



Microsoft
Research
微软亚洲研究院



Inertia-Guided Flow Completion and Style Fusion for Video Inpainting

Kaidong Zhang¹, Jingjing Fu², Dong Liu¹

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>



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

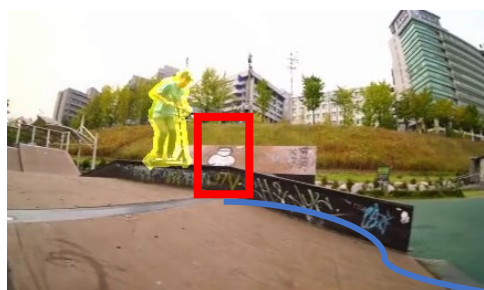


Output



Key Point in Video Inpainting

- Exploitation of **complementary video content**
- Maintenance of **spatiotemporal coherence**



T=10



T=15



T=20



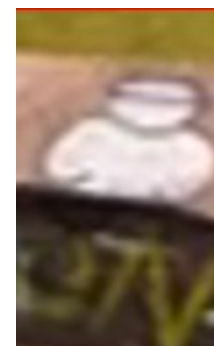
T=25



T=30



Output: T=20



Zoomed patch

Introduction | Method | Experiments | Conclusion

Related Works



Categories	Works	Long-range Modeling	Temporal Consistency	Details Maintenance	Large Resolution
CNN/RNN based	VINet [1], 3DGC [2], TSAM [3]		√		
Attention based	CPN [4], OPN [5], STTN [6], FFM [7]	√	√		
Flow Based	DFGVI [8], FGVC [9], Ours	√	√	√	√

- [1] Kim et al., Deep Video Inpainting. In CVPR 2019.
- [2] Chang et al., Free-form Video Inpainting with 3D Gated Convolution and Temporal PatchGAN. In ICCV 2019.
- [3] Zou et al. Progressive Temporal Feature Alignment Network for Video Inpainting. In CVPR 2021
- [4] Lee et al., Copy-and-Paste networks for Deep Video Inpainting. In ICCV 2019.
- [5] Oh et al., Onion-Peel Networks for Deep Video Completion. In ICCV 201
- [6] Zeng et al. Learning Joint Spatial-Temporal Transformations for Video Inpainting
- [7] Liu et al. FuseFormer: Fusing Fine-Grained Information in Transformers for Video Inpainting
- [8] Xu et al. Deep Flow-Guided Video Inpainting
- [9] Gao et al. Flow-edge Guided Video Inpainting

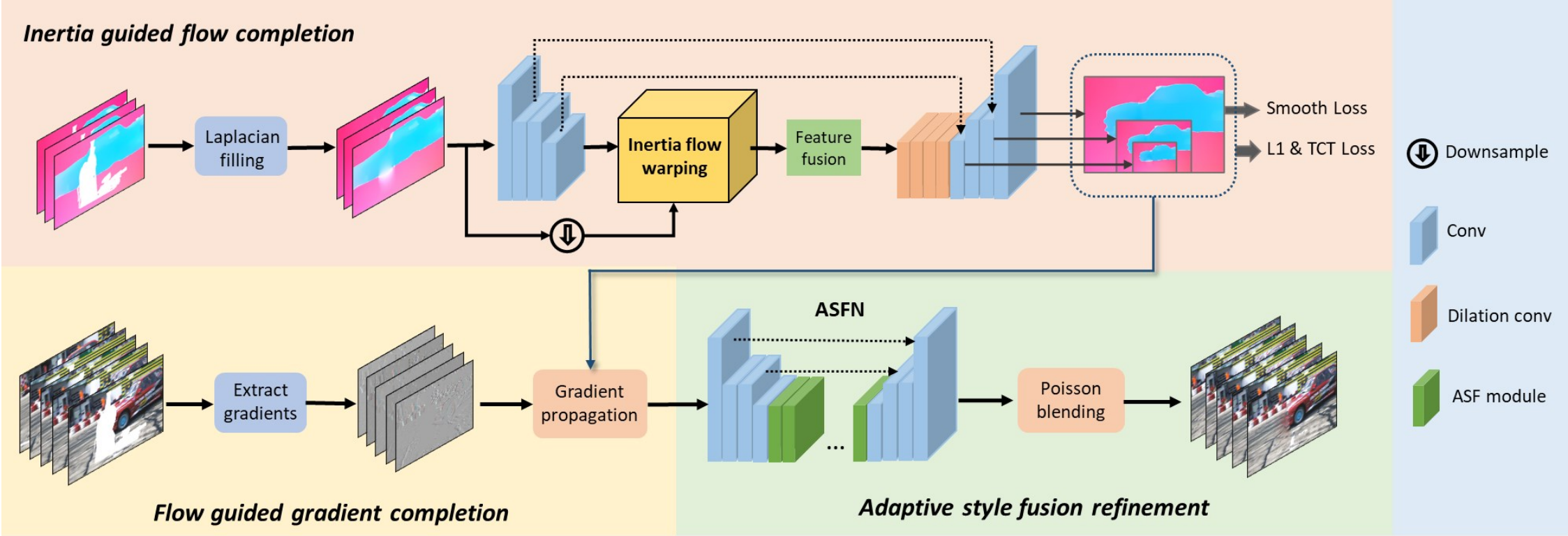
Motivation

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

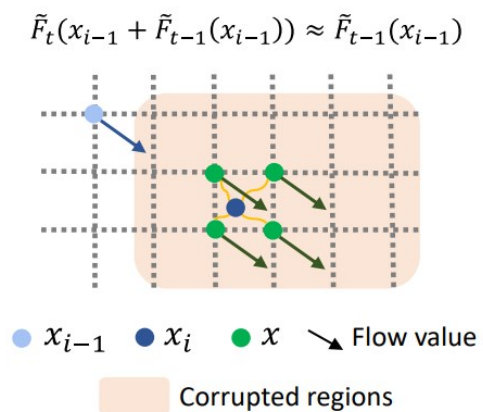
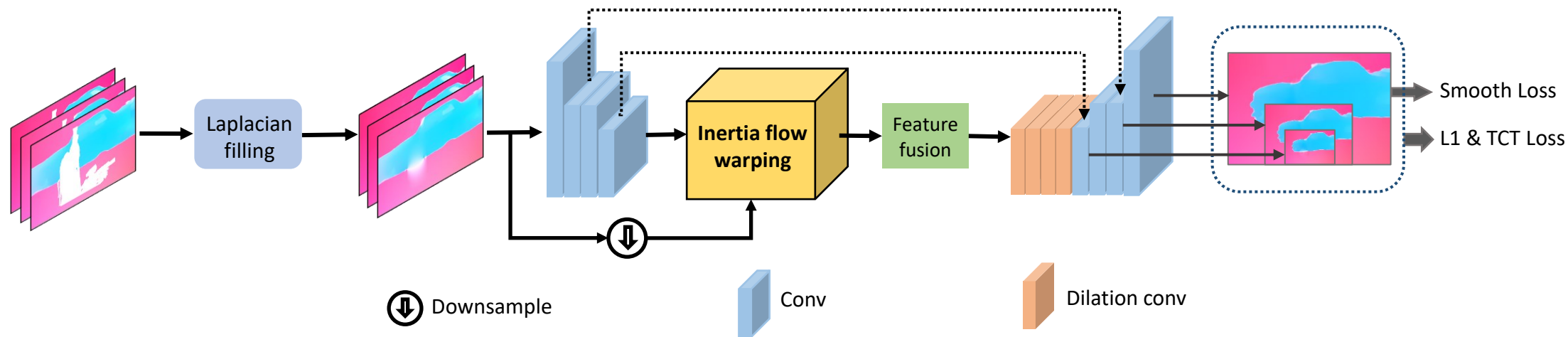


- After flow-guided content propagation, the **style difference** between different frames causes the **spatial inconsistency** between the filled regions and the valid regions. An **extra style correction component** is necessary to eliminate such difference.

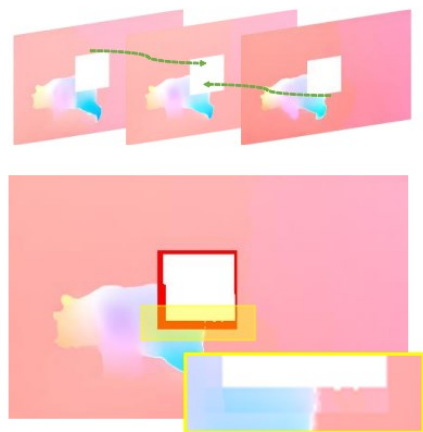
Img	Mean	Std
Bear/00010.jpg	88.27	44.67
Bear/00011.jpg	88.00	44.62
Bear/00021.jpg	86.54	43.10



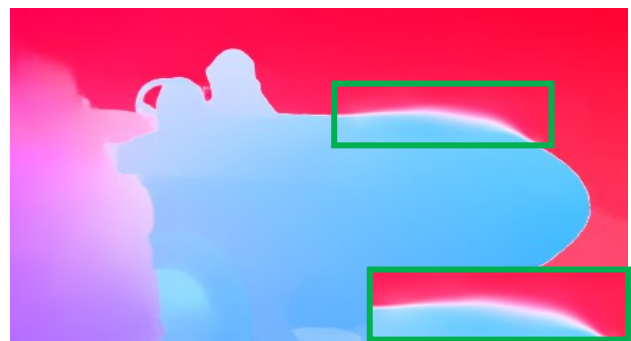
Inertia-Guided Flow Completion Network (IGFC)



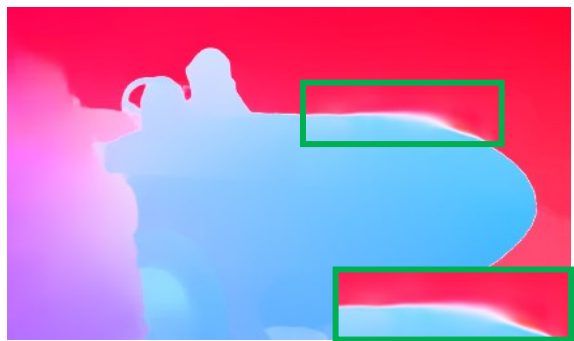
(a) Inertia assumption illustration



(b) Flow warping example



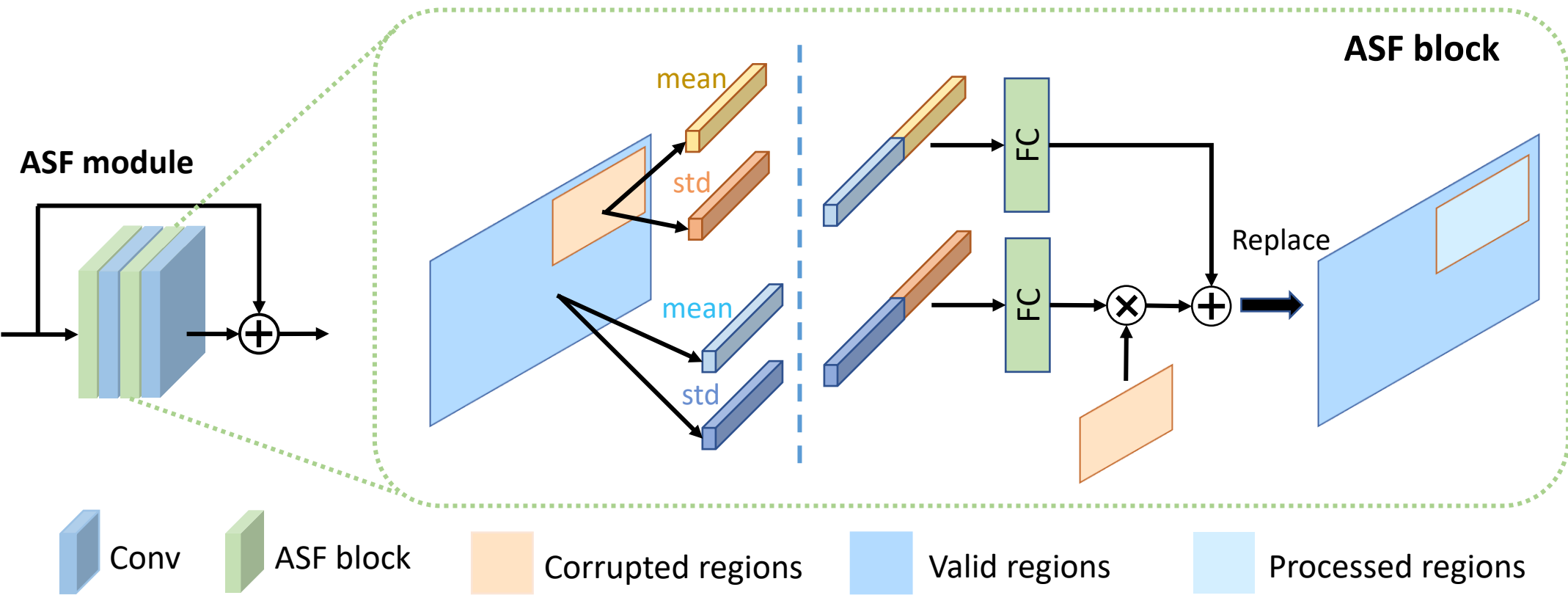
(a) w/o Inertia warp



(a) w/ Inertia warp

Inertia Flow Warping Module

Adaptive Style Fusion Network (ASFN)





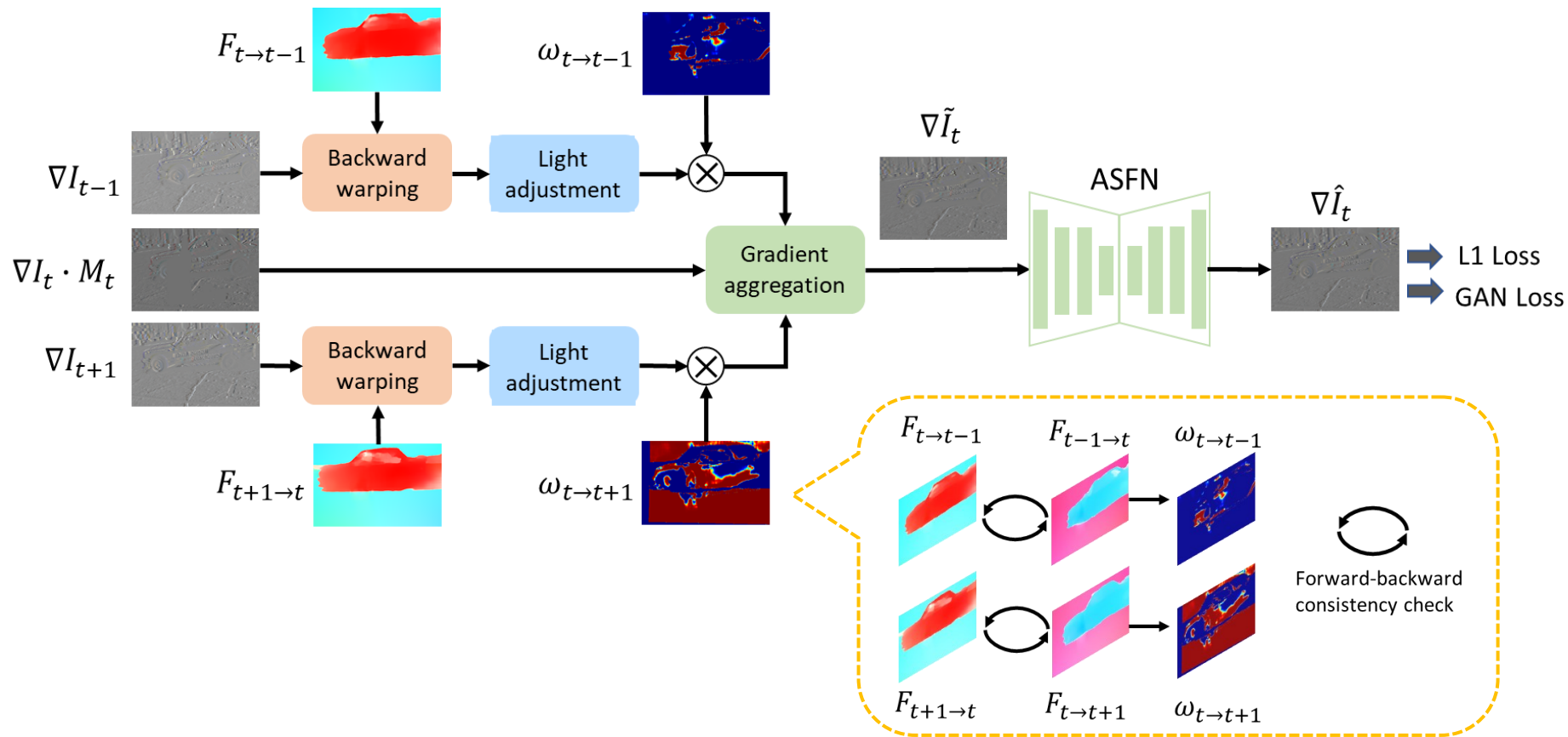
(a) Input

(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

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

3. Ternary census transform loss

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

ASFN

1. Reconstruction loss

$$L_{S_{hole}} = \left\| M_t \odot (\nabla I_t - \nabla \hat{I}_t) \right\|_1 / \|M_t\|_1$$

$$L_{S_{valid}} = \left\| (1 - M_t) \odot (\nabla I_t - \nabla \hat{I}_t) \right\|_1 / \|(1 - M_t)\|_1$$

$$L_{S_{rec}} = L_{S_{hole}} + L_{S_{valid}}$$

2. GAN loss

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

Discriminator loss

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

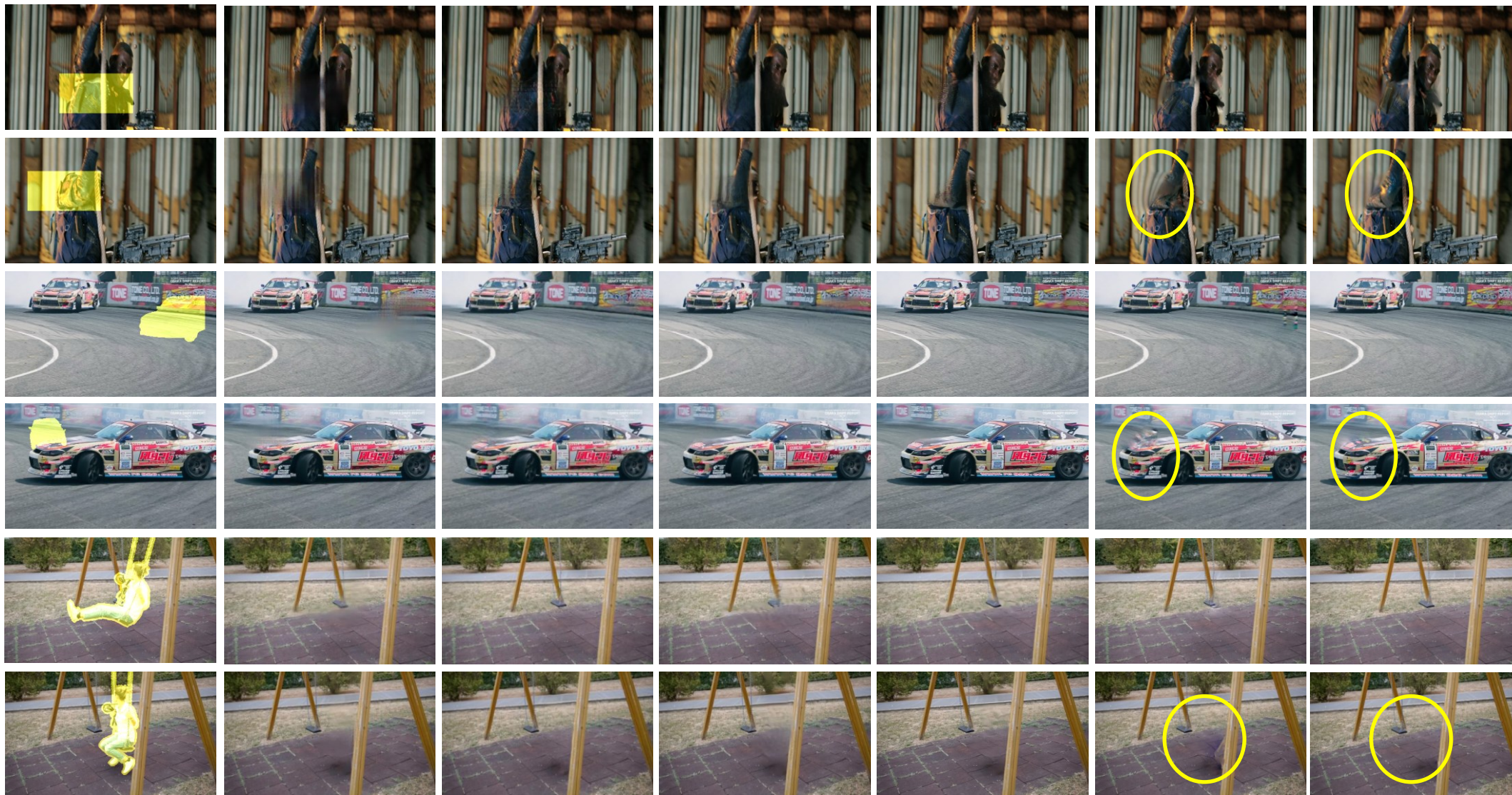
Adversarial loss

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	<u>34.19</u>	0.9510	0.0449	-	-	-
FGVC [8]	<u>33.94</u>	<u>0.9719</u>	<u>0.0259</u>	<u>32.14</u>	<u>0.9667</u>	<u>0.0298</u>	33.91	<u>0.9554</u>	<u>0.0360</u>	<u>34.23</u>	<u>0.9607</u>	<u>0.0345</u>
Ours	<u>34.79</u>	<u>0.9743</u>	<u>0.0225</u>	<u>33.23</u>	<u>0.9729</u>	<u>0.0247</u>	<u>35.16</u>	<u>0.9648</u>	<u>0.0304</u>	<u>35.40</u>	<u>0.9659</u>	<u>0.0303</u>

Introduction | Method | Experiments | Conclusion

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)



Demo video
(Youtube)



Demo video
(Bilibili)



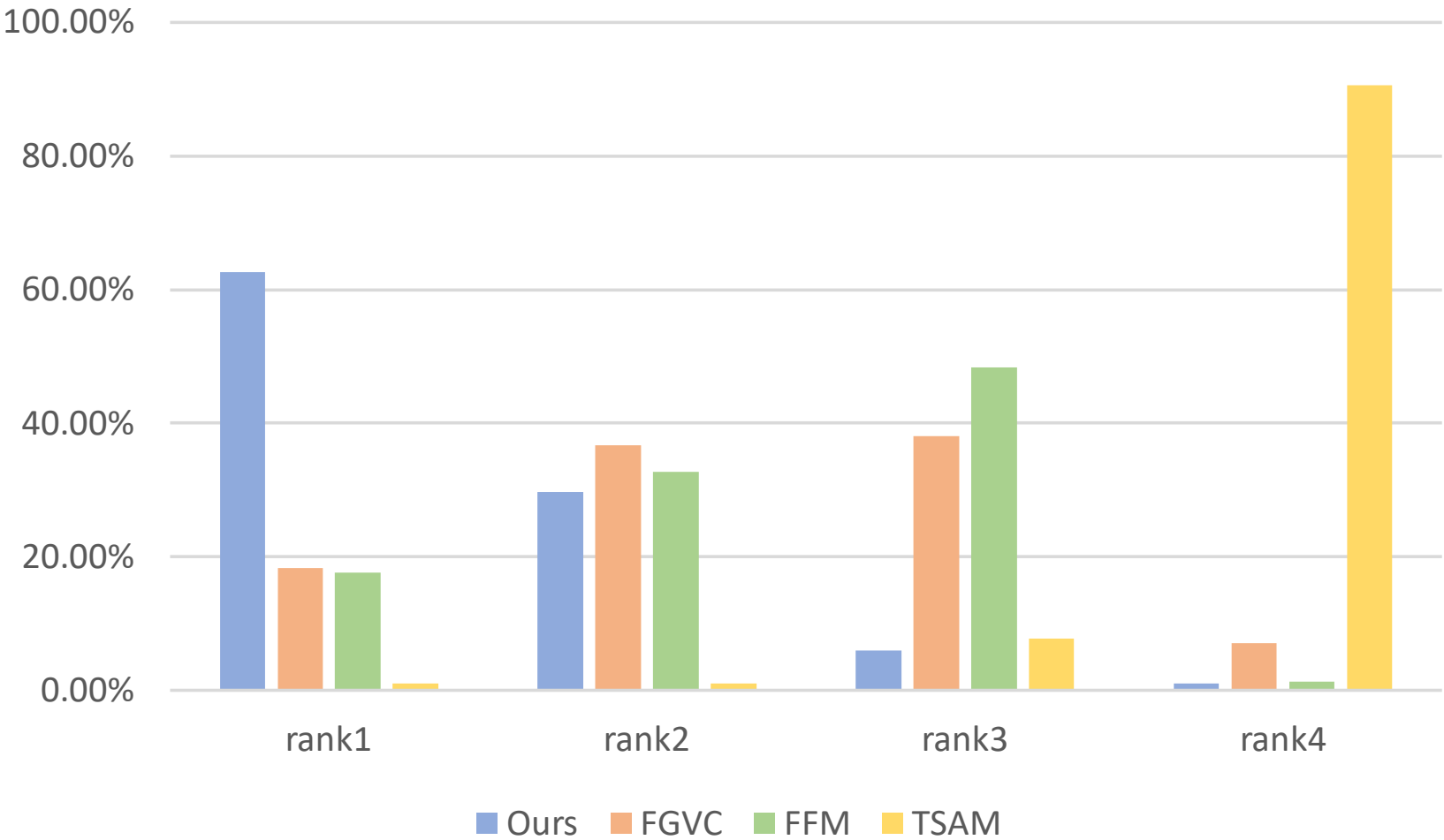
(a) Input

(b) DFGVI

(c) FGVC

(d) Ours

(e) GT



- 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>

