Microsoft Research 微软亚洲研究院

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, Hefei, China

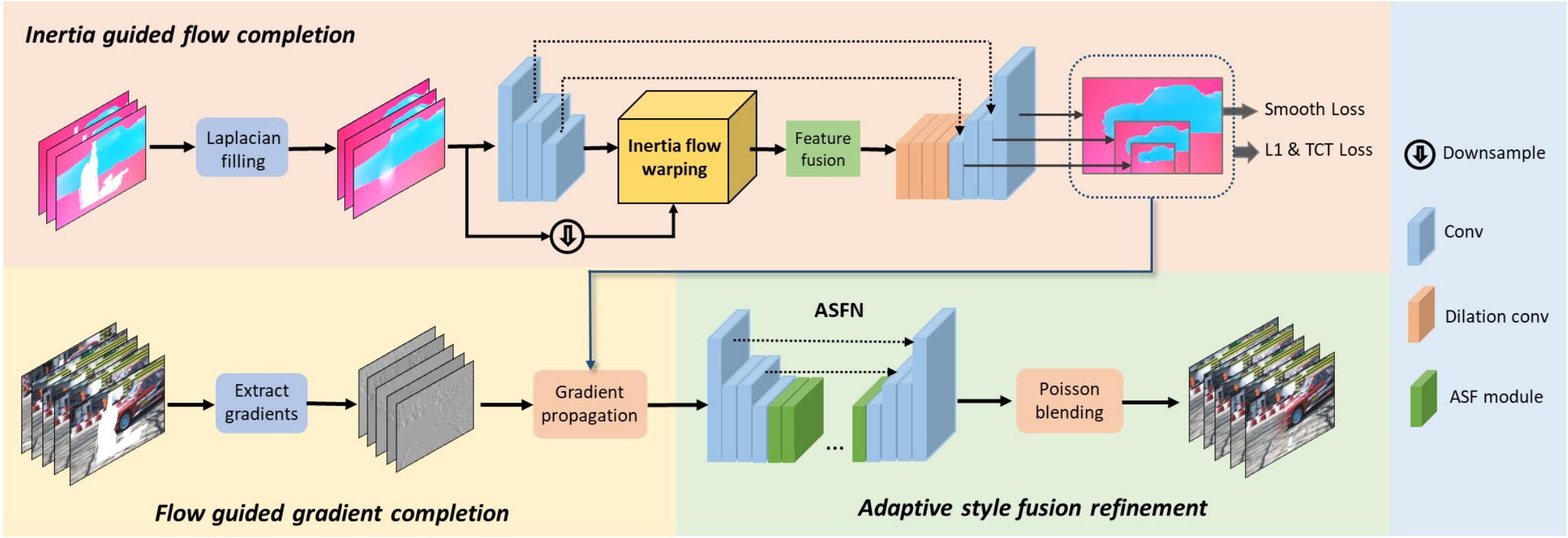
[2] Microsoft Research Asia, Beijing, China

Code is available at: https://github.com/hitachinsk/ISVI

Microsoft Research 微软亚洲研究院

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Inertia-Guided Flow Completion and Style Fusion for Video Inpainting



Video Inpainting

- Input: Video frames and frame-wise masks
- Output: Completed video frames with spatiotemporal coherence
- Application: Watermark removal, object removal, video retargeting, etc.

Input

Output

Key Point in Video Inpainting

- Exploitation of complementary video content
- Maintenance of spatiotemporal coherence

T=10 T=15 T=20 T=25 T=30

Output: T=20

Zoomed patch

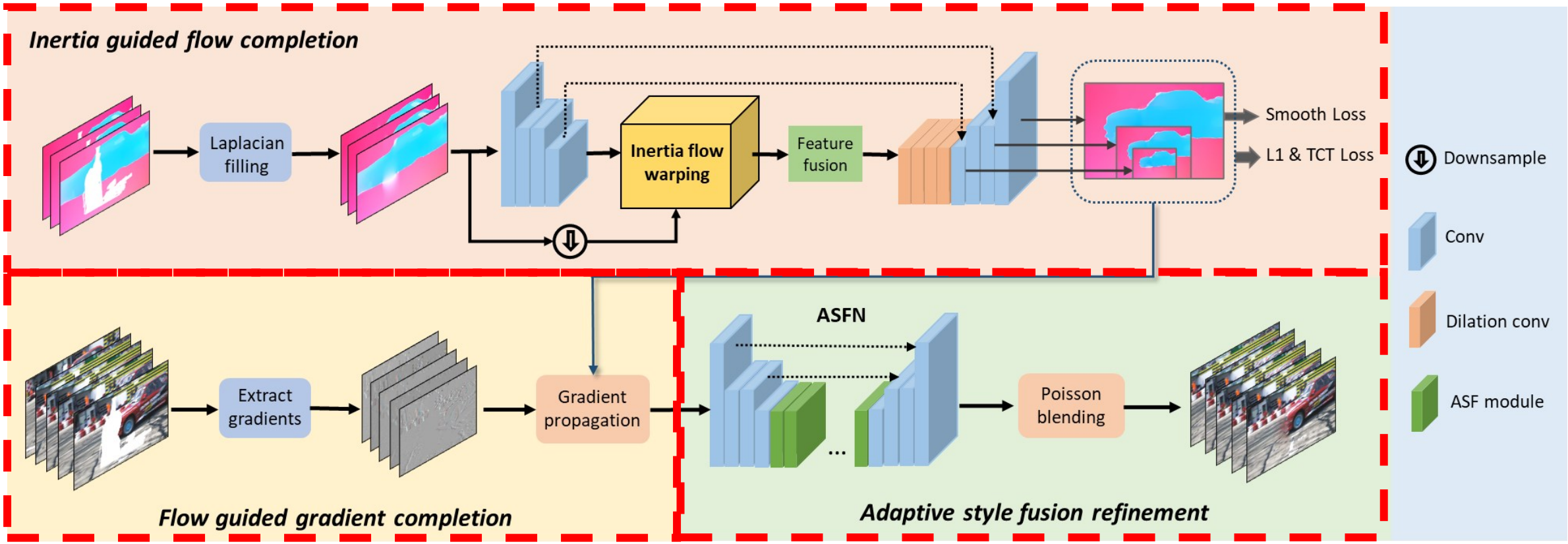
Motivation

- Inertia exists in any object, which causes the nearby optical flows correlated. Such context can be used for more accurate flow completion

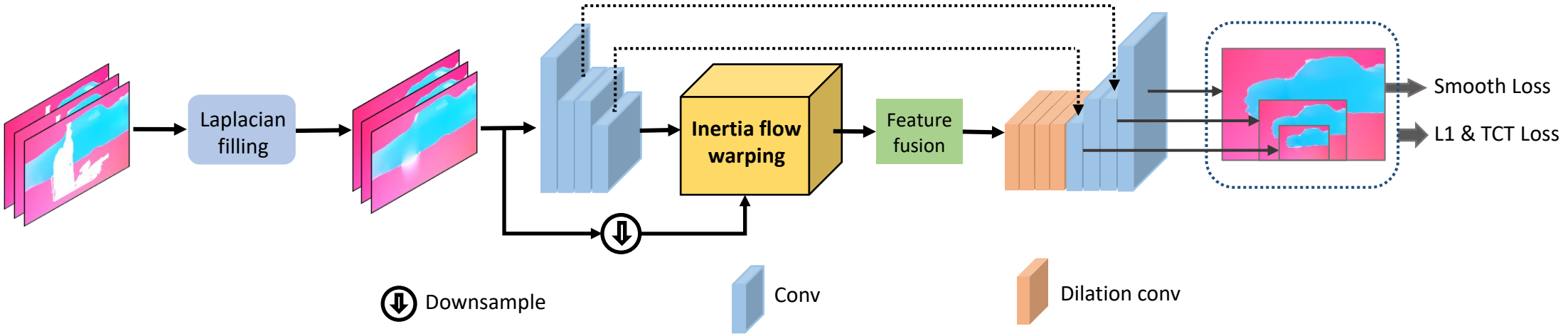
- After flow-guided content propagation, the style difference between different frames causes the spatial inconsistency between the filled regions and the valid regions. An extra style correction component is necessary to eliminate such difference.

Img	Mean	Std
Bear/00010.jpg	88.27	44.67
Bear/00011.jpg	88.00	44.62
Bear/00021.jpg	86.54	43.10

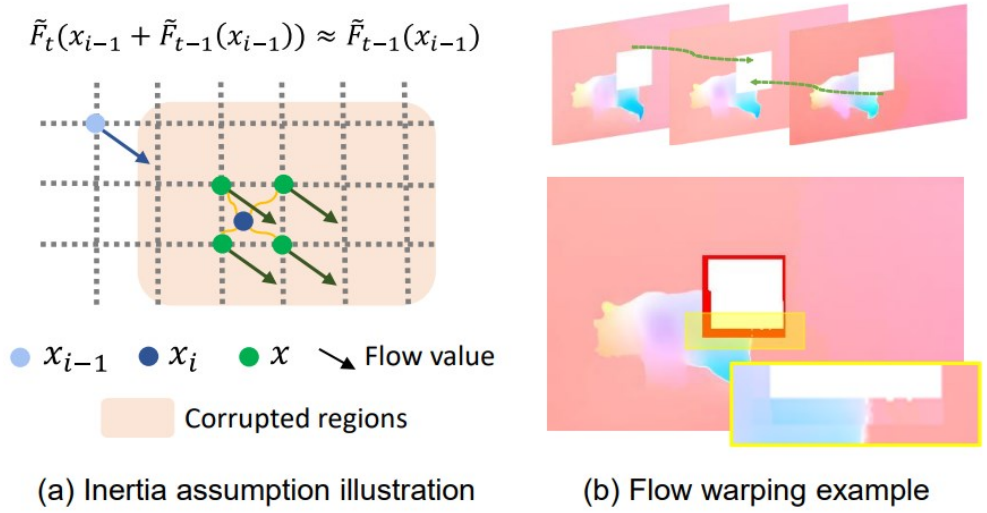
Overview


Inertia guided flow completion
Laplacian filling
Inertia flow warping
Feature fusion
Smooth Loss
L1 & TCT Loss
Downsample
Conv
Dilation conv
ASF module
Extract gradients
Gradient propagation
ASFN
Poisson blending
Flow guided gradient completion
Adaptive style fusion refinement

Inertia-Guided Flow Completion Network (IGFC)



Inertia Flow Warping Module



(a) Inertia assumption illustration

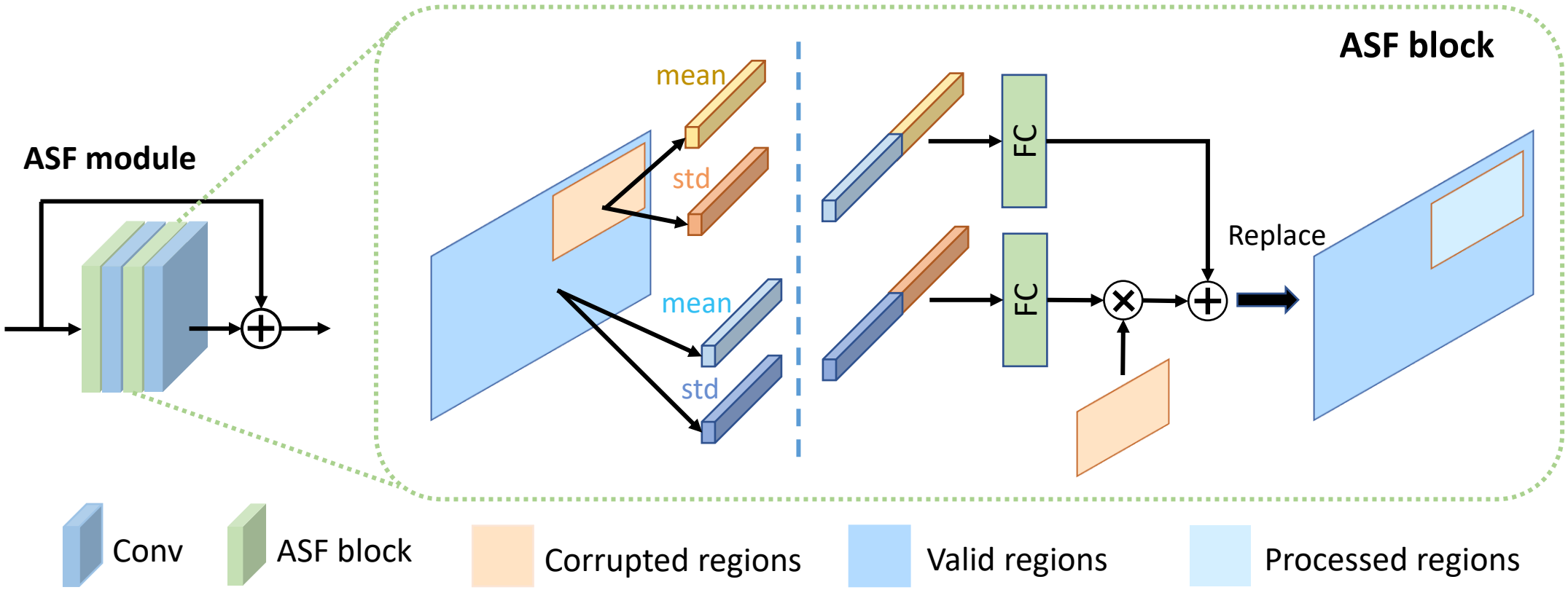
(b) Flow warping example

Visualization Results

(a) w/o Inertia warp

(a) w/ Inertia warp

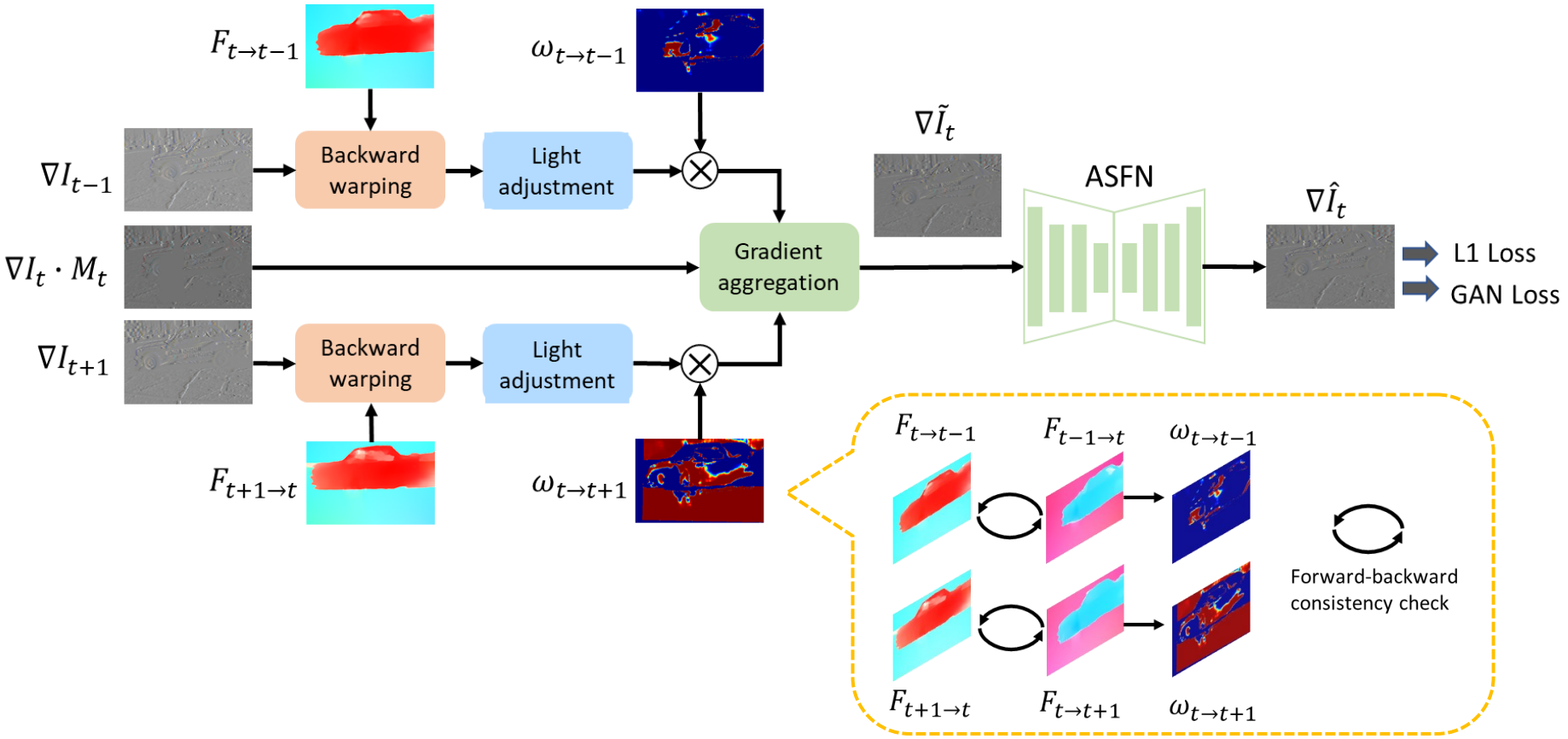
Adaptive Style Fusion Network (ASFN)



Adaptive Style Fusion Network (ASFN)

(a) Input (b) Without ASFN (c) With ASFN (d) GT

Data Simulation Pipeline



Optimization Objectives

IGFC

1. Flow reconstruction loss in hole and valid regions

$$L_{hole} = \left\| M_t \odot (F_t - \hat{F}_t) \right\|_1 / \|M_t\|_1$$
$$L_{valid} = \left\| (1 - M_t) \odot (F_t - \hat{F}_t) \right\|_1 / \|(1 - M_t)\|_1$$

2. Smoothness loss

$$L_{smooth} = \left\| \nabla \hat{F}_t \right\|_1 + \left\| \triangle \hat{F}_t \right\|_1$$

3. Ternary census transform loss L_{ter}

$$L = \lambda_1 L_{hole} + \lambda_2 L_{valid} + \lambda_3 L_{smooth} + \lambda_4 L_{ter}$$

ASFN

1. Reconstruction loss

$$Ls_{hole} = \left\| M_t \odot (\nabla I_t - \nabla \hat{I}_t) \right\|_1 / \|M_t\|_1$$
$$Ls_{valid} = \left\| (1 - M_t) \odot (\nabla I_t - \nabla \hat{I}_t) \right\|_1 / \|(1 - M_t)\|_1$$
$$Ls_{rec} = Ls_{hole} + Ls_{valid}$$

2. GAN loss
   - Discriminator loss

$$Ls_D = \mathbb{E}_{x \sim P_{\nabla I_t}(x)}[\mathrm{ReLU}(1 + D(x))] + \mathbb{E}_{z \sim P_{\nabla \hat{I}_t}(z)}[\mathrm{ReLU}(1 - D(z))]$$

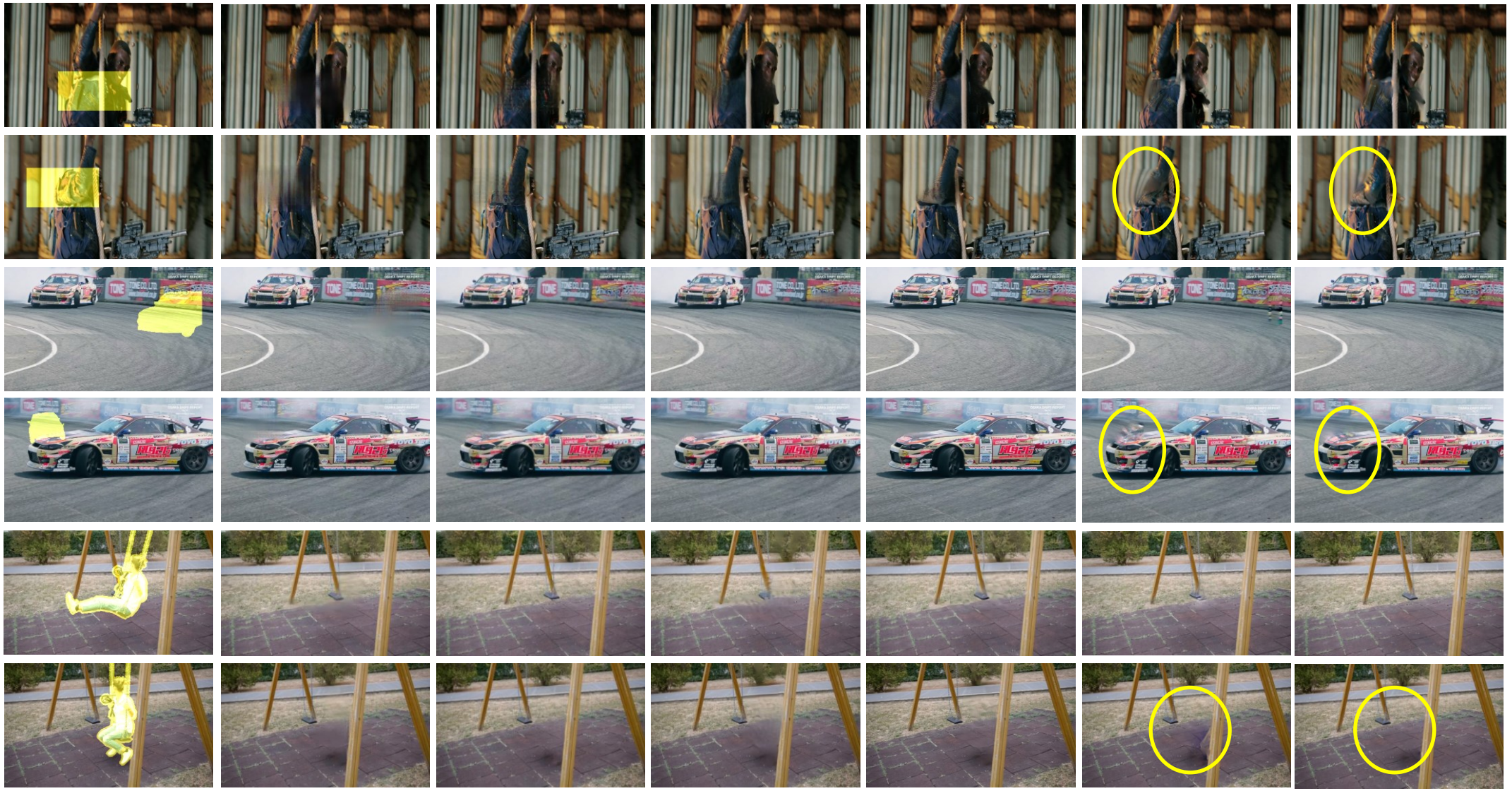
   - Adversarial loss

$$Ls_{adv} = -\mathbb{E}_{z \sim P_{\nabla \hat{I}_t(z)}}[D(z)]$$

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 Comparisons (frames and videos)

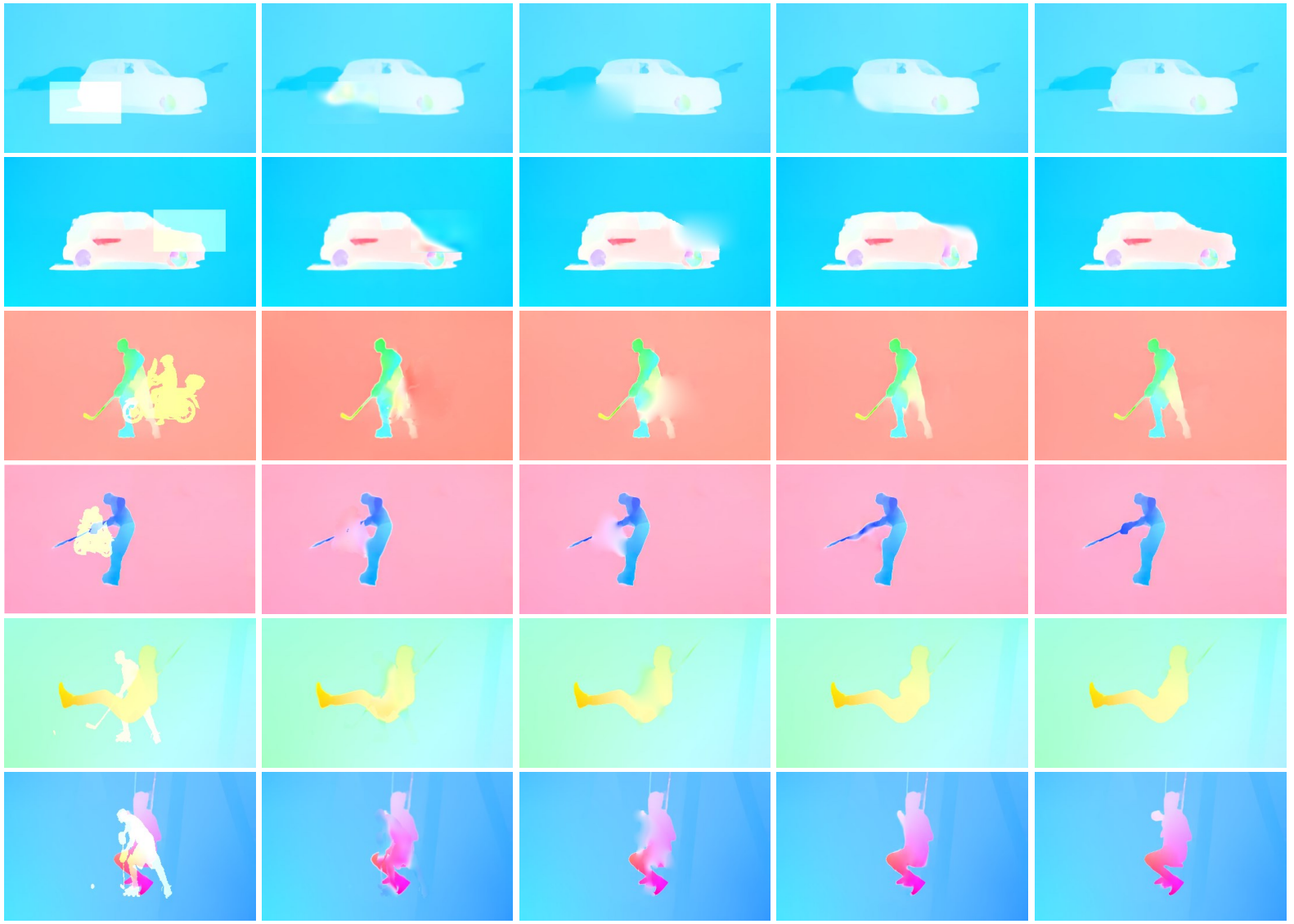


(a) Input (b) CPN (c) OPN (d) STTN (e) FuseFormer (f) FGVC (g) Ours

Demo video (Youtube)

Demo video (Bilibili)

Qualitative Comparisons (Flows)

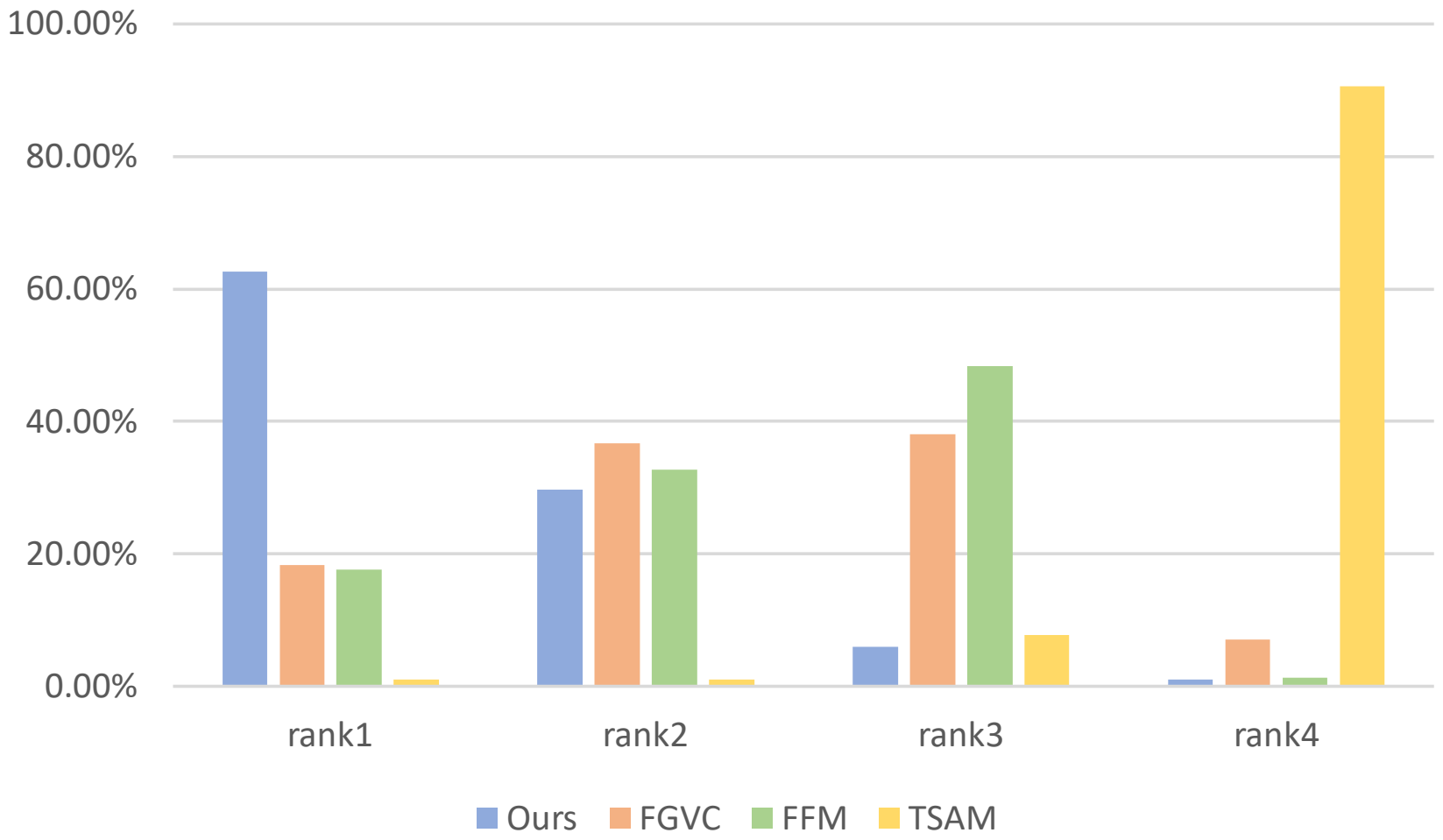


(a) Input (b) DFGVI (c) FGVC (d) Ours (e) GT

Demo video (Youtube)

Demo video (Bilibili)

User Study


100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
rank1
rank2
rank3
rank4
Ours
FGVC
FFM
TSAM

Conclusion

- Adopt inertia prior to exploit the complementary regions in nearby optical flows for the first time
- Design adaptive style fusion network (ASFN) to eliminate the style inconsistency after flow-guided content propagation for spatial coherence
- Design a novel data simulation pipeline to reduce training cost of ASFN
- Our method has excellent quantitative and qualitative performance

Welcome to visit our project page:

https://github.com/hitachinsk/ISVI