Computer Vision | Homework
Research Paper Review

Ruchi Manikrao Dhore - W1652116

Monday 20th February, 2023

1 Introduction

As part of this homework, we would study and review the paper *"Optimizing Video Prediction via Video Frame Interpolation"*. The authors of the research paper are *Yue Wu Qiang Wen Qifeng Chen*, The Hong Kong University of Science and Technology, published by *arXiv* for the proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17814-17823 in the year 2022. We will provide an overview of the research area, the main research question, and the significance of the research. The key findings and contributions of the research papers will be reviewed.

2 Rationale For Research

Many fields, such as robotics planning, autonomous driving, and video manipulation, can benefit from video prediction. Predicted films, for instance, can assist autonomous robots in better planning future actions given future visual data. *The majority of video prediction approaches require extra data, such as depth maps, semantic or instance maps, knowledge of the scene, and external training, which limits the applicability of their framework to a limited set of films.* For instance, due to the vastly different motion of a person moving dataset, it is challenging to deploy a video prediction model trained on a driving scene. These techniques perform worse and can't even be used in various videos in the real world when this extra information is lacking or of poor quality. **A technique that makes the framework applicable to any video was required.**

3 Major Objective Of Research

To propose new **optimization framework** for video prediction via video frame interpolation **to solve an extrapolation problem** based on an interpolation model.

4 Literature Survey

4.1 Group I | Video Prediction

Despite the use of deep voxel flow, as suggested by DVF, to synthesize future frames, there were difficulties connected with video prediction. Explicit modeling, limitations, and scene assumptions are required because video prediction in real-world scenarios is highly complicated and ambiguous. To enforce consistency and divide scenes into background and foreground identities, a number of techniques have been devised, including the use of depth maps, semantic maps, and semantic and instance segmentation. Video prediction was majorly researched by *Ziwei Liu 2017, Qi 2019, Gao 2019, Yue Wu 2020, Bei 2021 and Lee 2021.*

4.1.1 Strength

The main strength was video prediction utilizing extrapolated video frames.

4.1.2 Limitation

Due to the requirement for extra data, such as semantic or instance maps, depth maps, prior knowledge of the scene, or external training, their approach can only be applied to particular videos. For example, the application of a model trained on robot scenarios to driving scenes is challenging due to the significant differences in motion between these two domains.

4.1.3 Future Scope

The future scope is creating a framework applicable to any video.

4.2 Group II | Video Frame Interpolation

The difficulties of video prediction problems, is what drived the development of this approach. The goal of video frame interpolation is to interpolate interim frames in between subsequent input frames. It was majorly researched by *Wenbo Bao 2019, Wenbo Bao 2021, Simone Meyer 2018, Simone Meyer 2015, Simon Niklaus 2017, Zhewei Huang 2020 and Huaizu Jiang 2018.*

4.2.1 Strength

The use of video frame interpolation for video prediction was the key strength. These techniques can offer excellent interpolation performance even with complex motion and don't require any more data, such semantic maps and depth maps.

4.2.2 Future Scope

Nowadays, optimization-based approaches on test data are still competitive since learning-based methods suffer from the domain gap between training data and test data.

4.3 Group III | Optimization-based Methods

Optimization-based Methods was majorly researched by *Gatys 2016, Shaham 2019, Lei 2020 and Mildenhall 2020.* Gatys proposed the first method for neural style transfer. Shaham optimized a generative model on a single image, producing high-quality and diverse samples. Lei optimized a network to improve temporal consistency, while Mildenhall optimized a fully-connected network for novel view synthesis using a sparse set of input views.

4.3.1 Strength

The main strength was models of network optimization for fully connected systems.

4.3.2 Future Scope

The future scope is to suggest a system that not only addresses the domain gap issue but also gives users an instantaneous control during optimization.

5 Author's Approach

This paper presented a new optimization framework by using video frame interpolation (VFI) for video prediction.

5.1 Advantages of approach

- Due to the lack of requirement of semantic or instance maps, background knowledge of the situation, or outside training, the framework is very adaptable
- The technique can be used to anticipate videos in any scenario and at any resolution
- The method achieves exceptional performance on several datasets and surpasses modern video prediction techniques that call for more data
- With simply RGB frames as input, the approach significantly outperforms external learning techniques

6 Key Steps from Author's Approach

The fundamental stages of the author's methodology.

1. Formulation of problem and construction of E as objective function that measures image similarity
2. Construction of flow initialization equation to calculate rough flows
3. To propose Video Frame Interpolation
4. Identification and masking of occlusion areas

7 Mathematics Used

Let x_t represent the video frame at step t in time. The input to the framework includes two recent RGB frames x_{t-1} and x_t.

Goal is to predict the future frames:

$$\{\tilde{x}_{t+1}, \tilde{x}_{t+2}, \ldots\}$$

They used an interpolation network for videos that is designated as G. An objective function that gauges picture similarity is called constructed E. Here, the relationship between

$$x_{t-1}, \tilde{x}_{t+1}, \text{and } x_t$$

is constrained using a video frame interpolation network G.

$$\tilde{x}^*_{t+1} = \underset{\tilde{x}_{t+1}}{\operatorname{argmin}} E(G(x_{t-1}, \tilde{x}_{t+1}), x_t),$$

Followed by the optimization:

$$\tilde{f}^*_{t+1\to t} = \underset{\tilde{f}_{t+1\to t}}{\operatorname{argmin}} E(G(x_{t-1}, warp(x_t, \tilde{f}_{t+1\to t})), x_t).$$

1. Flow initialization: To perform any experiment, we need a seed value which is based on probability. The above equation can be optimized by starting with an approximate flow that is initialized using the negative flow of $f_{t\to t\text{-}1}$:

   $$\tilde{f}_{t+1\to t} = \delta(-f_{t\to t-1}).$$

   Yet, because the constraint towards $f_{t\to t\text{-}1}$ is indirect and the optimization method is challenging to converge, directly optimizing the above equation is still challenging.

2. Video Frame Interpolation Network: The above constraint leads to making use of network G's interim results. Provided following as the inputs

   $$x_{t-1} \text{ and } \tilde{x}_{t+1}$$

G generates optical flows of two directions and a mask of m_G

$$f^G_{t \to t-1},\ f^G_{t \to t+1}.$$

As part of this step, we calculate the following displacements:

(a) Displacement in image points

(b) Displacement in motion

Further, we convert these to pixel level to reduce the error minimization using the occlusion areas.

3. Flow Inpainting: In optical flow, there are always occlusion zones where some pixels do not have matching pixels in subsequent frames. In occlusion areas, the calculated optical flow is not accurate. In order to hide these areas, a threshold was selected, then flow inpainting was used to fill the gap. To fix the occlusions, an attempt to apply adaptive weights like Softmax Splatting was done.

$$\phi(\mathbf{p}) = \begin{cases} 1 & if \quad \left\| \Delta \tilde{f}_{t+1 \to t}(\mathbf{p}) \right\|_1 > \alpha, \\ 0 & \quad otherwise. \end{cases}$$

As shown in the above formula, 1 means the pixel would not be considered marking it as unwanted area and 0 means it is wanted area. Only later the author discovered that it is not appropriate for the architecture.

4. Implementation: The method for multi-frame prediction involves repeatedly optimizing the following frame. The author tried to combine the optimization of numerous optical flows and future frames, but discovered that it is more stable to optimize the following frame repeatedly.

8 Strengths Observed in this Approach

Following two strengths were observed with this approach:

1. Ability to outperform cutting-edge techniques

2. Agility to use any video of any resolution

9 Experimentation and Results

Several different datasets were used for experimentation based upon the information available. Many current methods for video prediction include extra presumptions such instance and semantic maps, which can increase prediction accuracy but limit generalization. The presented solution, however, gets over these limitations by framing the video prediction problem as an optimization problem, which does not call for any external training or presumptions regarding the data. The approach supposedly performs better than earlier approaches that make additional assumptions and is very broad. The supplement offers statistical analysis and visual comparisons.

1. Evaluation on Driving Datasets: For driving datasets where semantic information is available, the methodology was assessed and pertinent baselines because some baselines call for extra semantic maps.

    (a) Datasets: Cityscapes and KITTI datasets were used which contained driving sequences

    (b) Baselines: The initial group of techniques, such as PredNet, MCNet, and DVF, only accept RGB frames as input. The second class of methods, such as Vid2vid, Seg2vid, and FVS, call for extra data such as semantic maps and instance maps. Because they can only be used when this additional context is available, the later type of algorithms may not be as generalizable. The Multi-scale Structural Similarity Index Measure (MSSSIM) and LPIPS are the evaluation measures, and greater MS-SSIM and lower LPIPS indicate superior performance.

    (c) Quantitative results: The quantitative results show that the system can perform better in both short- and long-term video prediction without requiring external training on the Cityscapes and KITTI training sets. Some approaches employ hand-crafted loss functions, but these approaches could overfit to the main motion seen in the training set—the "zooming-in" effect—rather than learning the true motion.

    (d) Qualitative Results: Due to the datasets being recorded by a forward moving camera, DVF has a tendency to forecast motion through "zooming-in" rather of anticipating the genuine motion. To mimic the motion of moving cars, FVS creates a custom 2D affine transformation, however it is unable to capture complex motion, such as nonrigid deformation and 3D rotation. Additionally, when the assumptions made by DVF and FVS regarding semantic and instance segmentation are incorrect, their performance suffers. Cityscapes visual comparisons are shown in the supplement.

2. Evaluation on Diverse Datasets: The optimization structure of the method under discussion does not require external training, enabling it to be generalized to any video at any resolution.

    (a) Datasets: Several datasets, including DAVIS, Middlebury-Other, and Vimeo90K were used. The validation set of the DAVIS dataset has 30 sequences with resolutions of around 854 x 480. Ten films in the Middlebury dataset had resolutions of about 640 x 480. Every 10 video are taken to create a test set for the Vimeo90K dataset, which comprises 3782 triplets with a resolution of 448 256.

   (b) Baselines: The discussed method compares two latest external methods, DVF and DYAN, which take only RGB frames as input and can be applied to the datasets. Other methods that require additional assumptions cannot be compared because their assumptions do not hold. The authors test these two models on the datasets using their pre-trained model on UCF101.

   (c) Quantitative results: The quantative results are that the approach is still reliable for long-term forecasts. The test films and UCF-101 have a domain gap issue, which is a typical occurrence in video prediction tasks. To address this issue, the majority of video prediction systems develop unique models and make unique assumptions for each dataset. The VFI task's domain gap, however, does not appear to be a concern.

3. Ablation Study: Cityscapes is the subject of an analysis study to determine the significance of each element of the method. The effectiveness of next-frame prediction is used to rate several elements.

   (a) Initialization: It claims that because optical flow makes the optimization process easier, the optimization goal cannot be employed as the prediction frame.

   (b) Loss functions: Without $\mathrm{L}_{\mathrm{img}}$ supervised only by $\mathrm{L}_{\mathrm{cons}}$. Without $\mathrm{L}_{\mathrm{cons}}$ supervised only by $\mathrm{L}_{\mathrm{img}}$. The combination of $\mathrm{L}_{\mathrm{img}}$ and $\mathrm{L}_{\mathrm{cons}}$ produces the best performance, according to the findings. Compared to $\mathrm{L}_{\mathrm{interp}}$, $\mathrm{L}_{\mathrm{img}}$ serves as a simpler constraint. However, when dealing with substantial motion, long-term $\mathrm{L}_{\mathrm{img}}$ reduces the accuracy of VFI.

   (c) Flow inpainting: The performance suffers if we do away with the optical flow inpainting process since it effectively corrects incorrect flow data.

4. Convergence Analysis: A 256 x 512 resolution Cityscapes dataset was used for the convergence analysis. In the first 400 iterations, it was discovered that the optimization swiftly converged, with a steady improvement in the forecast outcomes. The outcome of the forecast continued to slowly improve even after 3000 repetitions.

10 Conclusion

A strategy is outlined for video prediction that, by framing the issue as a VFI-based optimization problem, addresses the domain gap issue. The method outperforms state-of-the-art techniques and can be used to any video at any resolution, however it takes longer to optimize each frame when using this method than when using other external learning-based techniques. The supplement provides a model size and inference time comparison between this method and other methods. The suggested method's gradient propagation inside the VFI network consumes the majority of run time, which implies the need for creating a future backbone for acceleration that is more effective.