Explicit Depth-Aware Blurry Video Frame Interpolation Guided by Differential Curves – supplementary materials

Zaoming Yan [1] Pengcheng Lei [1*] Tingting Wang[1] Faming Fang[1†] Junkang Zhang[1]
Yaomin Huang[1]
Haichuan Song[1†]
[1] School of Computer Science and Technology, and the KLATASDS
Shanghai Key Laboratory of MIP, East China Normal University, Shanghai, China.
[1]{yan_zaoming,pengchenglei1995}@163.com
[1]ymhuang@stu.ecnu.edu.cn [1]fmfang@cs.ecnu.edu.cn

Contents

1. Theoretical Supplement

1.1. Differential Curve in High Dimension

In high-dimensional space, based on the definition of differential curves, we present the derivative terms of differential curves in high-dimensional space:

$$\begin{bmatrix} r_1^{'}(s) \\ r_2^{'}(s) \\ r_3^{'}(s) \\ r_4^{'}(s) \end{bmatrix} = \begin{bmatrix} 0 & \kappa_{12}(s) & 0 & 0 \\ \kappa_{21}(s) & 0 & \kappa_{23}(s) & 0 \\ 0 & \kappa_{32}(s) & 0 & \kappa_{34}(s) \\ 0 & 0 & \kappa_{43}(s) & 0 \end{bmatrix} \begin{bmatrix} r_1(s) \\ r_2(s) \\ r_3(s) \\ r_4(s) \end{bmatrix}, \tag{1}$$

where $\kappa(s)$ is the curvature of high-dimensional slight curve $r(s)$. The desirable properties of this differential curve can be extended to various tasks.

*Equal contribution.
†Corresponding Author

1.2. Convergence Analysis

We provide the proof of convergence for the differential curve. First, a convergence analysis is conducted at s^*. Expanding $r(s)$ at s^* using a Taylor expansion and truncating at the first order yields the following expression:

$$r(s) = r(s^*) + \frac{r^{'}(s^*)}{1!}(s - s^*) + o(s^2). \tag{2}$$

We obtain the $|r(s) - r(s^*)|$ as follows:

$$|r(s) - r(s^*)| = |r(s^*) + \frac{r^{'}(s^*)}{1!}(s - s^*) + o(s^2)| \tag{3}$$

$$< |r^{'}(s^*)||s - s^*|. \tag{4}$$

We can derive the following formula:

$$|s - s^*| < \frac{\delta}{|r^{'}(s^*)|}, \delta > 0 \tag{5}$$

From the above equation, the following can be deduced:

$$|r(s) - r(s^*)| < \epsilon = \delta, \epsilon > 0 \tag{6}$$

This implies that when the condition (Eq. (5)) is satisfied, $r(s)$ will convergence to $r(s^*)$.

2. Additional Framework Analysis

2.1. Model parameters.

Tab. 1 illustrates that the model performance is not strongly correlated with parameters. Most other methods employ recurrent strategies to increase performance, leading to smaller parameter sizes but increased FLOPs.

Table 1. Model Information.

Method	Model Information		
	#params	FLOPs	time (s)
LDMVFI [1]	439M	9.9T	8.48
VIDIM [3]	357M	10.8T	9.28
QVI [9]	19M	1.3T	0.97
PRF [4]	11.4M	3.2T	0.76
Ours	88M	0.12T	0.124

Table 2. The method name with (G) denotes the generative model. The method name with (R) denotes the reconstruction model.

Method	LPIPS ↓	tOF (time-axis) ↓	PSNR
LDMVFI [1](G)	0.1356	17.28	26.301
VIDIM [3](G)	0.1495	20.11	24.07
QVI [9](R)	0.1986	8.91	32.54
Ours (R)	0.1683	7.63	35.54

2.2. Limitations and failures

It is worth noting that the processing ability of the DC-BMFI for discontinuous motion scenes needs to be improved, as shown in Fig. 1.

Figure 1. Failure cases in discontinuous motion.

3. Additional Framework Results

3.1. Comparison with DDPM-based Methods

The diffusion model uses stochastic differential equations to generate video. The disadvantage is that the object's motion in the generated video lacks physical constraints, such as motion continuity. Video frame interpolation is to reconstruct the real continuous motion trajectory.

For quantitative measurements, we tested some of the generative and reconstruction models on the UCF101 [5] dataset. Please see the table below: As we can see from Tab. 2, the generative model performs worse than the reconstructed model on the motion continuity metric (tOF).

3.2. More Visual Results

Fig. 2 presents the video frame interpolation results in the depth variations scene. Fig. 3 and Fig. 4 show more video multi-frame interpolation results. We have provided the 3D visualization of the interpolation results in an anonymous Git link https://baronzaomingyan.github.io/DC-BMVI/.

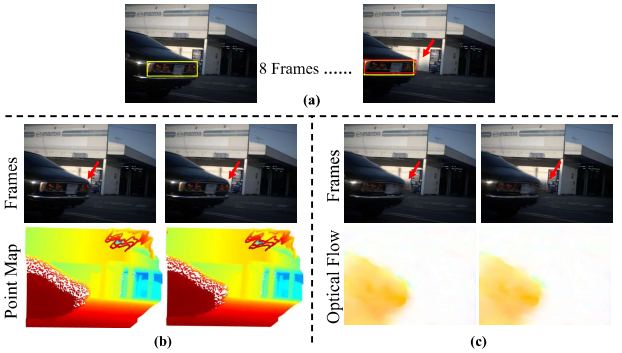

Figure 2. (a) Depth variation leads to changes in object size. (b) and (c) present the multi-frame video interpolation results of the 3D-based and 2D-based frameworks, respectively.

4. Additional Ablation Studies

To verify the impact of different backbones on the overall framework, We primarily evaluate this by replacing the ViT in UBNet and substituting the Flownet in MPNet.

4.1. UBNet

To evaluate the impact of different backbones on the transformation of frames into motion-space point maps and their influence on video frame interpolation. We replaced ViT with MoGe [7], DUSt3R [8], and conducted evaluations on the RBI test set. The results presented in Tab. 3 demonstrate that varying backbones have a negligible effect on the direct outcomes of video frame interpolation.

Table 3. The effect of various ViT configurations on blurry video frame interpolation.

ViT	MoGe [7]	DUSt3R [8]	PSNR(dB)	MiD (↓)
(a)	✓		31.03	76
(b)		✓	30.99	82

Table 4. The effect of different 2D optical flow configurations on blurry video frame interpolation.

FlowNet	IFNet [2]	RAFT [6]	PSNR(dB)	MiD (↓)
(a)	✓		31.03	76
(b)		✓	31.00	79

4.2. MPNet

To evaluate the effect of various 2D optical flow backbones as the foundation for MPNet and their influence on video frame interpolation, we employ IFNet [2] and RAFT [6] as

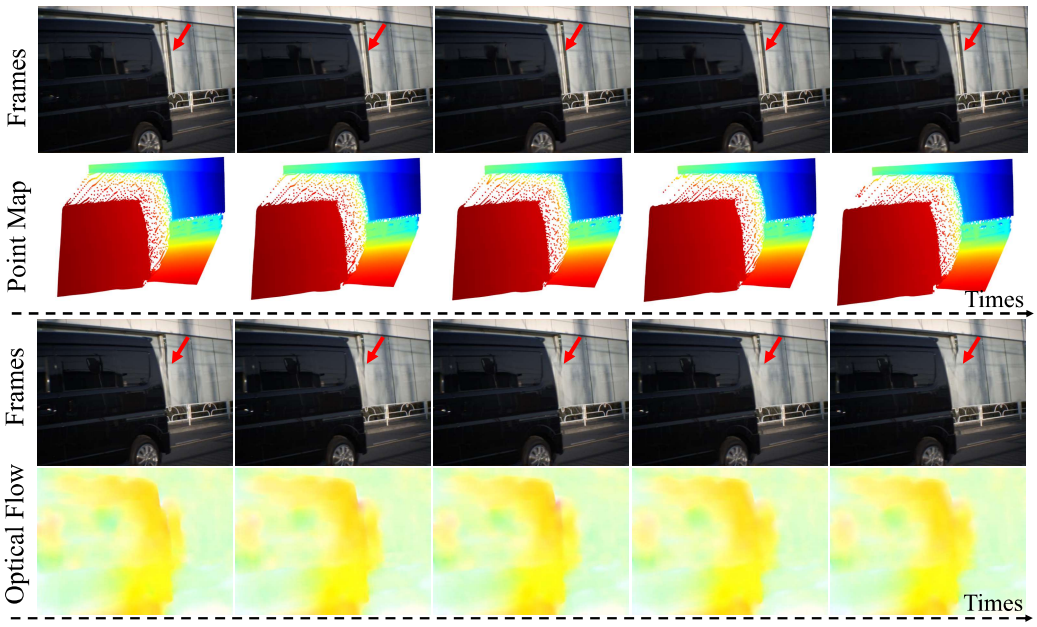

Figure 3. In the multi-frame video interpolation results, the top row presents the performance of the 3D-based framework, while the bottom row showcases the results of the 2D-based framework.

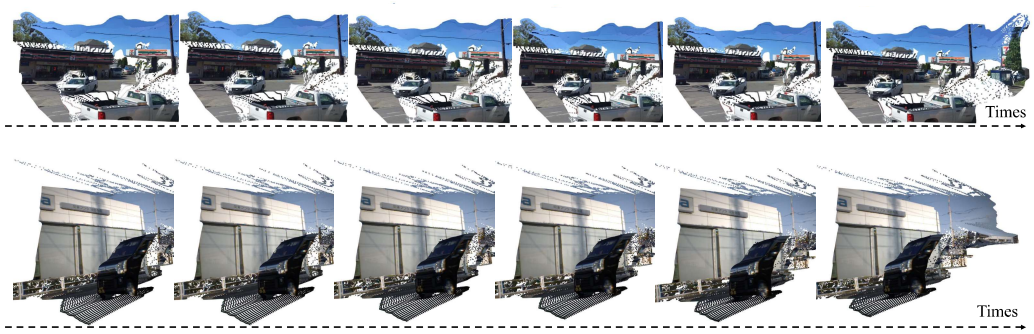

Figure 4. The multi-frame video interpolation results based on the 3D framework are presented.

the 2D optical flow models for MPNet. We conduct evaluations on the RBI test set. As shown in Tab. 4, the results suggest that different 2D optical flow models have a negligible impact on the motion-depth changes in video frame interpolation.

References

[1] Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1472–1480, 2024. 2

[2] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pages 624–642. Springer, 2022. 2

[3] Siddhant Jain, Daniel Watson, Eric Tabellion, Ben Poole, Janne Kontkanen, et al. Video interpolation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7341–7351, 2024. 2

[4] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Video frame interpolation and enhancement via pyramid recurrent framework. *IEEE Transactions on Image Processing*, 30:277–292, 2020. 2

[5] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 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] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu

Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision, 2024. 2

[8] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2

[9] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *NeurIPS*, 2019. 2