# Paper Reading

Dachun Kai

USTC

October 18, 2021

# XVFI*(ICCV2021, Oral)*

**XVFI: eXtreme Video Frame Interpolation**

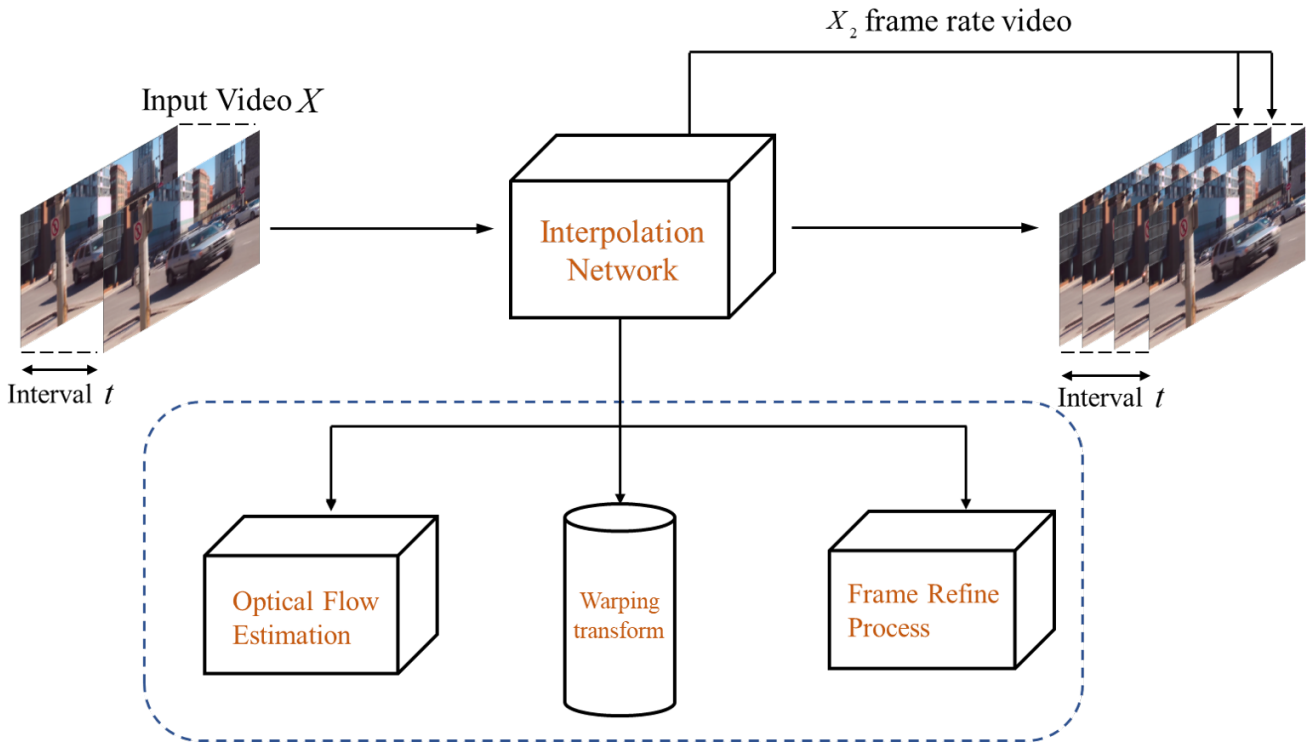
Hyeonjun Sim* Jihyong Oh* Munchurl Kim†

Korea Advanced Institute of Science and Technology

{flhy5836, jhoh94, mkimee}@kaist.ac.kr

# Task: Video Frame Interpolation

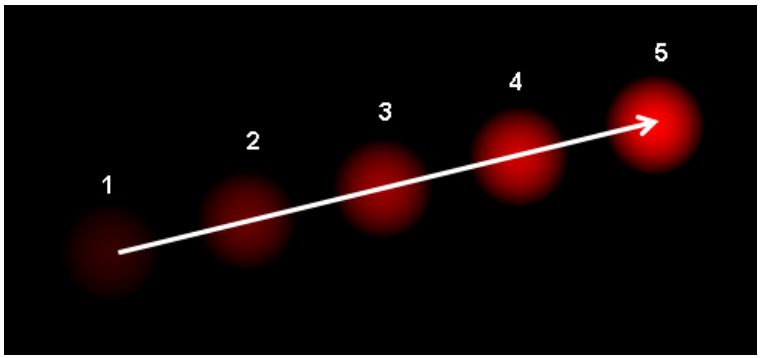
- A low-level vision task to increase frame rate of a video.
- Method: motion-based, kernel-based, phase-based, event-based
- Flow-based method framework:




- Evaluation metric: *PSNR*, *SSIM*, *IE*(Interpolation Error)

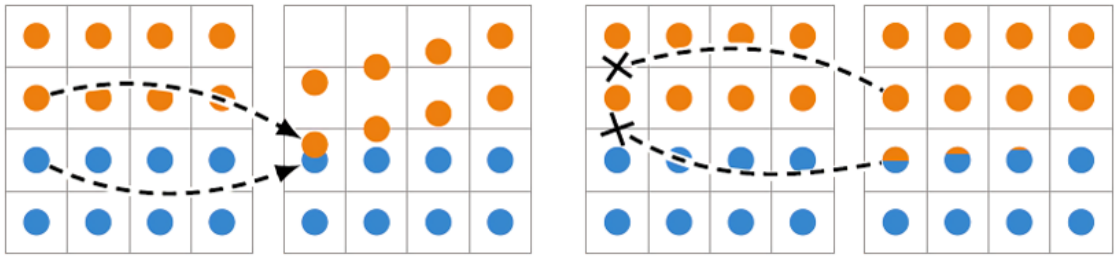
# Preliminaries: Optical Flow and Backward Warping

- Optical Flow depicts motion of objects in a visual scene.

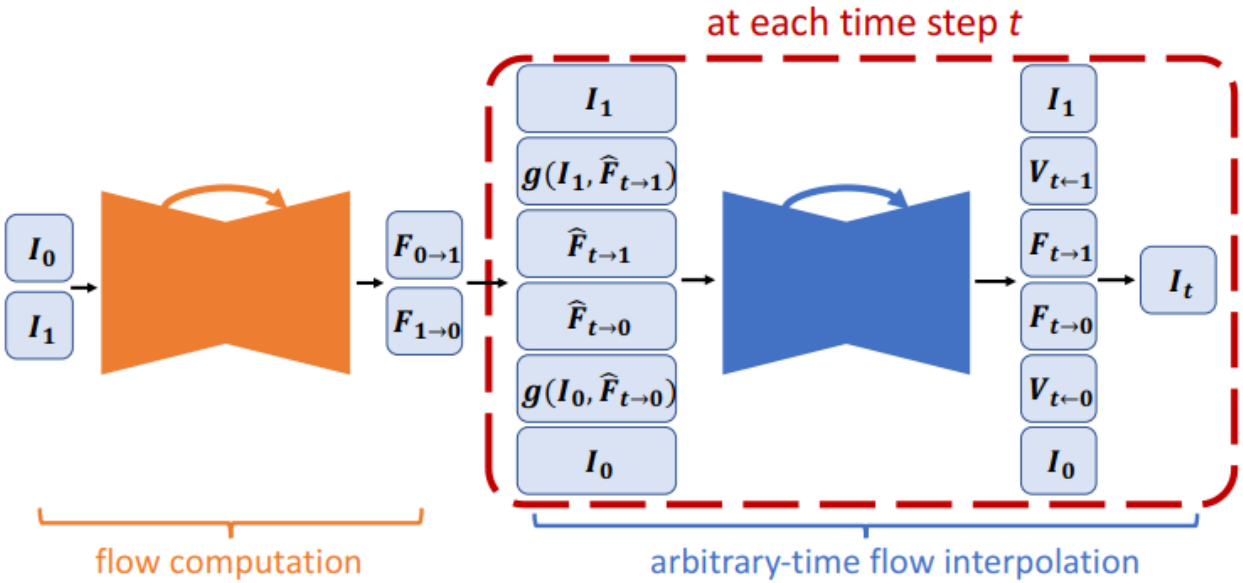



- Flow Example:

- Image Warping



# Milestone: Super SloMo[1]




## Ideas

$$\hat{F}_{t\to0} = -(1-t)tF_{0\to1} + t^2F_{1\to0}$$

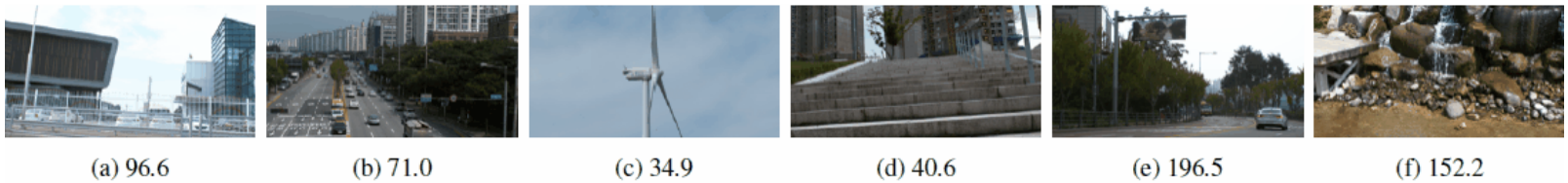
$$\hat{F}_{t\to1} = (1-t)^2F_{0\to1} - t(1-t)F_{1\to0}$$

$$\hat{I}_t = \frac{1}{Z} \odot \left((1-t)V_{t\leftarrow0} \odot g\left(I_0, F_{t\to0}\right) + tV_{t\leftarrow1} \odot g\left(I_1, F_{t\to1}\right)\right)$$

[1] Huaizu Jiang et al. "Super slomo: High quality estimation of multiple intermediate frames for video interpolation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 9000–9008.

# Problem Definition

- Motivation
  - Handle the VFI for $4K$ videos with large motion, computed by IRR[2].



(a) 96.6 (b) 71.0 (c) 34.9 (d) 40.6 (e) 196.5 (f) 152.2

[2] Junhwa Hur and Stefan Roth. "Iterative residual refinement for joint optical flow and occlusion estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5754–5763.

# Proposed X4K1000FPS Dataset

- Impressive $4K@1000fps$ dataset.


- Various objects: crowds, cars, trains, plants, boats$\cdots$
- Various places: stadiums, stations, beaches, rivers$\cdots$

# Proposed X4K1000FPS Dataset

- Impressive $4K@1000fps$ dataset.

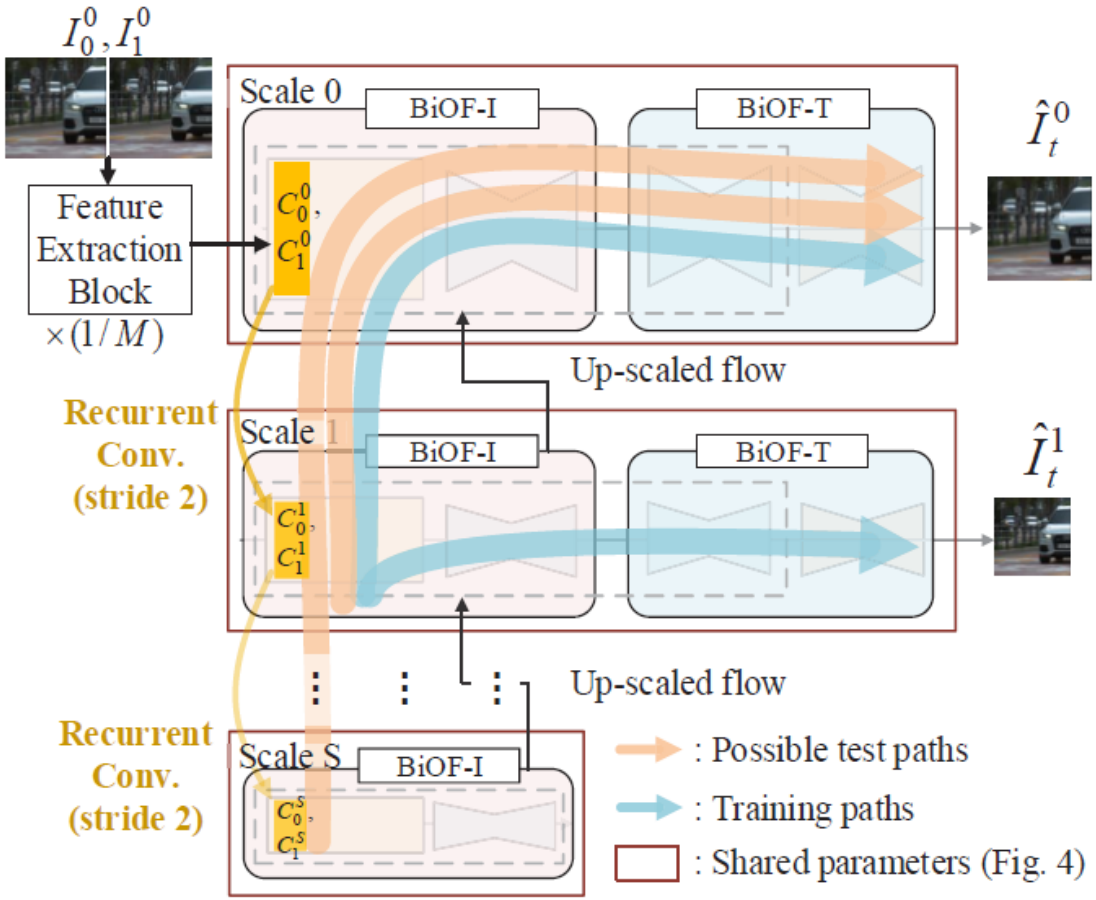
| Dataset | Occlusion [16] | | | Flow magnitude [16] | | |
|---|---|---|---|---|---|---|
| | $25_{th}$ | $50_{th}$ | $75_{th}$ | $25_{th}$ | $50_{th}$ | $75_{th}$ |
| Vimeo90K [48] | 6.8 | **11.9** | **18.1** | 3.1 | 4.9 | 7.1 |
| Adobe240fps [39] | 0.8 | 1.7 | 3.2 | 3.8 | 8.9 | 16.3 |
| X-TEST (ours) | 2.1 | 5.6 | 17.7 | **23.9** | **81.9** | **138.5** |
| X-TRAIN (ours) | **6.9** | 10.1 | 15.7 | 5.5 | 18.0 | 59.5 |

$25_{th}$, $50_{th}$ and $75_{th}$ represent percentiles of each datset.

- Detailedly analysis: occlusion, flow magnitude
- Select extreme scenes for X-TRAIN and X-TEST.

# Proposed Method

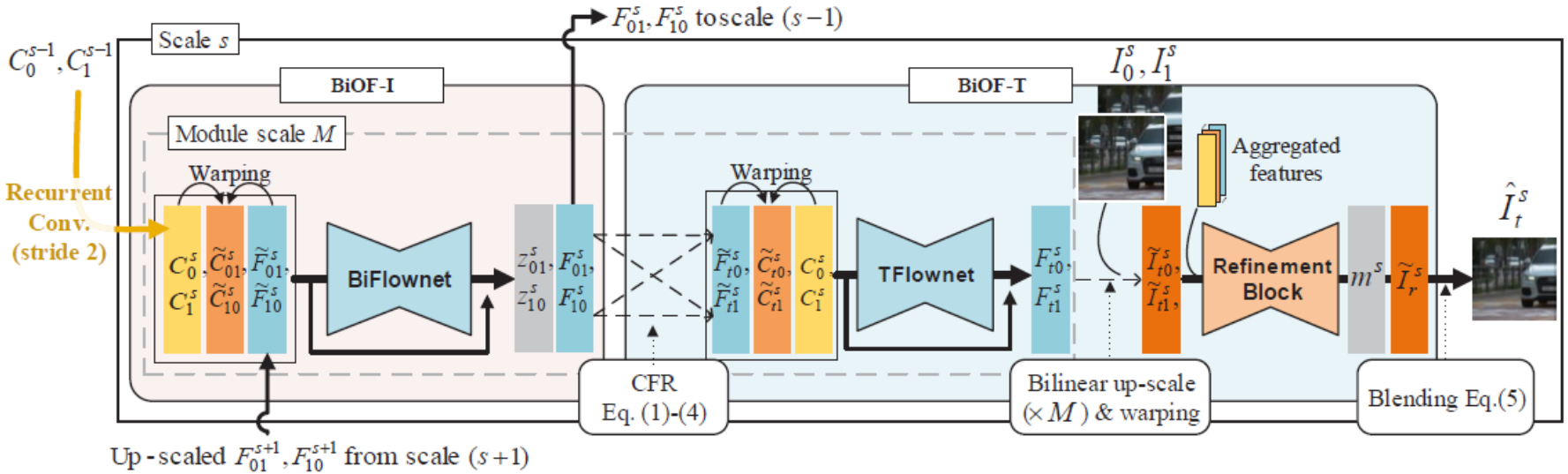
- Feature Extraction: strided conv + residual block




- Scale Adaptivity: Shared parameters, adapt to resolution of $I_0$, $I_1$

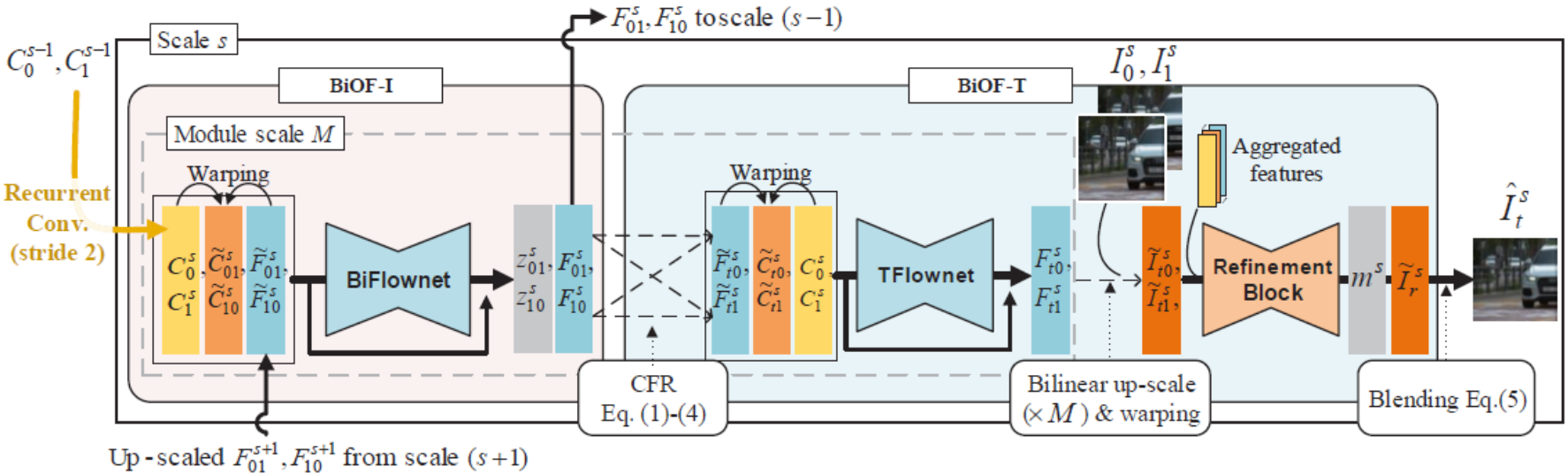
# Proposed Method

- Arch of XVFI-Net in scale s.



# Proposed Method

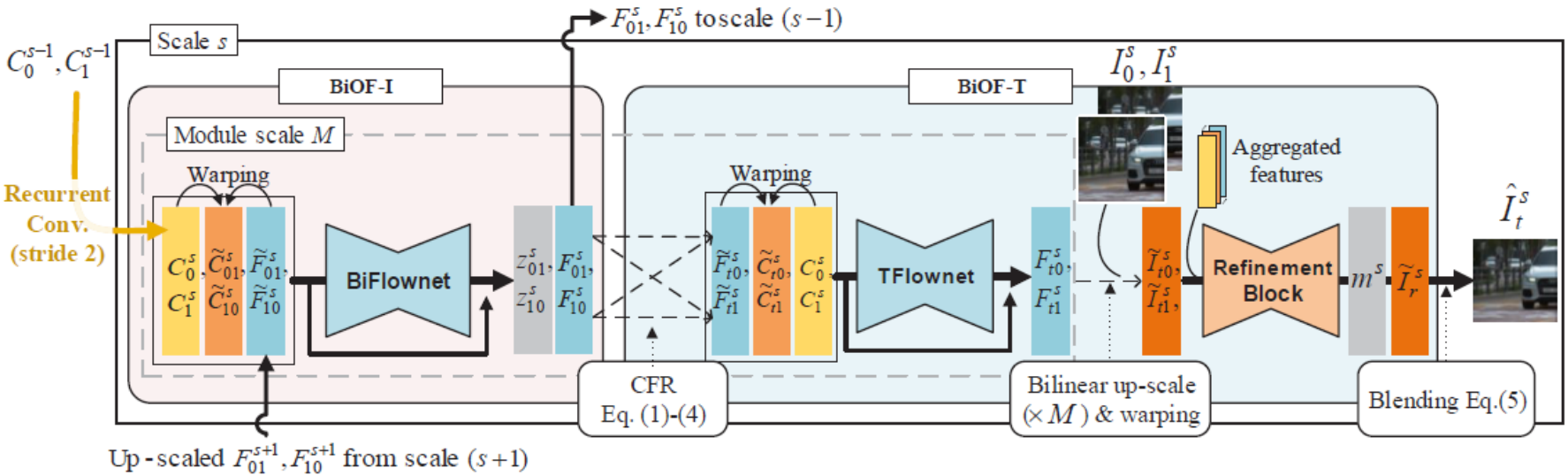
- Arch of XVFI-Net in scale s.




- BIOF-I module
  - Input: context info $C_0^{s-1}$, $C_1^{s-1}$, $s+1$-scale Flow $F_{01}^{s+1}$, $F_{10}^{s+1}$
  - Output: $s$-scale Flow $F_{01}^s$, $F_{10}^s$

# Proposed Method

- Arch of XVFI-Net in scale s.



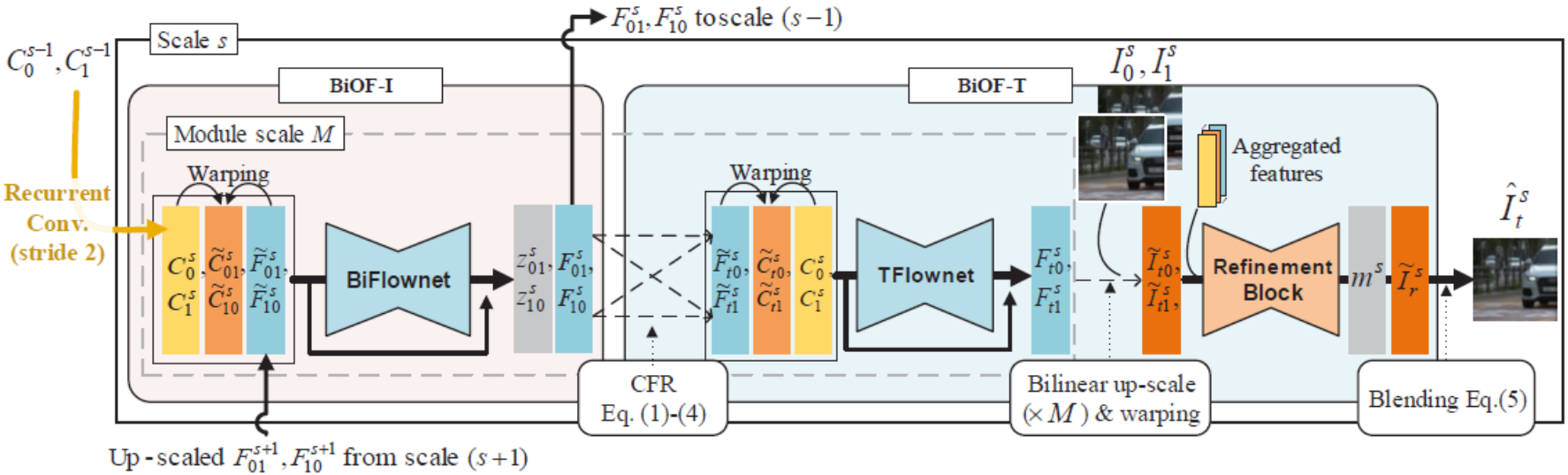

- BIOF-T module:

$$\tilde{F}_{t0}^{x} = \frac{(1-t)\sum_{\mathcal{N}_0} w_0 \cdot (-F_{0t}^{y}) + t\sum_{\mathcal{N}_1} w_1 \cdot F_{1\cdot(1-t)}^{y}}{(1-t)\sum_{\mathcal{N}_0} w_0 + t\sum_{\mathcal{N}_1} w_1} \quad (1)$$

$$\tilde{F}_{t1}^{x} = \frac{(1-t)\sum_{\mathcal{N}_0} w_0 \cdot F_{0\cdot(1-t)}^{y} + t\sum_{\mathcal{N}_1} w_1 \cdot (-F_{1t}^{y})}{(1-t)\sum_{\mathcal{N}_0} w_0 + t\sum_{\mathcal{N}_1} w_1} \quad (2)$$

# Proposed Method

- Arch of XVFI-Net in scale s.




- BIOF-T module:

$$\mathcal{N}_0 = \left\{ \mathrm{y} \mid \ \mathrm{round}\ \left( \mathrm{y} + F_{0t}^{\mathrm{y}} \right) = \mathrm{x} \right\} \tag{3}$$

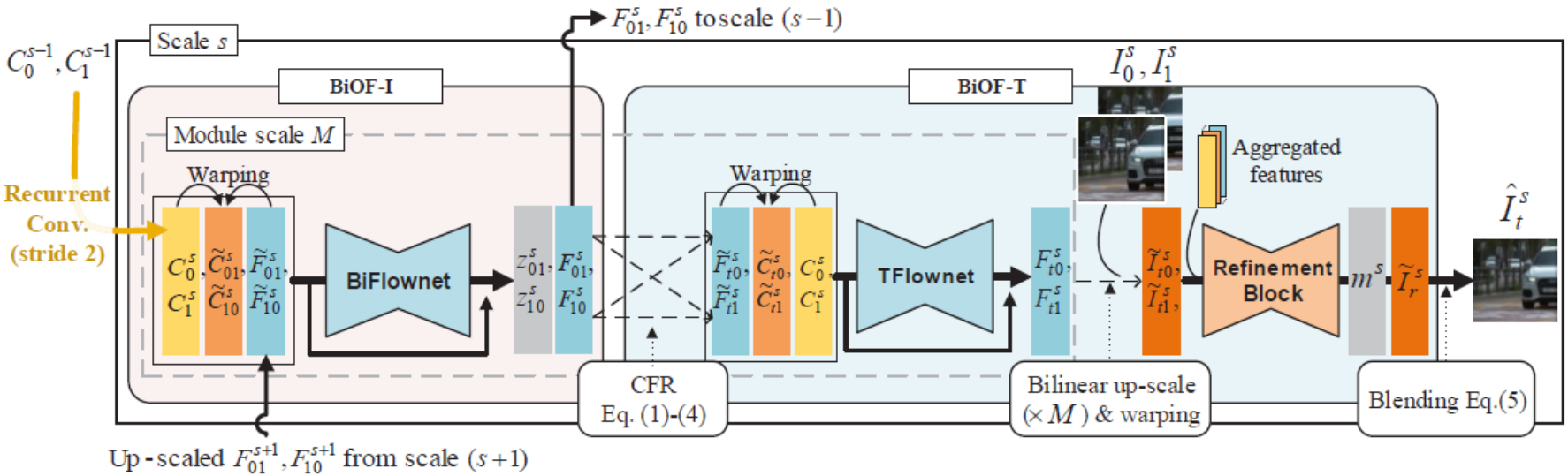
$$\mathcal{N}_1 = \left\{ \mathrm{y} \mid \ \mathrm{round}\ \left( \mathrm{y} + F_{1t}^{\mathrm{y}} \right) = \mathrm{x} \right\} \tag{4}$$

$$w_i = z_i^{\mathrm{y}} \cdot G\left( \left| \mathrm{x} - \left( \mathrm{y} + F_{it}^{\mathrm{y}} \right) \right| \right) \tag{5}$$

$$G(d) = e^{-d^2/\sigma^2} \tag{6}$$

# Proposed Method

- Arch of XVFI-Net in scale s.

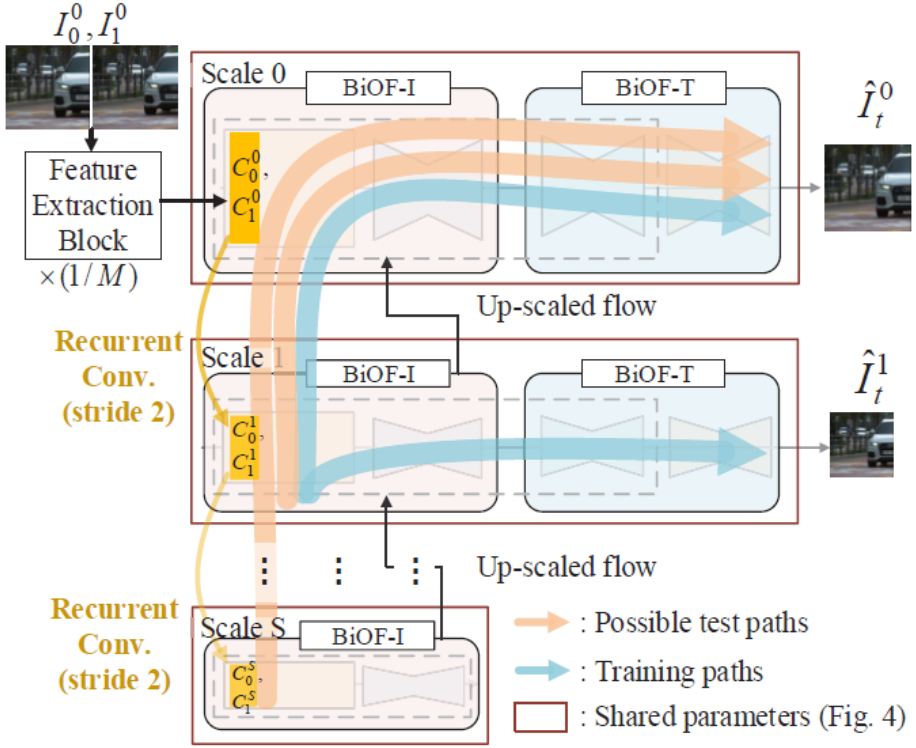



- BIOF-T module:

$$\hat{I}_t^s = \frac{(1-t)\cdot m^s \cdot \tilde{I}_{t0}^s + t\cdot(1-m^s)\cdot \tilde{I}_{t1}^s}{(1-t)\cdot m^s + t\cdot(1-m^s)} + \tilde{I}_r^s$$

where $m^s$ is occlusion mask and $\tilde{I}_r^s$ is residual image.

# Proposed Method

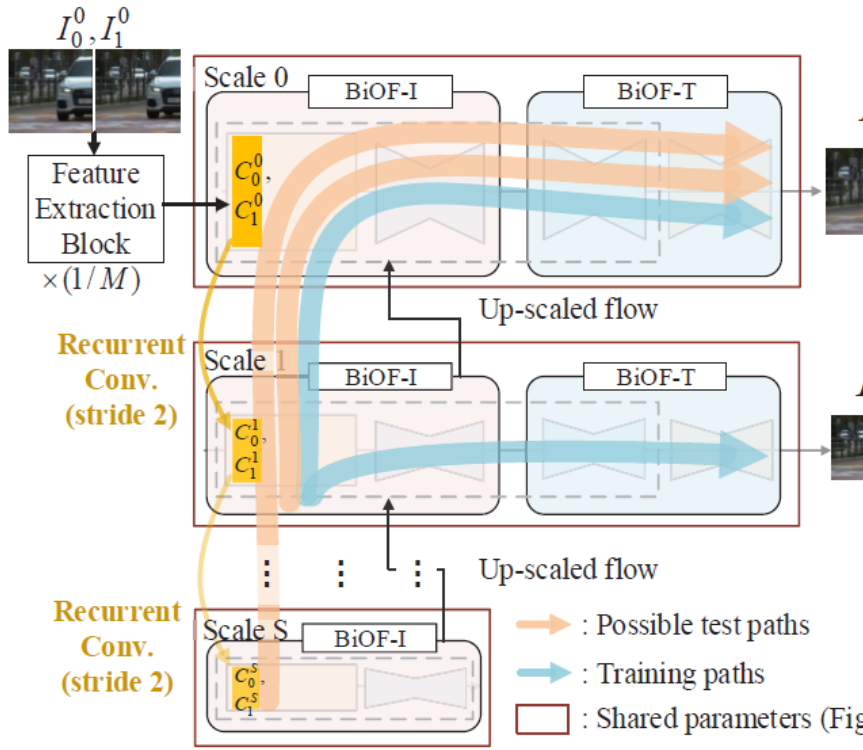



## Loss Function

- $\mathcal{L}_{\text{total}} = \mathcal{L}_r + \lambda_s \cdot \mathcal{L}_s$
- Multi-scale reconstruction loss: $\mathcal{L}_r = \sum_{s=0}^{S_{trn}} \left\| \hat{I}_t^s - I_t^s \right\|_1$
- Edge-aware smoothness loss: $\mathcal{L}_s = \sum_{i=0,1} \exp\left(-e^2 \sum_c \left|\nabla_x I_{t_c}^0\right|\right)^\top \cdot \left|\nabla_x F_{ti}^0\right|$

# Model Analysis

- Shared parameters→lightweighted.
- Inference can start from any scale level.
- Coarse-to-fine strategy, but start from low spatial resolution.
- Interpolate a frame at any time instance from two input frames.



# Comparison to SOTA

- Settings
  - Retrain SOTA on X-TRAIN with sub $f$, original with sub $o$.
  - XVFI-Net train($S_{trn} = 3$) and test($S_{tst} = 3$ *or* 5).
- Results:

| Methods (×N) | X-TEST (PSNR/SSIM/tOF) | Adobe240fps (PSNR/SSIM) | #P ↓ | $R_t$ ↓ |
|---|---|---|---|---|
| AdaCoF$_o$ (×5.8) | 23.90/0.727/6.89 | 25.26/0.785 | 21.8 | **0.005** |
| AdaCoF$_f$ [25] | 25.81/0.772/6.42 | 25.21/0.791 | 21.8 | **0.005** |
| FeFlow$_o$ (×5.3) | 24.00/0.756/6.59 | 25.18/0.785 | 102.5 | 1.681 |
| FeFlow$_f$ [13] | 25.16/0.783/6.54 | 24.17/0.780 | 102.5 | 1.681 |
| DAIN$_o$ (×9.3) | 26.78/0.807/3.83 | 29.89/0.911 | 24 | 1.375 |
| DAIN$_f$ [4] | 27.52/0.821/3.47 | 29.99/0.910 | 24 | 1.375 |
| Ours ($S_{tst}$=3) | 28.86/0.858/2.67 | **30.29/0.912** | **5.5** | 0.074 |
| Ours ($S_{tst}$=5) | **30.12/0.870/2.15** | 30.18/0.911 | **5.5** | 0.075 |

×N: The ratio of number of iterations of the original version to that of -retrained version in the fair condition. #P: The number of parameters (M).
$R_t$: The runtime on 1024×1024-sized frames in sec.
**RED**: Best performance, BLUE: Second best performance.

Table 2. Quantitative comparisons on both X-TEST (4K) and Adobe240fps (HD) [39] for multi-frame interpolation (×8).

# Comparison to SOTA

- Settings
  - Retrain SOTA on X-TRAIN with sub $f$, original with sub $o$.
  - XVFI-Net train($S_{trn} = 3$) and test($S_{tst} = 3$ *or* 5).
- Results:

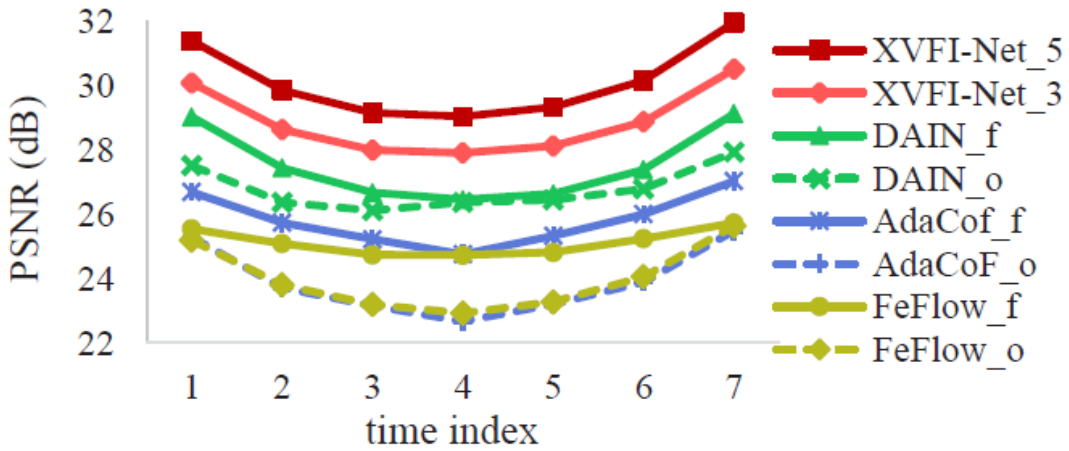



Figure 5. PSNR profiles for multi-frame interpolation results (×8) on X-TEST.

# Comparison to SOTA

- Qualitive Results:

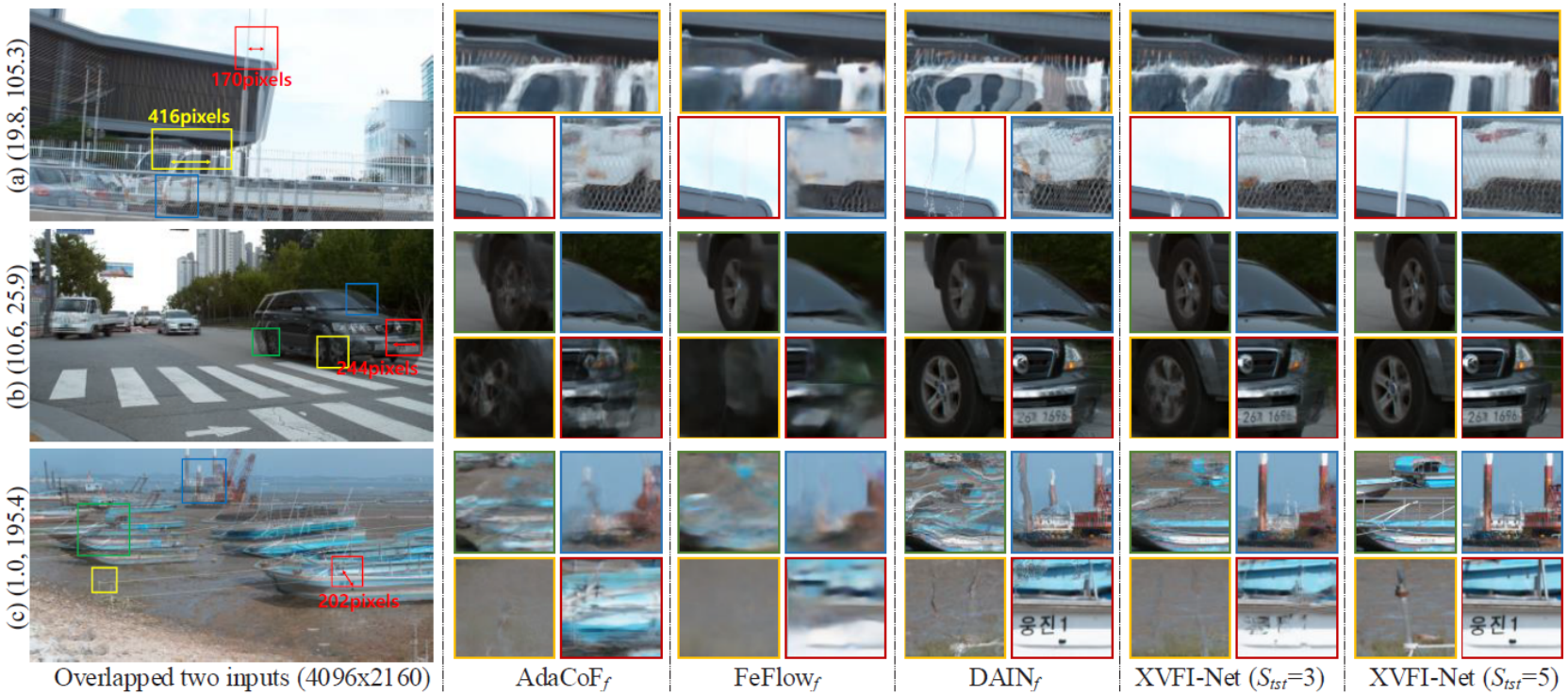
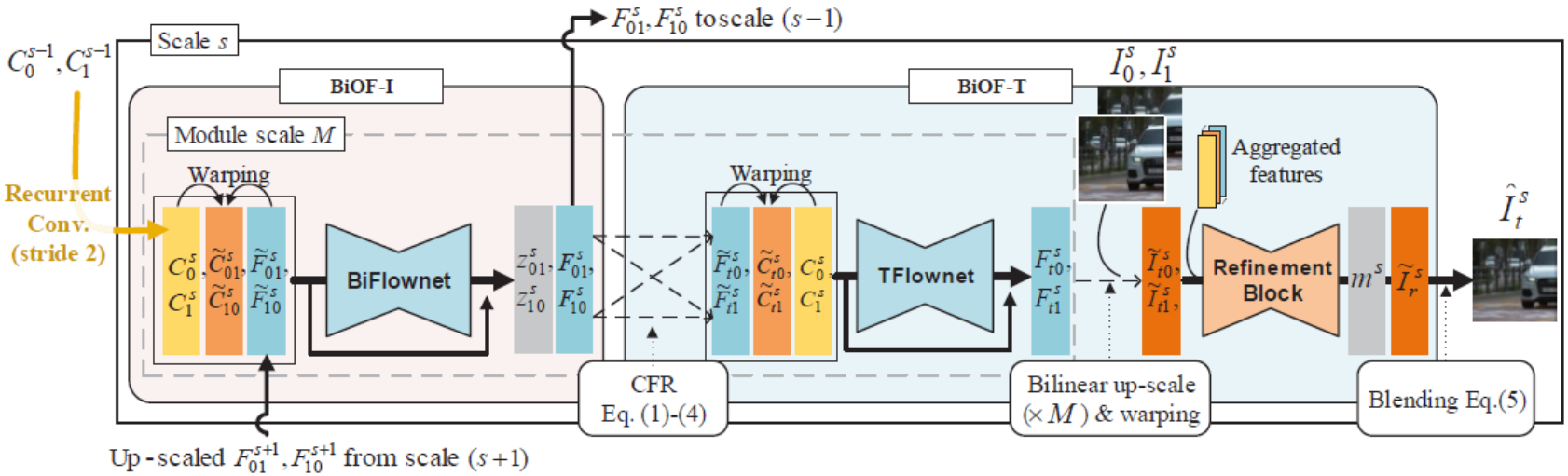



Figure 6. Visual comparisons for VFI results ($t = 0.5$) on X-TEST for our and retrained SOTA methods with X-TRAIN. (*,*): occlusions and optical flow magnitudes between the two input frames measured by [16], respectively. *Best viewed in zoom.*

# Ablation Studies



# Ablation Studies

- Flow Approximation

| Methods \ Metrics | EPE↓ | PSNR↑ | SSIM↑ | tOF↓ |
|---|---|---|---|---|
| (a) Linear comb. | 0.0752 | 28.73 | 0.8518 | 2.89 |
| (b) Flow reversal | 0.0892 | 28.30 | 0.8425 | 2.98 |
| (c) CFR (ours) | **0.0721** | **28.86** | **0.8582** | **2.67** |

- Adjustable scalability

| $S_{trn}$ \ $S_{tst}$ | (PSNR(dB)↑ / SSIM[44]↑ / tOF[8]↓) | | |
|---|---|---|---|
| | 1 | 3 | 5 |
| 1 | 26.85/0.806/4.90 | **28.40/0.852/3.46** | 27.14/0.842/3.69 |
| 3 | 23.61/0.729/6.56 | 29.22/0.863/2.68 | **30.35/0.879/1.98** |
| 5 | 22.37/0.699/6.71 | 23.70/0.724/6.39 | **29.48/0.864/2.08** |

**RED**: Best performance of each row

# Ablation Studies

- Robustness on LR-LFR dataset

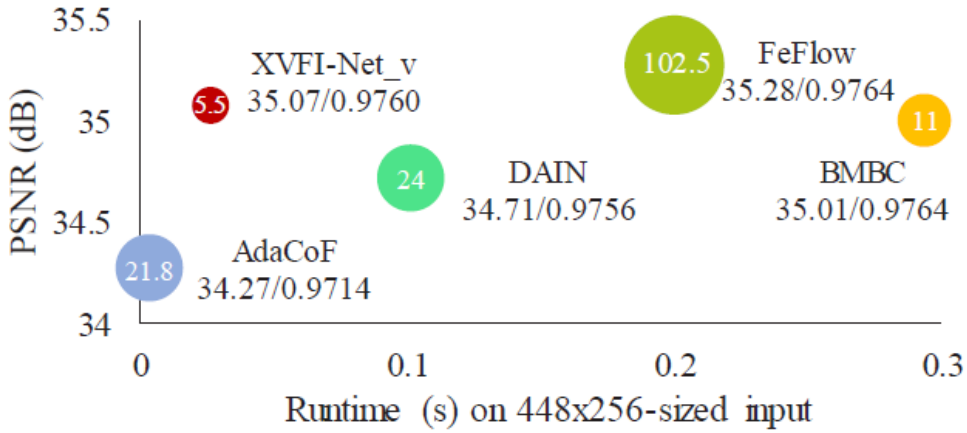



Figure 8. PSNR/SSIM vs runtime (s) on Vimeo90K [48] with model size (M) indicated in each circle.

# Failure cases

- Cases on 4K patches, with very large flow magnitude(196.5)

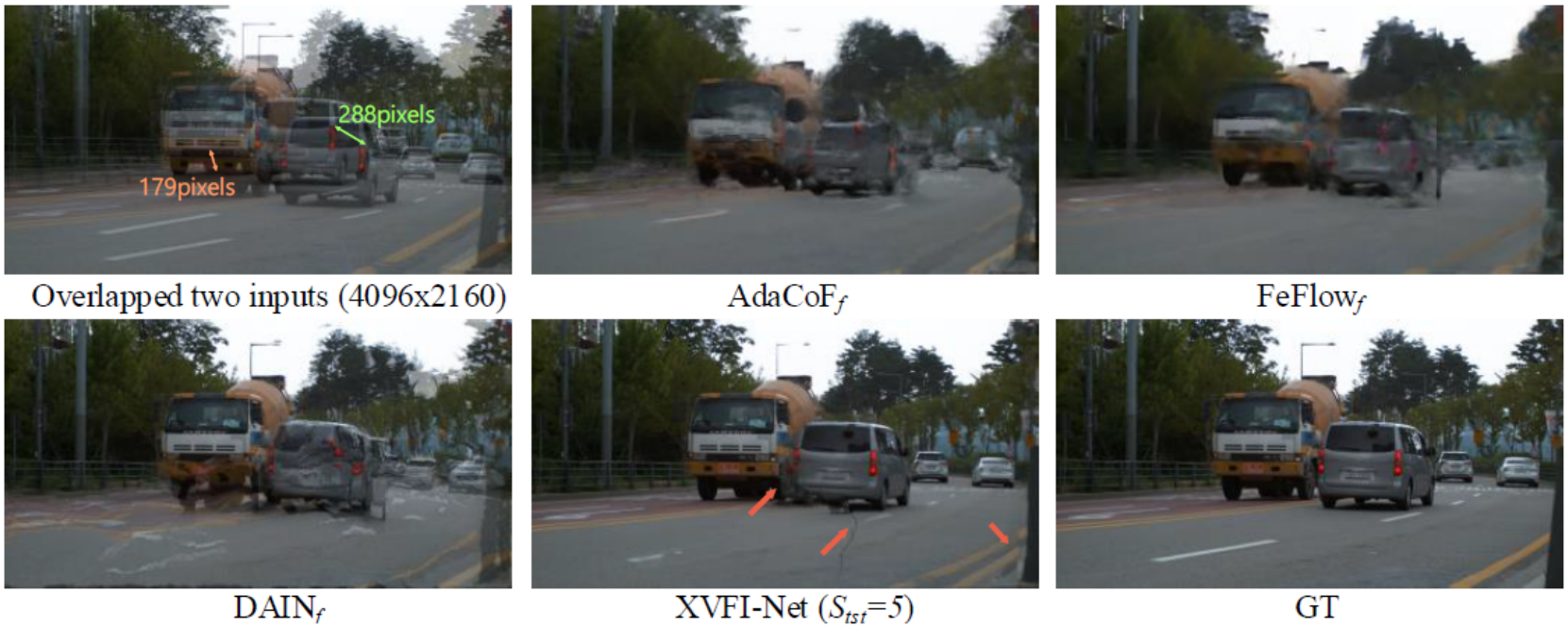



Figure 6. Failure cases of 4K result ($t = 0.5$) on X-TEST for our and *retrained* SOTA methods with X-TRAIN, including the corresponding ground truth. *Best viewed in zoom.*

# Failure cases

- Cases on croped patches


Figure 7. Failure cases of cropped results ($t = 0.5$) on X-TEST for our and *retrained* SOTA methods with X-TRAIN, including the corresponding ground truth. *Best viewed in zoom.*

- Analysis: attribute to flow and warping algorithm.

# Conclusion

- Contributed HR-HFR X4K1000FPS dataset is very valuable for research.
- XVFI-Net showed SOTA performance on HR dataset, also robust to LR-LFR dataset.
- Extend VFI for more recent real-world applications with HR video.

*Thanks, Q & A*