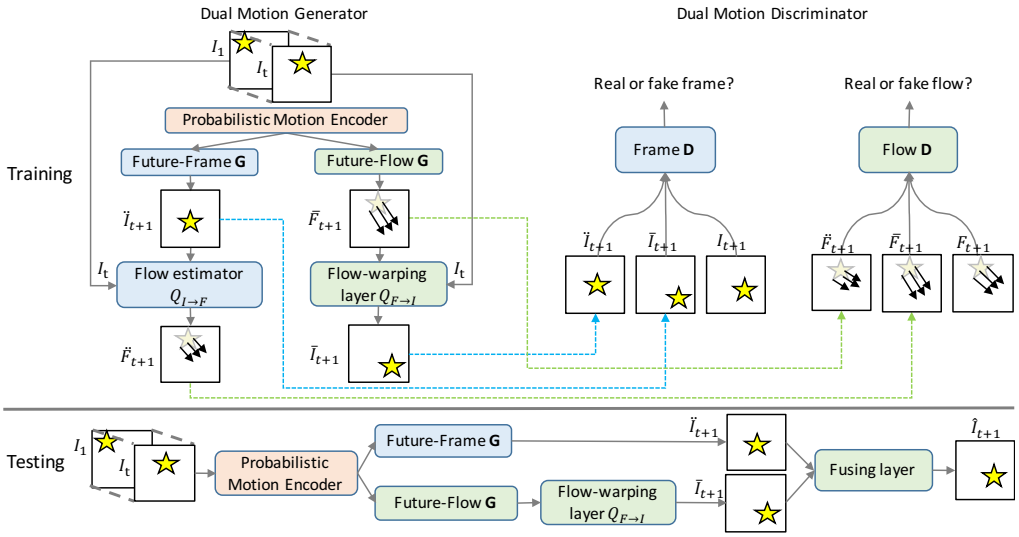

Figure 1. The proposed dual motion GAN jointly solves the future-frame prediction and future-flow prediction tasks with a dual adversarial learning mechanism. A video sequence $I_1, \ldots, I_t$ is first fed into a probabilistic motion encoder E to obtain a latent representation z. The dual motion generators ("Future-frame G" and "Future-flow G" on the left) decode z to synthesize future frames and flows. The dual motion discriminators ("Frame D" and "Flow D" on the right) learn to classify between real and synthesized frames or flows, respectively. The flow estimator $Q_{I\to F}$ takes the predicted frame $\ddot{I}_{t+1}$ and real frame I_t to estimate the flow $\ddot{F}_{t+1}$, which is further judged by "Flow D". The flow-warping layer $Q_{F\to I}$ warps the real frame I_t with the predicted flow $\bar{F}_{t+1}$ to generate the warped frame $\bar{I}_{t+1}$, which is then evaluated by "Frame D". The testing stage is shown in the bottom row.

sian. The encoder E outputs the mean maps $E_\mu(\mathbf{v})$ and the variance maps $E_{\sigma^2}(\mathbf{v})$, where the distribution of the latent code z is given by $q(z|\mathbf{v}) = \mathcal{N}(z|E_\mu(\mathbf{v}), E_{\sigma^2}(\mathbf{v}))$. The architecture of E is detailed in Section 3.2. The frame prediction is obtained as $\ddot{I}_{t+1} = G_I(z \sim q(z|\mathbf{v}))$, and the corresponding estimated flow is $\ddot{F}_{t+1} = Q_{I\to F}(\ddot{I}_{t+1}, \mathbf{v})$. The flow prediction is calculated as $\bar{F}_{t+1} = G_F(z \sim q(z|\mathbf{v}))$, and the corresponding warped frame is $\bar{I}_{t+1} = Q_{F\to I}(\bar{F}_{t+1}, \mathbf{v})$. Note that the flow-warping layer $Q_{F\to I}$ does not have parameters to be optimized. We train the VAE by minimizing a variational upper bound of a negative log-likelihood function:

$$\begin{aligned}\mathcal{L}_{\text{VAE}}(E, G_I, G_F, Q_{I\to F}) = \mathbb{E}_{z\sim q(z|\mathbf{v})}(\\ -\log p_{G_I}(I_{t+1}|z) - \log p_{Q_{I\to F}}(F_{t+1}|G_I(z))\\ -\log p_{G_F}(F_{t+1}|z) - \log p_{Q_{F\to I}}(I_{t+1}|G_F(z))\\ +\mathbf{KL}(q(z|\mathbf{v})||p(z)),\end{aligned} \tag{1}$$

where KL is the Kullback-Leibler divergence that penalizes deviation of the distribution of the latent code from the prior distribution $p(z) = \mathcal{N}(z|0, I)$. L1 distance [18, 16] is imposed on the future-frame prediction $\ddot{I}_{t+1}$ and warped frame prediction $\bar{I}_{t+1}$. We thus model the conditional distribution $p_{G_I}(I_{t+1}|z) = \exp(-||I_{t+1} - G_I(z)||_1)$ and $p_{Q_{F\to I}}(I_{t+1}|G_F(z)) = \exp(-||I_{t+1} - Q_{F\to I}(G_F(z), \mathbf{v})||_1)$. Following common practice in flow estimation [24], we adopt the average End Point Error (EPE) Δ_{EPE} to optimize the future-flow prediction and flow estimation. We thus compute two conditional distributions of flows as $p_{G_F}(F_{t+1}|z) = \exp(-\Delta_{\text{EPE}}(F_{t+1}, G_F(z)))$ and $p_{Q_{I\to F}}(F_{t+1}|G_I(z)) = \exp(-\Delta_{\text{EPE}}(F_{t+1}, Q_{I\to F}(G_I(z), \mathbf{v})))$. Hence, minimizing the negative log-likelihood term is equivalent to minimizing L1 distance between the predicted frame and the true frame, and the EPE loss between the predicted and the true flow.

Adversarial Dual Objective: The generators G_I, G_F and the discriminators D_I, D_F form two dual generative adversarial networks, and enables the dual motion GAN to generate sharper and more realistic frame and flow predictions. As discussed in the new Wasserstein GAN (WGAN) [1], the original GAN [6] suffers from several training difficulties such as mode collapse and instable convergence. We thus follow the proposed training strategies in Wasserstein GAN (WGAN) [1] that theoretically remedy these problems by minimizing an approximated Wasserstein distance. The dual motion GAN can be trained by jointly solving the learning problems of the VAE and two dual GANs:

$$\begin{aligned}\min_{E,G_I,G_F,Q_{I\to F}} \max_{D_I,D_F} \mathcal{L}_{\text{VAE}}(E, G_I, G_F, Q_{I\to F})\\ +\lambda\mathcal{L}^I_{\text{GAN}}(G_I, G_F, D_I)\\ +\lambda\mathcal{L}^F_{\text{GAN}}(G_F, G_I, D_F, Q_{I\to F}).\end{aligned} \tag{2}$$

where λ balances the VAE loss and two dual GAN losses. The generators G_I, G_F and flow estimator $Q_{I\to F}$ try to minimize the whole objective against the adversarial discriminators D_I and D_F that try to maximize it. The discriminator D_I learns to distinguish real frames I_{t+1} from the predicted frames $\ddot{I}_{t+1}$ and warped frames $\bar{I}_{t+1}$. Similarly, D_F learns to distinguish real flows F_{t+1} from pre-

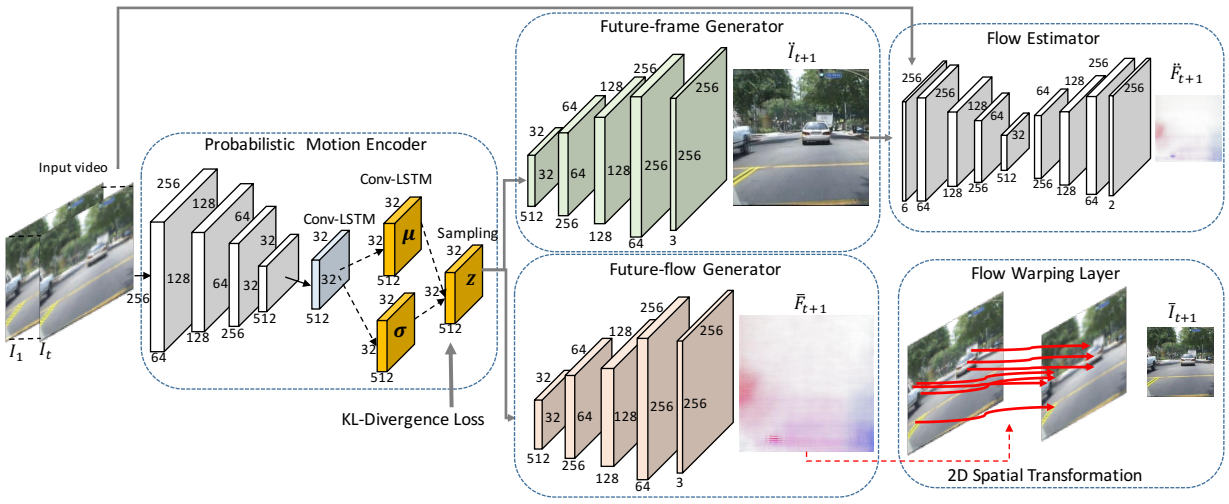

Figure 2. The dual motion generator. Each frame in a given sequence is recurrently fed into the probabilistic motion encoder E, which includes four convolutional layers, one intermediate ConvLSTM layer, and two ConvLSTM layers that produce the mean maps and variance maps for sampling z. Next, the future-frame generator G_I and future-flow generator G_F decode z to produce a future-frame $\ddot{I}_{t+1}$ and future flow $\bar{F}_{t+1}$, respectively. The flow estimator $Q_{I\to F}$ then generates the estimated flow $\ddot{F}_{t+1}$ between I_t and $\ddot{I}_{t+1}$. The flow-warping layer $Q_{F\to I}$, which performs differential 2D spatial transformation, warps I_t into $\bar{I}_{t+1}$ according to $\bar{F}_{t+1}$.

dicted flows $\bar{F}_{t+1}$ and estimated flows $\ddot{F}_{t+1}$. Let $p(\mathbf{v})$ and $p(\mathcal{F})$ denote the true data distributions of true frames and true flows. The dual GAN objective functions are given by:

$$\begin{aligned}\mathcal{L}^I_{\text{GAN}}(G_I,G_F,D_I) &= \mathbb{E}_{I_{t+1}\sim p(\mathbf{v})} D_I(I_{t+1}) \\ &- \frac{1}{2}\mathbb{E}_{z\sim q(z|\mathbf{v})} D_I(G_I(z)) \\ &- \frac{1}{2}\mathbb{E}_{z\sim q(z|\mathbf{v})} D_I(Q_{F\to I}(G_F(z))), \\ \mathcal{L}^F_{\text{GAN}}(G_F,G_I,D_F,Q_{I\to F}) &= \mathbb{E}_{F_{t+1}\sim p(\mathcal{F})} D_F(F_{t+1}) \\ &- \frac{1}{2}\mathbb{E}_{z\sim q(z|\mathbf{v})} D_F(G_F(z)) \\ &- \frac{1}{2}\mathbb{E}_{z\sim q(z|\mathbf{v})} D_F(Q_{I\to F}(G_I(z))).\end{aligned} \tag{3}$$

Our adversarial dual objective functions differ from the standard GAN objective function in that the samples for each discriminator come from two different distributions depicted by two dual generators. For $\mathcal{L}^I_{\text{GAN}}$, two synthesized distributions are p_{G_I} and $p_{Q_{F\to I}}$ functioning on the distribution p_{G_F}. Optimizing $\mathcal{L}^I_{\text{GAN}}$ encourages both G_I and $Q_{F\to I}$ to output frames resembling true frames from $p(\mathbf{v})$, which can further serve as feedback signals to the distribution p_{G_F}. Similarly, optimizing $\mathcal{L}^F_{\text{GAN}}$ encourages both G_F and G_I to output flows resembling true flows from $p(\mathcal{F})$. During training, the dual motion GAN uses true frames I_{t+1} in the video to supervise future-frame prediction, and the optical flows F_{t+1} estimated by EpicFlow [25] to supervise future-flow prediction. We choose the traditional EpicFlow since it does not require annotated flows for training, yet achieves the best results on flow estimation.

Learning: Inheriting from GAN, the optimization of dual motion GAN can be seen as a two-player game—the first player consisting of an encoder and two generators, and the second player consisting of two adversarial discriminators. The first player's objective is to defeat the second player and also to minimize the VAE losses. Following WGAN [1], we apply an alternating gradient update scheme, performing five gradient descent steps on D_I and D_F, and one step on $G_I, G_F, Q_{I\to F}$. We use minibatch SGD and apply the RMSprop solver [29]. The λ is empirically set to 0.001, and the learning rate is set to 0.0001. We train the model for roughly 40 epochs. In order to have parameters of D_I and D_F lie in a compact space, we clamp the weights to a fixed box [0.01, 0.01] after each gradient update. We apply batch normalization [10] and set the batch size to 1, which has been termed "instance normalization" and demonstrated to be effective in image generation tasks.

3.2. Network Architectures

The detailed networks for generators and discriminators are provided in Figure 2 and Figure 3, respectively. For simplicity, the pooling layers, batch normalization layers, and ReLU layers after the intermediate convolutional layers are omitted in the figures.

Probabilistic Motion Encoder: The exact motions of objects in real-world videos are often unpredictable and have large variations due to intrinsic ambiguities. Existing works [37, 32, 12, 8] often learn a whole latent vector z for all objects in the scene. A shortcoming of these models is that they cannot distinguish the particular motion pattern for each pixel location of distinct objects. We thus extend the variational autoencoder to generate a spatial joint distribution conditioned on the input frames.

Formally, to accommodate our dual motion GAN for an

arbitrary number of input frames, we design a recurrent probabilistic motion encoder E (Figure 2) to learn variational motion representations z that encode past motion patterns and also model the uncertainty in motion fields. As presented in Section 3.1 (VAE), the encoder E generates the mean maps $E_\mu(\mathbf{v})$ and the variance maps $E_{\sigma^2}(\mathbf{v})$. Specifically, each frame I_t in $\mathbf{v}$ is recurrently passed into four convolutional layers to obtain a compact 32×32 feature map with 512 dimensions. Next, one Convolution LSTM (ConvLSTM) layer [9, 36, 14] is employed for sequence modeling where the memory cells essentially act as an accumulator of the spatial state information. The resulting features map is further fed into two convolution LSTM layers to predict the mean maps $E_\mu(\mathbf{v})$ and variance maps $E_{\sigma^2}(\mathbf{v})$, respectively. Compared to conventional convolution layers, ConvLSTM determines the future state of a certain cell in the grid by incorporating the inputs and past states of its local neighbors. We use a 4×4 kernel and 512 hidden states for all ConvLSTM layers. Finally, the latent motion representation z is sampled from $\mathcal{N}(z|E_\mu(\mathbf{v}), E_{\sigma^2}(\mathbf{v}))$.

Dual Motion Generator: The future-frame generator G_I decodes the shared motion representation z to produce a future-frame prediction $\ddot{I}_{t+1}$ with RGB channels. Similarly, future-flow generator G_F decodes z to produce a future-flow prediction $\bar{F}_{t+1}$ with two channels that represent the horizontal and vertical components of the optical flow field. Both generators use five deconvolutional layers with 3×3 kernels. The flow estimator $Q_{I \to F}$, which consists of four convolutional layers and four deconvolutional layers, estimates flow maps $\ddot{F}_{t+1}$ between the previous frame I_t and the future-frame prediction $\ddot{I}_{t+1}$. The flow-warping layer $Q_{F \to I}$ is a warp operator that generates $\bar{I}_{t+1}$ by warping the previous frame I_t according to the predicted flow field $\bar{F}_{t+1}$ using bilinear interpolation. We follow the differential spatial transformation layer used in [24, 11] to define this warping operator.

Dual Motion Discriminator: As shown in Figure 3, to optimize the adversarial dual objective in Eqn. (2), the frame discriminator D_I takes a $3 \times 256 \times 256$ image as input and produces one output value, while the flow discriminator D_F takes a $2 \times 256 \times 256$ flow map from a real data or generated distribution. Following the Wasserstein GAN [1], we drop the Sigmoid operation in the final prediction in order to remedy the mode collapse problem in vanilla GANs [6].

4. Experiments

In this section, we first present the main comparisons on video prediction tasks, including next frame and multiple frame predictions. We then present ablation studies on our model. In addition, we demonstrate the generalization capabilities of our model through extensive experiments on flow prediction, flow estimation, and unsupervised representation learning.

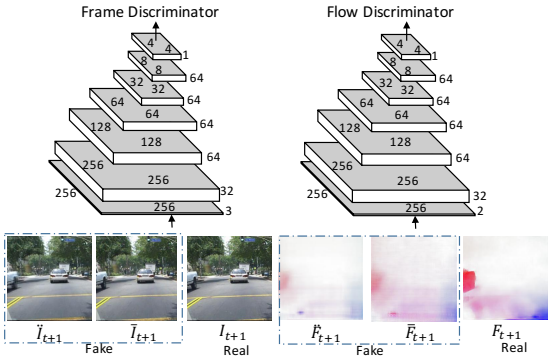

Figure 3. Architectures of the two dual motion discriminators. The frame and flow discriminators learn to classify between real and synthesized frames and flows, respectively.

4.1. Comparisons on Video Prediction

Experimental Settings. We evaluate the video prediction capabilities of our model on complex real-world sequences. First, we experiment on car-mounted camera videos, since these videos span a wide range of settings and are characterized by rich temporal dynamics, including both the self-motion of the vehicle and the motion of other objects in the scene. Following the state-of-the-art PredNet [18], models are trained using raw videos from the KITTI dataset [5] and evaluated on the test partition of the Caltech Pedestrian dataset [3]. Following PredNet's [18] procedure, we sample sequences of 10 frames from the City, Residential, and Road categories, with 57 recording sessions used for training and 4 used for validation. In total, the training set consisted of roughly 41K frames. In order to further validate our model's generalization capability, we evaluate the trained model on 500 raw 1-minute clips from YouTube, collected using the keywords "dashboard videos". Second, following [20] and [16], we train models on the generic consumer videos from UCF101 [27], and evaluate on the UCF-101 [27] and THUMOS-15 [7] test sets. To compare with current state-of-the-art [16] models, we also use two previous frames as the input to predict the next future frame.

We used the metrics MSE [18], PSNR, and SSIM [35] to evaluate the image quality of video frame prediction, where higher values of PSNR and SSIM indicate better results. The implementations are based on the public Torch7 platform on a single NVIDIA GeForce GTX 1080. The details of our optimization procedure and hyperparameters are presented in Section 3. Our dual motion GAN takes around 300ms to predict one future frame and flow given a sequence of 10 previous frames (as in the Caltech test set).

Comparison on Caltech and YouTube Clips. Table 1 reports the quantitative comparison with the state-of-the-art models BeyondMSE [20] and Prednet [18] on the video next-frame prediction task. We obtain the results of BeyondMSE [20] by training a model that minimizes the loss

Table 1. Performance (MSE and SSIM) of video frame prediction on Caltech and YouTube clips after training on KITTI dataset.

	Caltech		YouTube Clip	
Method	MSE	SSIM	MSE	SSIM
CopyLast	0.00795	0.762	0.01521	0.785
BeyondMSE [20]	0.00326	0.881	0.00853	0.820
PredNet [18]	0.00313	0.884	0.00679	0.858
Ours frame w/o GAN	0.00307	0.880	0.00833	0.826
Ours frame GAN	0.00291	0.883	0.00793	0.836
Ours flow w/o GAN	0.00292	0.884	0.00778	0.839
Ours flow GAN	0.00289	0.887	0.00701	0.843
Ours frame+flow w/o GAN	0.00269	0.892	0.00617	0.859
Ours w/o motion encoder	0.00262	0.895	0.00583	0.863
Ours future-flow (testing)	0.00255	0.896	0.00601	0.866
Ours future-frame (testing)	0.00260	0.893	0.00613	0.862
Ours (full)	**0.00241**	**0.899**	**0.00558**	**0.870**

Table 2. Performance (PSNR and SSIM) of video frame prediction on UCF-101 and THUMOS-15.

	UCF-101		THUMOS-15	
Method	PSNR	SSIM	PSNR	SSIM
BeyondMSE [20]	28.2	0.89	27.8	0.87
EpicFlow [25]	29.1	0.91	28.6	0.89
DVF [16]	29.6	0.92	29.3	0.91
Nextflow [26]	29.9	-	-	-
Ours (full)	**30.5**	**0.94**	**30.1**	**0.92**

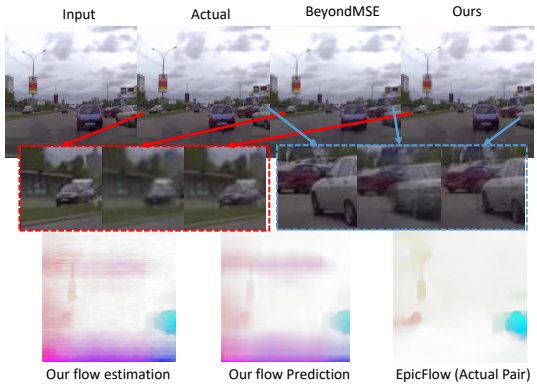

Figure 4. Qualitative results on a YouTube clip. We highlight the predicted regions of two vehicles approaching in opposite directions in red and blue boxes for better comparison.

functions in BeyondMSE [20] (ADV+GDL), and replaces the backbone network with our frame generator, except for the motion autoencoder. Our model significantly outperforms both baselines, achieving a MSE of 2.41×10^3 and SSIM of 0.899, compared to 3.13×10^3 and 0.884 of Prednet [18], and 3.26×10^3 and 0.881 of BeyondMSE [20].

We show qualitative comparisons on the Caltech Pedestrian dataset and YouTube clips in Figure 5 and Figure 4, respectively. Our model makes fairly accurate and high-resolution predictions in a wide range of scenarios. For instance, in Figure 4, the model is able to predict the motions of two vehicles and their shadows as they approach from different directions, as well as handle the stationary vehicle. We also show the future-flow prediction and the estimated flow between the input frame and predicted frame. Our model gives reasonable and comparable future flows with that of Epicflow [25] estimated between the actual frame pair.

Comparison on UCF-101 and THUMOS-15. Table 2 shows the comparisons of four state-of-the-art methods on UCF-101 [27] and THUMOS-15 [7]. We directly compare the results reported in DVF [16]. BeyondMSE [20] directly hallucinates pixel values while EpicFlow [25], DVF [16], and Nextflow [26] extrapolate future frames by predicting intermediate flows. Our dual motion GAN, which combines the merits of both frame-based and flow-based models via a dual-learning mechanism, achieves the best performance.

Multiple frame prediction. In multiple frame prediction, the models' predictions are used as input to the network for predicting subsequent frames. As discussed in [20, 18], model-based video prediction methods tend to break down fairly quickly when extrapolating long-term frames, as deviations of the predictions unavoidably accumulate over time. We report the performance comparison with BeyondMSE [20] in Figure 6 and qualitative results of our model in Figure 7. Our model again shows better performance after five time steps. Despite some blurriness, our model still captures some key structures in its extrapolations after the fifth time step, benefiting from the long-term memorization capabilities of the recurrent motion encoder.

4.2. Ablation studies

We report comparisons on our model variants in Table 1.

Future-frame Generator. "Ours frame w/o GAN" only includes the probabilistic motion encoder and future-frame generator, and thus the final prediction is $\ddot{I}_{t+1}$. Benefiting from the motion uncertainty modeling by the probabilistic motion encoder, "Ours frame w/o GAN" with L1 loss achieves better results than BeyondMSE [20] with L1, GDL, and ADV losses. Adding the frame discriminator and jointly optimizing the adversarial loss gives us another variant "Ours frame GAN", which generates sharper and more realistic predictions, and slightly outperforms PredNet [18].

Future-flow Generator. The future-flow generator learns to predict optical flows $\bar{F}_{t+1}$, which are then passed through a warping layer to get the final frame prediction $\bar{I}_{t+1}$. "Ours flow w/o GAN" and "Ours flow GAN" both obtain better prediction performance than "Ours frame w/o GAN" and "Ours frame GAN", which speaks to the superiority of learning motion fields that are more consistent with the natural characteristics of the videos.

Adversarial Dual Objective. We verify the advantages of combining the merits of both the future-frame generator and the future-flow generator. "Ours frame+flow w/o GAN"

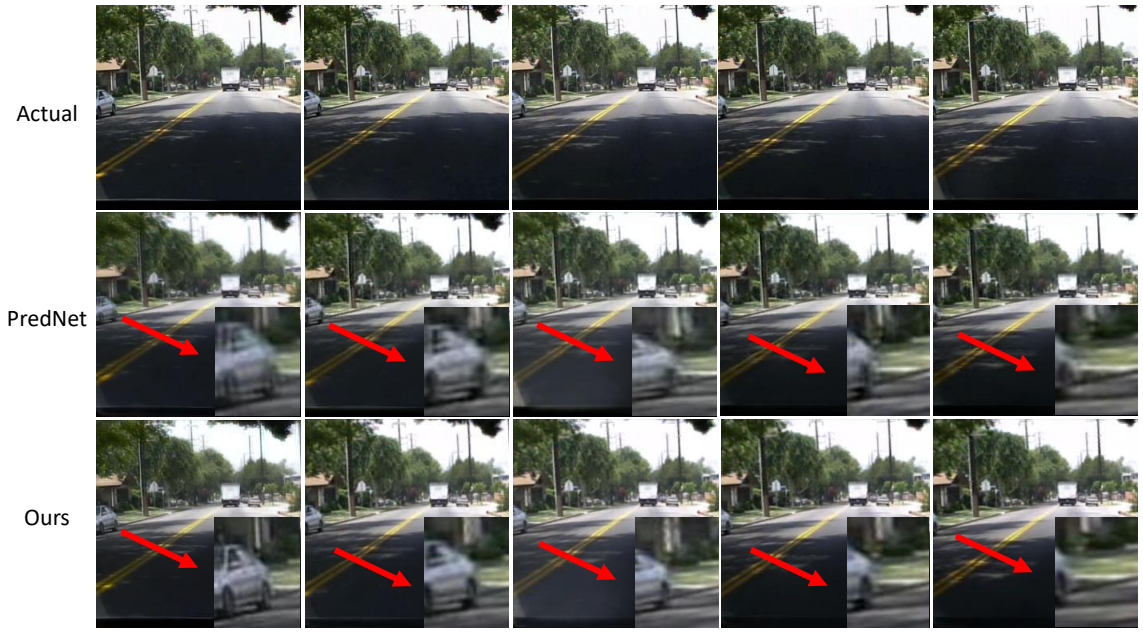

Figure 5. Qualitative comparisons with Prednet [18] for next-frame prediction on car-cam videos from the Caltech dataset.

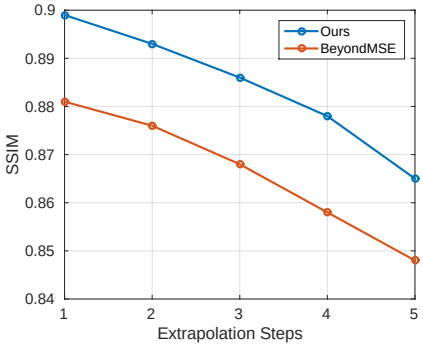

Figure 6. Performance comparison for multiple-frame prediction on the Caltech dataset.

obtains remarkably lower MSE and higher SSIM over the single objective-based models "Ours flow GAN" and "Our frame GAN". Our full model has the best performance since the adversarial discriminators help judge the fidelity of both frame and flow predictions in a dual manner.

Probabilistic Motion Encoder. We also evaluate the effect of eliminating the motion encoder ("Ours w/o motion encoder"). The significant performance drop compared to our full model can be observed, particularly on the YouTube clips. A possible reason is that the raw YouTube videos have more diverse motion patterns. The motion uncertainty of each object at different spatial locations can be effectively captured by the inferred probabilistic motion maps.

Performances of Two Generators During Testing. The aforementioned studies use differently trained models for each setting. We also experiment on how the predictions from two generators differ from each other during testing. "Ours future-flow (testing)" and "Ours future-frame (testing)" both show significantly better results than "Ours frame GAN" and "Ours flow GAN" due to the trained dual model that can mutually improve the frame and flow predictions. We find that fusing the two predictions from the two generators obtains the best results, as shown in Figure 8.

4.3. Flow Prediction and Estimation

Although we have already verified the effectiveness of flow prediction on improving future-frame prediction, we further quantitatively evaluate the "by-product" flow predictions and flow estimation performance. We compare our models with state-of-the-art models [2, 25, 4]. Following DVF [16], we train on the UCF-101 dataset and evaluate on the KITTI flow 2012 dataset [5]. The future-flow prediction module generates the predicted flows of test frames given previous frames, while the flow estimation module takes the true previous frame and test frame as inputs to estimate intermediate flows. Table 3 reports the average endpoint error (EPE) over all the labeled pixels. Our dual motion GAN only uses the flows predicted by EpicFlow [25] as the supervision for training the flow estimator and future-flow prediction module, which is thus an unsupervised method. Both the performances of our flow prediction and flow estimation are competitive with existing methods. "Ours (flow GAN)", in which future-frame prediction is eliminated during training, is inferior to our full model "Ours (flow prediction)", but is better than the prior flow-based method DVF [16]. Our model is capable of encoding essential motion information, benefiting from the joint optimization of the primal video frame prediction and dual flow prediction tasks. In addition, forecasting future flows is more challenging than flow estimation given two true frames. We provide visualization results by flow prediction and estimation in Figure 8.

Figure 7. Multiple-frame prediction results of our model on Caltech sequences for five time steps.

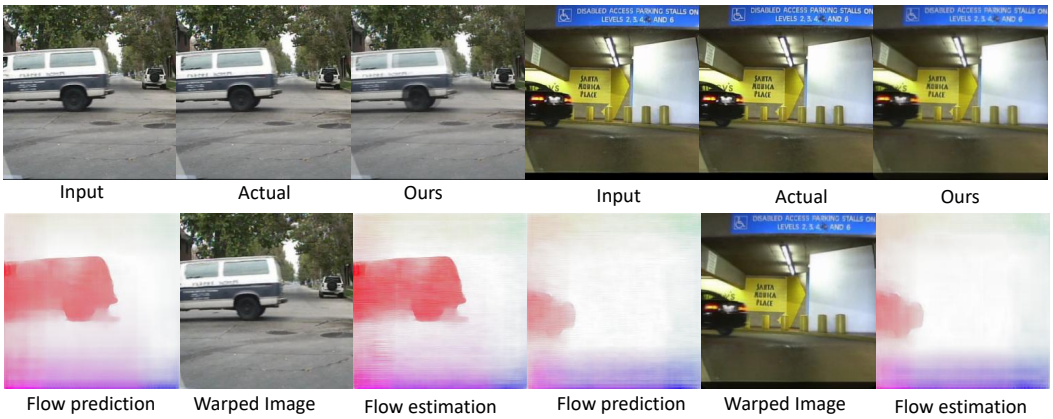

Figure 8. Some example future-frame prediction and future-flow prediction results of our model on two sequences from the KITTI dataset.

Table 3. Endpoint error of flow estimation and prediction on the KITTI dataset. Here, lower values indicate better performance.

Method	EPE
Flownet [4] (supervised)	9.1
EpicFlow [25] (unsupervised)	3.8
DVF [16] (unsupervised)	9.5
Ours (flow GAN) (unsupervised)	9.3
Ours (flow prediction) (unsupervised)	**8.9**
Ours (flow estimation) (unsupervised)	**7.6**

4.4. Unsupervised Representation Learning

To show the effectiveness of our model on unsupervised video representation learning, we replace the future-frame and future-flow generators with one fully-connected layer and one softmax loss layer appended to the probabilistic motion encoder. Our model is then fine-tuned and tested with an action recognition loss on the UCF-101 dataset (split-1), following [20, 16]. This is equivalent to treating the future-frame and future-flow prediction tasks as pre-training. As demonstrated in Table 4, our model outperforms random initialization by a large margin and also shows superior performance compared to other approaches.

5. Conclusion and Future Work

We proposed a dual motion GAN that simultaneously solves the primal future-frame prediction and future-flow prediction tasks via a dual adversarial training mechanism. The probabilistic motion encoder learns to capture spatial motion uncertainty, while the dual adversarial discriminators and generators send feedback signals to each other to generate realistic flows and frames that are implicitly coherent with each other. Extensive experiments on video frame prediction, flow prediction, and unsupervised video representation learning demonstrate the contributions of our model to motion encoding and predictive learning. As future work, we plan to explicitly model the multi-agent dependencies so as to be able to handle real-world videos with complex motion interactions.

Table 4. Classification accuracy of action recognition on UCF-101.

Method	Accuracy
Unsupervised Video [34]	43.8
Shuffle&Learn [21]	50.2
DVF [16]	52.4
Ours	**55.1**

Acknowledgement

This work is funded by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3, 4, 5

[2] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *TPAMI*, 33(3):500–513, 2011. 7

[3] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, pages 304–311, 2009. 2, 5

[4] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 7, 8

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2, 5, 7

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 1, 3, 5

[7] A. Gorban, H. Idrees, Y. Jiang, A. R. Zamir, I. Laptev, M. Shah, and R. Sukthankar. Thumos challenge: Action recognition with a large number of classes. In *CVPR workshop*, 2015. 5, 6

[8] P. Goyal, Z. Hu, X. Liang, C. Wang, and E. Xing. Nonparametric variational auto-encoders for hierarchical representation learning. *arXiv preprint arXiv:1703.07027*, 2017. 4

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

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *JMLR*, 2015. 4

[11] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015. 5

[12] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2, 4

[13] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing. Recurrent topic-transition gan for visual paragraph generation. *arXiv preprint arXiv:1703.07022*, 2017. 1

[14] X. Liang, L. Lin, X. Shen, J. Feng, S. Yan, and E. P. Xing. Interpretable structure-evolving lstm. *arXiv preprint arXiv:1703.03055*, 2017. 5

[15] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *TPAMI*, 33(5):978–994, 2011. 2

[16] Z. Liu, R. Yeh, X. Tang, Y. Liu, and A. Agarwala. Video frame synthesis using deep voxel flow. In *arXiv preprint arXiv:1702.02463*, 2017. 1, 2, 3, 5, 6, 7, 8

[17] G. Long, L. Kneip, J. M. Alvarez, H. Li, X. Zhang, and Q. Yu. Learning image matching by simply watching video. In *ECCV*, pages 434–450. Springer, 2016. 2

[18] W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *ICLR*, 2017. 1, 2, 3, 5, 6, 7

[19] Z. Luo, B. Peng, D. Huang, A. Alahi, and L. Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *CVPR*, 2017. 2

[20] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016. 1, 2, 5, 6, 8

[21] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, pages 527–544, 2016. 8

[22] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh. Action-conditional video prediction using deep networks in atari games. In *NIPS*, pages 2863–2871, 2015. 1

[23] V. Patraucean, A. Handa, and R. Cipolla. Spatio-temporal video autoencoder with differentiable memory. *arXiv preprint arXiv:1511.06309*, 2015. 1, 2

[24] A. Ranjan and M. J. Black. Optical flow estimation using a spatial pyramid network. *arXiv preprint arXiv:1611.00850*, 2016. 3, 5

[25] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *CVPR*, pages 1164–1172, 2015. 4, 6, 7, 8

[26] N. Sedaghat. Next-flow: Hybrid multi-tasking with next-frame prediction to boost optical-flow estimation in the wild. *arXiv preprint arXiv:1612.03777*, 2016. 1, 2, 6

[27] 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. 2, 5, 6

[28] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, pages 843–852, 2015. 1, 2

[29] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012. 4

[30] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, pages 98–106, 2016. 2

[31] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *NIPS*, pages 613–621, 2016. 2

[32] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*, pages 835–851, 2016. 2, 4

[33] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *CVPR*, pages 3302–3309, 2014. 2

[34] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, pages 2794–2802, 2015. 8

[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[36] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, pages 802–810, 2015. 5

[37] T. Xue, J. Wu, K. Bouman, and B. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NIPS*, pages 91–99, 2016. 1, 2, 4