Optimizing Video Prediction via Video Frame Interpolation

Ruchi Manikrao Dhore | W1652116

Introduction

- Video prediction has many practical applications, such as robotics planning, autonomous driving, and video manipulations
- Most video prediction methods require additional information about the scene
- These requirements limit the applicability of the methods to specific videos only
- A need for a video prediction method that can be applied to any video

Objective

- To propose new optimization framework for video prediction via video frame interpolation to solve an extrapolation problem based on an interpolation model
- To optimize optical flow by using video frame interpolation
- To optimize image level distance and consistency constraint between the predicted flow

Extrapolation vs Interpolation

- Extrapolation is the process to make predictions about future frames beyond the end of the input sequence
- Interpolation is the process of filling in missing frames within the input sequence

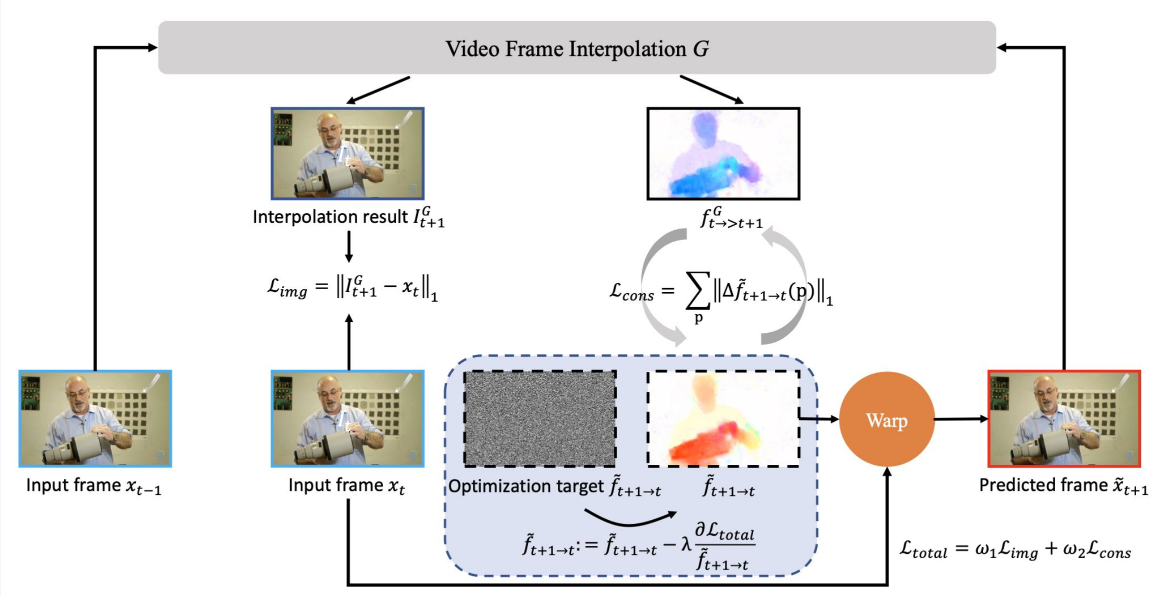
Literature Survey

	Video Prediction	Video Frame Interpolation	Optimization-based Methods
Main idea	Use of extrapolation and parameters semantic maps, depth dimension map	Worked on interpolation using three algorithms, learning methods were used	Optimization methods applied on algorithms explored in video frame interpolation

Problem Formulation

- In time vector dimension, let's say x_t is the video frame at any given time t
- Input being two RGB frames x_{t-1} and x_t , the idea is to predict the motion with accuracy in future frames x_{t+1}, x_{t+2}, …

Main idea

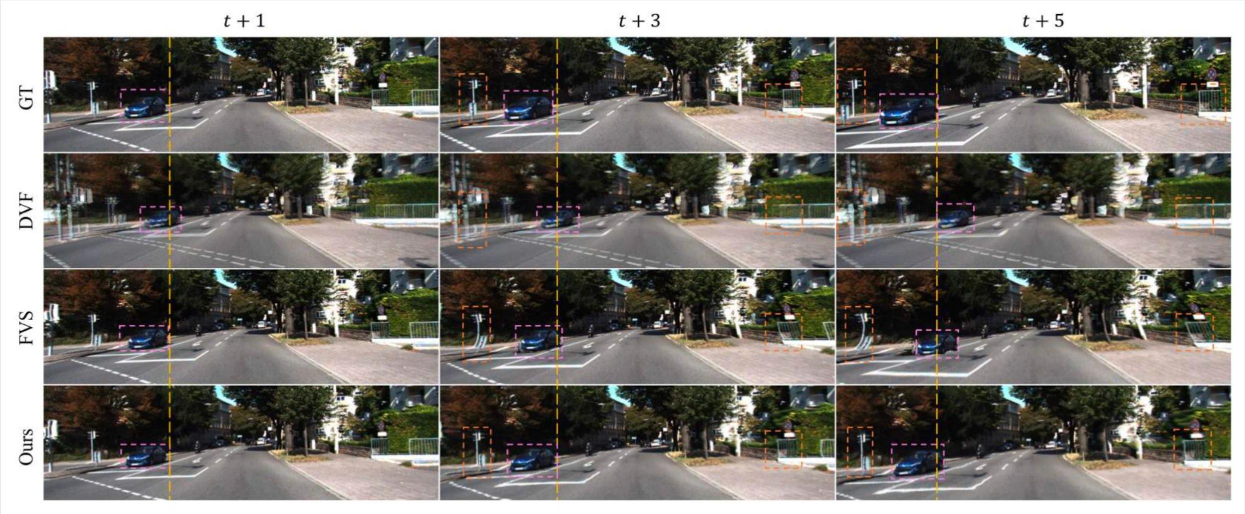


Video Frame Interpolation G
Interpolation result I_{t+1}^{G}
$f_{t\to>t+1}^{G}$
$\mathcal{L}_{img} = \|I_{t+1}^{G} - x_t\|_1$
$\mathcal{L}_{cons} = \sum_{p} \|\Delta \tilde{f}_{t+1\to t}(p)\|_1$
Input frame x_{t-1}
Input frame x_t
Optimization target $\tilde{f}_{t+1\to t}$
$\tilde{f}_{t+1\to t}$
Warp
Predicted frame $\tilde{x}_{t+1}$
$\tilde{f}_{t+1\to t} := \tilde{f}_{t+1\to t} - \lambda \frac{\partial \mathcal{L}_{total}}{\tilde{f}_{t+1\to t}}$
$\mathcal{L}_{total} = \omega_1 \mathcal{L}_{img} + \omega_2 \mathcal{L}_{cons}$

Mathematics

1 $\tilde{x}^*_{t+1} = \underset{\tilde{x}_{t+1}}{\text{argmin}}\, E(G(x_{t-1}, \tilde{x}_{t+1}), x_t),$		
Flow initialization	Video Frame Interpolation Network	Flow implanting
2 $\tilde{f}_{t+1\to t} = \delta(-f_{t\to t-1}).$	3 4 5 $\begin{aligned} I^G_{t-1} &= warp(x_{t-1}, f^G_{t\to t-1}), \\ I^G_{t+1} &= warp(\tilde{x}_{t+1}, f^G_{t\to t+1}), \\ I^G_t &= I^G_{t-1} \times m^G + I^G_{t+1} \times (1 - m^G) \end{aligned}$ 6 $\mathcal{L}_{img} = \left\lVert I^G_{t+1} - x_t \right\rVert_1$ 7 $\mathcal{L}_{cons} = \sum_{\mathbf{p}} \left\lVert \Delta \tilde{f}_{t+1\to t}(\mathbf{p}) \right\rVert_1$ 8 $\begin{aligned} \Delta \tilde{f}_{t+1\to t}(\mathbf{p}) &= \mathbf{p} - (\mathbf{p}' + f^G_{t\to t+1}(\mathbf{p}')) \\ \mathbf{p}' &= \mathbf{p} + \tilde{f}_{t+1\to t}(\mathbf{p}). \end{aligned}$	9 $\phi(\mathbf{p}) = \begin{cases} 1 & if \quad \left\lVert \Delta \tilde{f}_{t+1\to t}(\mathbf{p}) \right\rVert_1 > \alpha, \\ 0 & otherwise. \end{cases}$

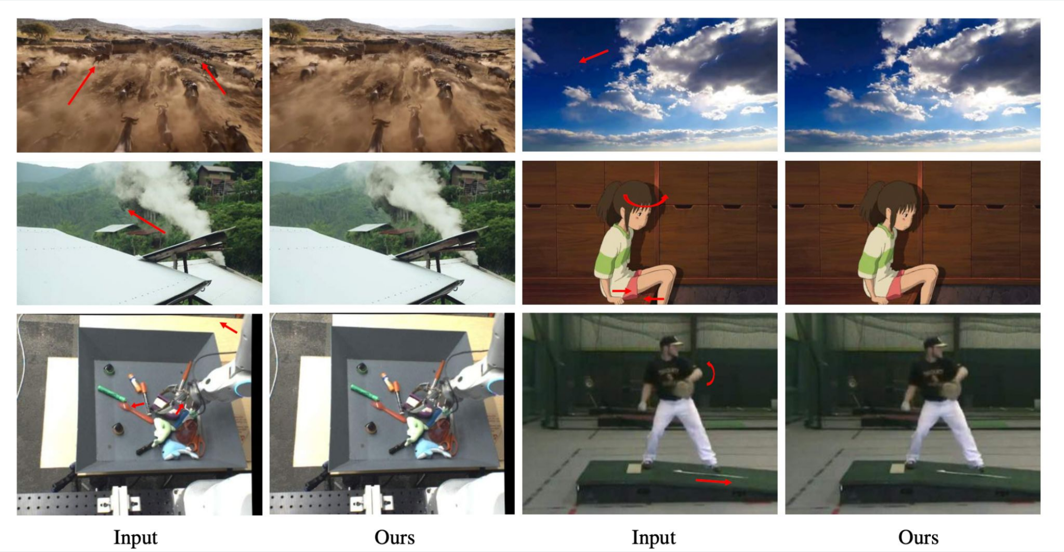
Experimentation & Results | Driving datasets



Experimentation & Results | Driving datasets

	Input	Cityscapes MS-SSIM (×1e−2)↑ t+1	t+3	t+5	Cityscapes LPIPS (×1e−2)↓ t+1	t+3	t+5	KITTI MS-SSIM (×1e−2)↑ t+1	t+3	t+5	KITTI LPIPS (×1e−2)↓ t+1	t+3	t+5
External learning methods													
PredNet [22]	RGB	84.03	79.25	75.21	25.99	29.99	36.03	56.26	51.47	47.56	55.35	58.66	62.95
MCNET [42]	RGB	89.69	78.07	70.58	18.88	31.34	37.34	75.35	63.52	55.48	24.05	31.71	37.39
DVF [20]	RGB	83.85	76.23	71.11	17.37	24.05	28.79	53.93	46.99	42.62	32.47	37.43	41.59
Vid2vid [43]	RGB+S.	88.16	80.55	75.13	10.58	15.92	20.14	N/A	N/A	N/A	N/A	N/A	N/A
Seg2vid [31]	RGB+S.	88.32	N/A	61.63	9.69	N/A	25.99	N/A	N/A	N/A	N/A	N/A	N/A
FVS [45]	RGB+S.+I.	89.10	81.13	75.68	8.50	12.98	**16.50**	79.28	67.65	60.77	18.48	24.61	30.49
Optimization methods													
Ours	No external training	**94.54**	**86.89**	**80.40**	**6.46**	**12.50**	17.83	**82.71**	**69.50**	**61.09**	**12.34**	**20.29**	**26.35**

Experimentation & Results | Diverse datasets



Experimentation & Results | Diverse datasets

	DAVIS				Middlebury				Vimeo90K	
	MS-SSIM ($\times$1e$-$2)$\uparrow$		LPIPS ($\times$1e$-$2)$\downarrow$		MS-SSIM ($\times$1e$-$2)$\uparrow$		LPIPS ($\times$1e$-$2)$\downarrow$		MS-SSIM ($\times$1e$-$2)$\uparrow$	LPIPS ($\times$1e$-$2)$\downarrow$
	t+1	t+3	t+1	t+3	t+1	t+3	t+1	t+3	t+1	t+1
External learning methods										
DVF [20]	68.61	55.47	23.23	34.22	83.98	65.54	13.57	25.70	92.11	7.73
DYAN [18]	78.96	70.41	13.09	21.43	92.96	83.91	7.98	15.03	N/A	N/A
Optimization methods										
Ours	**83.26**	**73.85**	**11.40**	**18.21**	**94.49**	**87.96**	**6.07**	**10.82**	**96.75**	**3.59**

Experimentation & Results | Comparison

External learning methods				
	External training	Semantic	Instance	Depth
PredNet [22]	✓	×	×	×
MCNET [42]	✓	×	×	×
DVF [20]	✓	×	×	×
Vid2vid [43]	✓	✓	×	×
Qi *et al*. [35]	✓	✓	×	✓
Seg2vid [31]	✓	✓	×	×
FVS [45]	✓	✓	✓	×
HVP [15]	✓	✓	×	×
SADM [4]	✓	✓	×	×
Optimization methods				
Ours	×	×	×	×

Conclusion

- Ability to apply to any video at any resolution
- Outperforms state-of-the-art methods in terms of accuracy

Limitations

- The optimization process takes more time than other external learning-based methods
- Most of the run time of their model is spent on gradient propagation inside the VFI network

Thank you!