Inertia-Guided Flow Completion and Style Fusion for Video Inpainting

Kaidong Zhang[1] Jingjing Fu[2] Dong Liu[1]

[1] University of Science and Technology of China [2] Microsoft Research Asia

Introduction

Goal

- Input: Video frames and frame-wise masks
- Output: multiple inpainting solutions

Key Points

- Exploitation of complementary video content
- Maintenance of spatiotemporal coherence

Motivations

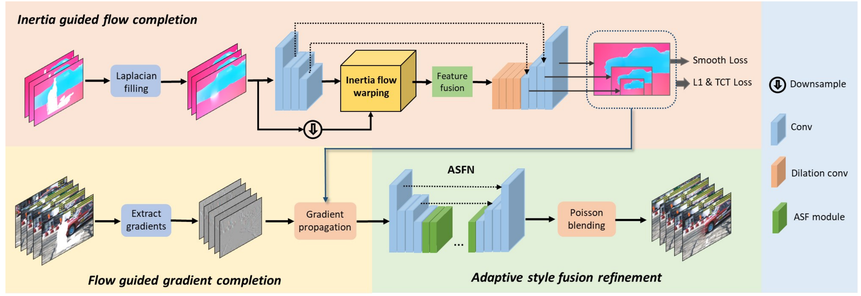
- The existence of inertia makes the nearby optical flows relevant
- The style difference across frames causes spatial incoherence

Contributions

- We introduce the inertia prior to model the inherent correlation within optical flow sequences and design a novel Inertia-Guided Flow Completion Network for more accurate flow completion.
- We propose the Adaptive Style Fusion Network (ASFN) to refine the warped gradients in the warped regions to alleviate the spatial incoherence caused by style variation across different frames.
- We establish a data simulation pipeline for ASFN training, which degrades the data preparation cost significantly for more efficient training.
- Experiments on Youtube-VOS and DAVIS datasets show the superiority of our method on various type of masks quantitatively and qualitatively.

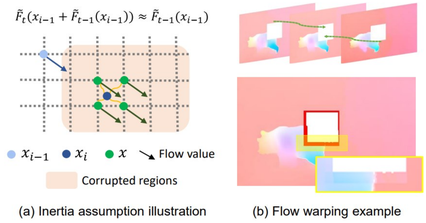
Method

Pipeline



Inertia Prior

Inertia assumption assumes the motion pattern between corresponding points across nearby optical flows is identical. Such assumption empowers the exploitation of complementary regions between nearby optical flows.

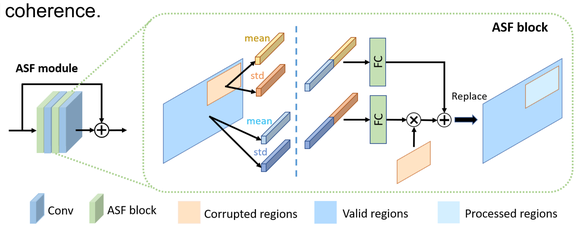


(a) Inertia assumption illustration

(b) Flow warping example

Adaptive Style Fusion Network (ASFN)

Since the style information in valid regions is known. We extract its mean value and standard deviation to correct the corresponding style information extracted from the invalid counterparts. After style correction, we modulate the corrupted regions with the correct style information to achieve spatial coherence.



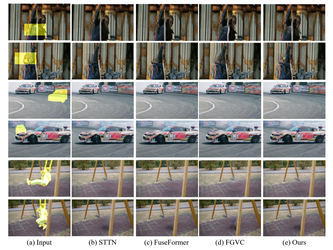
Experiments

Quantitative Analysis

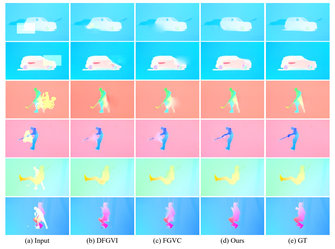
Method	Youtube-VOS			DAVIS								
				square			object			960×600		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
VINet [18]	29.83	0.9548	0.0470	28.32	0.9425	0.0494	28.47	0.9222	0.0831	-	-	-
DFGVI [48]	32.05	0.9646	0.0380	29.75	0.9589	0.0371	30.28	0.9254	0.0522	29.10	0.9249	0.0564
CPN [20]	32.17	0.9630	0.0396	30.20	0.9528	0.0489	31.59	0.9332	0.0578	-	-	-
OPN [29]	32.66	0.9647	0.0386	31.15	0.9578	0.0443	32.40	0.9443	0.0413	-	-	-
3DGC [5]	30.22	0.9607	0.0410	28.19	0.9439	0.0485	31.69	0.9396	0.0535	-	-	-
STTN [54]	32.49	0.9642	0.0400	30.54	0.9540	0.0468	32.83	0.9426	0.0524	-	-	-
TSAM [57]	31.62	0.9615	0.0314	29.73	0.9505	0.0364	31.50	0.9344	0.0478	-	-	-
FFM [25]	33.73	0.9704	0.0297	31.87	0.9652	0.0340	34.19	0.9510	0.0449	-	-	-
FGVC [8]	33.94	0.9719	0.0259	32.14	0.9667	0.0298	33.91	0.9554	0.0360	34.23	0.9607	0.0345
Ours	34.79	0.9743	0.0225	33.23	0.9729	0.0247	35.16	0.9648	0.0304	35.40	0.9659	0.0303

Qualitative Analysis

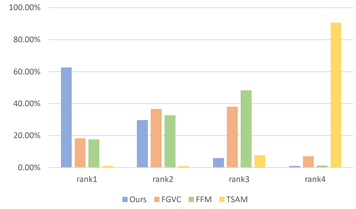
- Frames



- Flows



- User Study



- Video Results

Youtube Bilibili

Code and pre-trained models are available at:
https://github.com/hitachinsk/ISVI