Conditional Retrospective Cycle GAN for Video Predictions

Mayuresh Bhosale
Clemson University
Greenville, USA
mbhosal@clemson.edu

Prakhar Gupta
Clemson University
Greenville, USA
prakhag@clemson.edu

Abstract

Video frame prediction is crucial in the field of computer vision, with applications such as autonomous vehicles, robotics, and video refinement. Existing research approaches struggle with maintaining longer-length video predictions considering motions which ultimately result in pixel hallucinations or blurry video frame. This is often due to combined errors from improper motion modeling. In this paper, we present a novel framework for predicting video frames that leverages kinematic motion constraints and optical flow to incorporate motion adherence. The framework provides reliable predictions even in intricate, non-linear motion situations. Various tests on benchmark datasets show notable gains in visual consistency and prediction accuracy with addition of kinematic constraints when compared with the baseline models. The proposed framework bridges the gap in between the physics informed prediction the learned representations

1. Introduction

Video frame prediction has an important role to predict the dynamically changing systems such as Autonomous vehicles for safety enhancement, which is addressed by some advanced machine learning algorithms such as RNNs, GANs, Diffusion models, etc. Although the accuracy to predict the video frames has improved, they suffer from pixel hallucinations and accurate long-term video predictions. These are majorly since the predictions do not leverage the kinematics information that is extant in the real-world such as vehicle and pedestrian motion.

Main contributions: We build upon the ideas from Retrospective Cycle GAN (RCGAN) [3]. They established great performance compared to the SOTA with their forward and backward temporal consistency idea for training the generator.

1. Optical Flow-based constraints: Preserves the spatio-temporal motion consistency that utilizes pixel-wise fine grain tracking. We used pretrained model RAFT for predicting the optical flow.
2. Kinematic Velocity and Acceleration based constraints: Embedded kinematics equation to restrict the pixel wise motion in video frame prediction via velocity and acceleration. Executed this via additional loss term in the generator loss function.

By evaluating the two approaches to the baseline RCGAN was outperformed by Optical Flow based RCGAN. Moreover, Kinematic-RCGAN performed superior than all other approaches with better quantitative and qualitative results. More detailed results and discussion can be found at project website: conditional rcgan

2. Relevant Work

Vehicle and pedestrian motions are guided by the physical laws and to accurately predict the videos, its important to include physical a biological property in videos which is in general challenging. This research work introduces a new architecture based on the encoder decoder with convolutional LSTM to predict the long-term video sequence aka PhyLoNet [7] which encompasses the physics laws to represent the realistic motion behaviors. The paper is an addition to the previous research PhyDNet where in PhyLoNet now, has network architectural differences with an encoder decoder structure with a new concept of a relative flow loss. Where PhyLoNet claims to have better results in long term video predictions with a small input video sequence.
Another work proposes an end-to-end unsupervised learning video prediction network model called as Generative Differential-Assisted Discriminative Network, GDDNet. The motion-based information in the video is learned through the attention mechanisms and the variations in the videos are captured by the difference generation method. This research mainly focuses on enhancing the detailed representation in the long-term video prediction learned from the short-term motion cells. Although this method shows efficiently predicts the video sequence, there is room for improvement in blurriness of the frames.
In [4] a novel Dual GAN based architecture is used to pre-

dict the next frames reducing the blurriness with dual objective by pixel wise flow refinement and frame-based refinement. The architecture encompasses of, a Probabilistic Encoder, two GANs one as a Flow Generator GAN followed by a Frame wrapping layer and another as a Frame Generator GAN followed by a Flow Estimator, and two adversarial discriminators, i.e. a frame Discriminator and a Flow Discriminator. The dual nature of predicting the flow and fake enhances the network's ability to reduce the pixel hallucinations and blurriness. Similarly, [2] proposes dual objective based on the encoder decoder framework that handles the dynamics as well as enhances the spatial correlations. However, the both of these research works struggles predicting the large video sequences with details where [2] additionally requires heavy tuning of of motion cell hyper parameters. PredNet [5] on the other hand is inspired by predictive coding theories from neuroscience. It is designed to learn temporal structure in video frames, enabling the prediction of future frames by decoupling the representation of each frame into prediction and error layers. Each layer in the network makes localized predictions, passing forward only deviations. The research showed better generalization for learned representations like steering angle prediction. But, their predictions under-perform when in complex and rapidly changing scenes.

The most related work on which our work builds is [3] Retrospective Cycle GAN (RCGAN), of which network architecture aims to predict the accurate temporal sequences of the video frames with a GAN architecture. A single generator predicts the sequences in forward and reverse cycles allowing the generator to double verify its consistency. The architecture encompasses two discriminators, one as a frame discriminator that differentiates in between real and fake individual frames and another as a sequence discriminator that determines real and fake sequences of video frames generated. The generator loss function consists of image reconstruction loss, Laplacian of gaussian to prevent the edges in the image, and adversarial loss. This allows the entire model architecture to predict very long sequences of the video accurately. Additionally, we draw our inspiration from the optical flow model RAFT [6] that estimates deep optical flow in the video frame sequences.

3. Approach

We develop 3 models over the baseline RCGAN architecture to improve blurring in the later frames. In an attempt to condition the GAN on physics directly or indirectly we employ two methods. In the first method, we compute optical flow for the frame sequences with RAFT [6] which is shown to perform well on KITTI dataset [1] that we use here. The difference in the optical flow is used to condition the generations. The loss function is as in (1).

$$l_1 = L_{image} + \lambda_1 \cdot L_{LoG} + \lambda_2 \cdot L^{Adv}_{frame} + \lambda_3 \cdot L^{Adv}_{sequence} + \lambda_4 \cdot L_{flow} \quad (1)$$

Where, L_{image} is the image reconstruction loss, $\cdot L_{LoG}$ is the Laplacian of Gaussian loss that helps preserve the edge details and the rest are adversarial losses with respective weights λ_1 to be tuned. The condition added here is $\cdot L_{flow}$ which is a optical flow loss constraint. In the second approach, velocity and acceleration for each pixel is computed during training and is compared to the ground truth sequence. The velocity and acceleration losses for each pixel are given as in 2. The resulting loss function is given in (4).

$$L_{vel} = \|v_{predicted} - v_{real}\| \quad (2)$$

$$L_{acc} = \|a_{predicted} - a_{real}\| \quad (3)$$

$$l_2 = L_{image} + \lambda_1 \cdot L_{LoG} + \lambda_2 \cdot L^{Adv}_{frame} + \lambda_3 \cdot L^{Adv}_{sequence} + \lambda_5 \cdot L_{vel} + \lambda_6 \cdot L_{acc} \quad (4)$$

And in the third approach, we include both kinematics and flow loss in the loss function (5).

$$l_3 = L_{image} + \lambda_1 \cdot L_{LoG} + \lambda_2 \cdot L^{Adv}_{frame} + \lambda_3 \cdot L^{Adv}_{sequence} + \lambda_5 \cdot L_{vel} + \lambda_6 \cdot L_{acc} + \lambda_4 \cdot L_{flow} \quad (5)$$

Where, $\cdot L_{vel}$ and $\cdot L_{acc}$ are kinematic losses added to as a constraint. With the network architecture from RCGAN, we predict the 5 frame output video frames from 5 frame input sequences to the model.

4. Experimental Results and Discussion

We train and evaluate over five input and predicted output sequences across baseline RCGAN model and 3 other novel conditioned RCGAN models. To run the simulation on limited memory access we reduce original sequence prediction to five consecutive frames. Additionally, we Cropped and downsized images to 128x128 from KITTI Raw dataset [1] for city driving with normalization that leads to improved learning performance. Small image sequences (18) and batch sizes were used to train the model to accommodate 16GB memory of V100 GPU.

We evaluated video frame predictions for 4 models:

- Baseline Retrospective Cycle GAN-RCGAN
- Approach 1: Kinematics Conditioned RCGAN
- Approach 2: Optical Flow Conditioned RCGAN

- Approach 3: Optical Flow + Kinematics Conditioned RCGAN

For the evaluation purposes across all the qualitative results aare show in the format of Fig. 1 (baseline RCGAN model evaluation), where the upper row of images denotes predicted 5 frame sequences and the second row of images denotes corresponding ground truth video frames. As observed, the transition from 4th to 5th frame prediction by the baseline model is blurred marginally high.

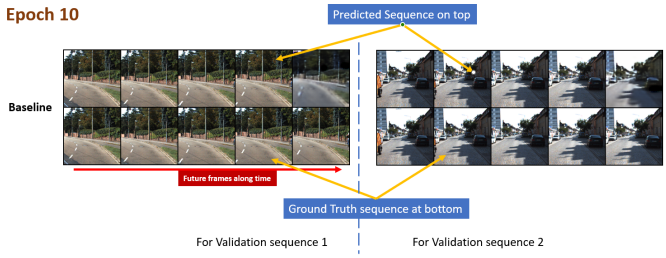

Figure 1. Baseline model at 10 epochs

Epoch No.	Model	PSNR	SSIM	MSE
10	Baseline	22.09	0.719	6.29
	Approach-1	22.95	0.763	5.26
	Approach-2	23.22	0.764	4.83
	Approach-3	22.66	0.737	5.46
100	Baseline	23.76	0.790	4.9
	Approach-1	26.20	0.850	2.48
	Approach-2	25.43	0.839	3.10
	Approach-3	25.93	0.838	2.647

Table 1. Quantitative: Baseline, Flow Conditioned RCGAN, and Kinematics RCGAN model at 10 and 100 epochs

From the Fig. 2, at epoch 10 we can already observe the bluriness in the conditioned GAN approaches is lesser than the Baseline RCGAN model. Where, the Kinematic constrained RCGAN predictions are visibly similar to the Flow based RCGAN. All of the models struggle to predict accurately transitioning from 4th to 5th image prediction with blurriness added.

Similarly when we look at Table 1, For epoch 10, all three conditioned RCGAN models outperforms the baseline RCGAN (PSNR 22.09, SSIM 0.719, MSE 6.29) suggesting its effectiveness. At epoch 10, the performance of Kinematic RCGAN (PSNR 23.22, SSIM 0.764, MSE 4.83) and Flow RCGAN (PSNR 22.95, SSIM 0.763, MSE 5.26) performs very similar with kinematic RCGAN having a small edge over Flow RCGAN. When both of the Kinematic and Flow RCGAN models are combined in Approach 3, the performance decreases suggesting the conditions are working against each other when combined or more training data is required.

This also suggests that more number of epochs are needed to train the models to better the performance. As observed from training after 100 epochs, Conditional RCGANs showed less blurry video frames Fig. 3. Additionally, we also evaluated the model combing the optical flow and kinematics RCGAN and found out that the qualitative and quantitative performance was not better than the Kinematic conditioned RCGAN (PSNR 26.20,SSIM 0.85, MSE 2.48).

Detailed observation from Fig. 4 suggests marginal improvement in the details from Kinematic conditioning. The road markings and vehicle edges are identified clearly even in the last frame of prediction sequence suggests that motion constraints enhances the accuracy.

5. Conclusion and Future Work

We developed a conditional GAN model to restrict pixel movements to realistic ones using optical flow or kinematic velocity constraints. This helps reducing blurring over the baseline model. One main limitation is evaluating model over long term prediction. In future, we intend to apply constraints to the objects tracked to make the predictions more accurate, especially in long term sequence

References

[1] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2

[2] Huilin Huang and YePeng Guan. Danet: A spatio-temporal dynamics and detail aware network for video prediction. *Neurocomputing*, page 128023, 2024. 2

[3] Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1811–1820, 2019. 1, 2

[4] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *proceedings of the IEEE international conference on computer vision*, pages 1744–1752, 2017. 1

[5] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016. 2

[6] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[7] Cuong Than, Derek Ruths, and Luay Nakhleh. Phylonet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC bioinformatics*, 9:1–16, 2008. 1

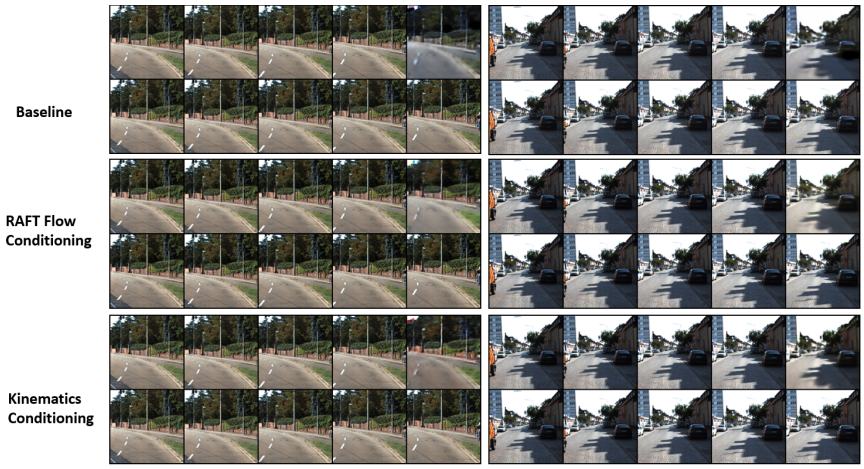

Figure 2. Qualitative: Baseline, Flow Conditioned RCGAN, and Kinematics RCGAN model at 10 epochs

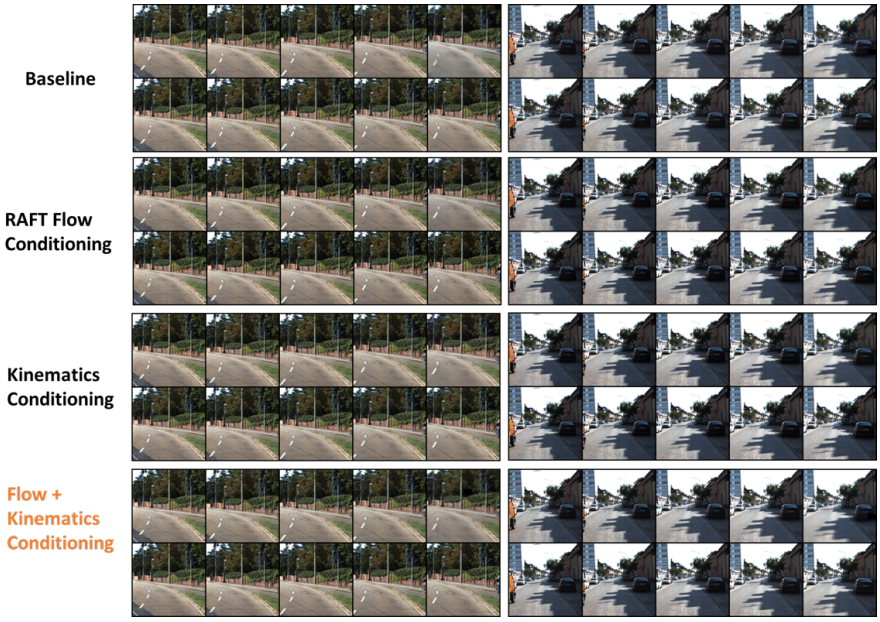

Figure 3. Qualitative: Baseline, Flow Conditioned RCGAN, and Kinematics RCGAN model at 100 epochs

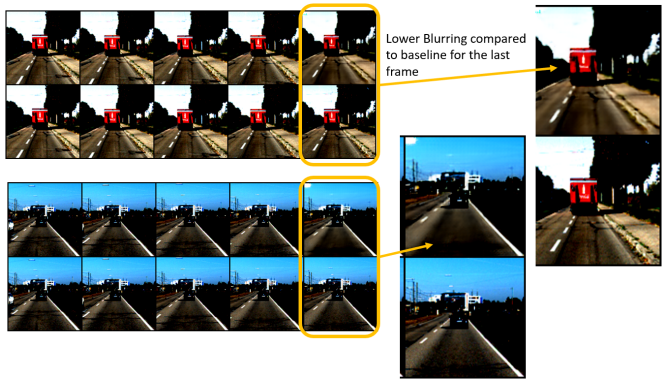

Figure 4. Qualitative: Kinematics RCGAN model