

Supplementary for JFFRA : Joint Flow And Feature Refinement Using Attention For Video Restoration

Ranjith Merugu²

ranjith.merugu@stonybrook.edu

Mohammed Sameer Suhail^{* 1}

mo.suhail@samsung.com

Akshay P Sarashetti^{* 1}

akshay.p@samsung.com

Venkata Bharath Reddy Reddem^{* 1}

r.reddy@samsung.com

Pankaj Kumar Bajpai¹

pankaj.b@samsung.com

Amit Satish Unde¹

amit.unde@samsung.com

In this supplementary material, we have included more details about design choices, ablation study, visual representation of mutual refinement, evaluation metric that was used to check the effectiveness of temporal loss, task-wise performance and complexity comparison, impact of different optical flow computation methods, and visual results.

1. Joint Flow and Feature Refinement

Our approach introduces a robust joint and iterative refinement mechanism for video restoration and optical flow estimation. Unlike traditional methods, we leverage a cost volume to iteratively refine optical flow, ensuring it adapts dynamically to the evolving frame features and vice versa. Simultaneously, JFFR module enhances the video features using attention mechanisms followed by convolution blocks, allowing the network to focus on critical spatio-temporal details. This tightly coupled refinement pipeline ensures that both flow estimation and feature enhancement mutually benefit each other at every iteration, resulting in superior restoration quality, improved temporal consistency, and flow accuracy. This unique integration of cost volumes and attention mechanisms and joint iterative refinement of flow and features distinguishes our method from existing approaches. More visual results can be found below.

2. Distinctive Design Choices of JFFRA

Design Uniqueness and Justification. Table 1 highlights that JFFRA is the only framework to unify all critical components needed for robust video restoration. Unlike BasicVSR++ [2] and EDVR [13], which either lack joint refinement or tightly coupled flow-feature feedback, JFFRA performs explicit, iterative refinement of both flow and features, ensuring continual correction of alignment and restoration errors. While transformer-based methods

like VRT [7] and RVRT [6] leverage attention and multi-scale processing, they do not implement joint flow-feature refinement or cost-volume-based flow updates, and thus are less effective in handling large or complex motion. FMA-Net [16] introduces joint refinement, but its reliance on memory modules makes it fragile and computationally heavy. Specifically, FMA-Net does not perform truly iterative joint refinement at every scale; instead, it relies on separate feature propagation and motion modules without explicit progressive feedback updating. Its scale-wise coupling remains limited, lacking an explicit mechanism to progressively align and refine both flow and features across coarse-to-fine stages, and it lacks both attention mechanisms and occlusion-aware loss. TOFlow [15] and similar approaches use flow for alignment but do not iteratively refine flow and features together, nor do they address temporal flicker or occlusion. In contrast, JFFRA uniquely combines mutual, scale-wise iterative refinement, attention-guided alignment, cost-volume-based flow correction, and an explicit occlusion-aware temporal loss, all within a lightweight design that avoids memory bottlenecks. This comprehensive integration directly addresses the limitations seen in all previous methods, resulting in superior robustness, temporal consistency, and restoration quality across diverse and challenging video scenarios.

3. Complexity vs. Performance Analysis

Our method achieves significantly better performance across multiple restoration tasks and even outperforms task-specific SOTA, as reported in Table 2 and Table 3. For video denoising, it can be seen that the proposed JFFRA achieves the best quality with lower complexity than ShiftNet [5] and VRT [7]. This performance is consistent even in the video deblurring task, and the proposed method sets a new standard while computationally being lower complex than ShiftNet [5], VRT [7], and BSSNet [11].

^{*}Equal contribution

¹Samsung R&D Institute India, Bangalore

²Stony Brook University

Table 1. Comparison of JFFRA with prior video restoration methods. Only JFFRA integrates all key components including iterative refinement, attention, occlusion-aware loss, and lightweight design without memory.

Method	Joint Flow-Feature Refinement	Attention Mechanism	Occlusion-Aware Loss	No Memory / Recurrent Modules	Bidirectional Flow Update	Flow-Feature Sync	Lightweight Design
BasicVSR++ [2]	×	×	×	✓	✓	×	✓
EDVR [13]	×	×	×	✓	×	×	×
FMA-Net [16]	✓	×	×	×	×	✓	×
TOFlow [15]	×	×	×	✓	×	×	✓
VRT [7]	×	✓	×	✓	✓	×	×
RVRT [6]	×	✓	×	×	✓	×	×
JFFRA (Ours)	✓	✓	✓	✓	✓	✓	✓

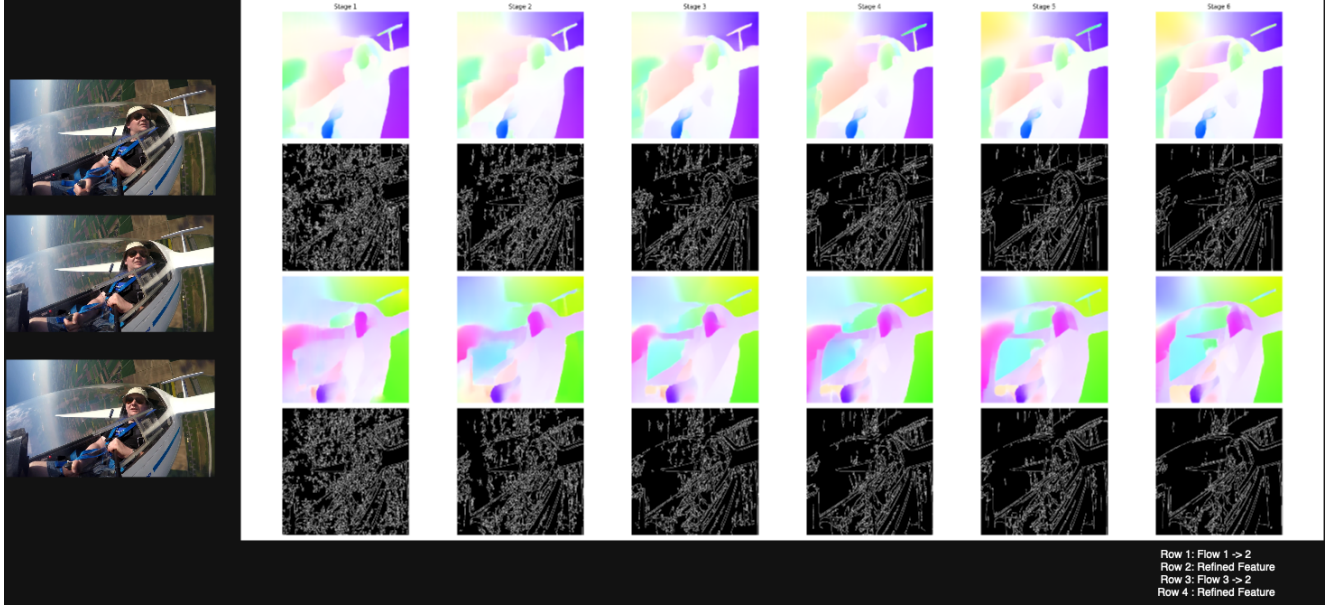


Figure 1. Visualization of progressive joint flow and feature refinement in JFFRA. This figure illustrates the step-by-step refinement process at different scales across multiple JFFR blocks, using the DAVIS dataset with added noise ($\sigma = 30$) to demonstrate robustness. Row 1 shows intermediate flow from frame 1 to frame 2. Row 2 depicts the corresponding refined feature maps after feature alignment. Row 3 shows flow from frame 3 to frame 2. Row 4 presents the refined features guided by these flows. This visualization highlights how JFFRA’s iterative feedback mechanism progressively improves both motion alignment and feature consistency, ultimately enabling sharper, temporally coherent video restorations even in the presence of significant noise.

Table 2. Performance and complexity analysis of various methods for video denoising on DAVIS dataset ($\sigma = 30$) [4] with patch size of 256×256 .

Method	Publication	Params (M)	GFLOPs	PSNR
Tempformer [9]	ECCV’22	22.3	185.2	35.66
RVRT [6]	NeurIPS’22	13.6	88.8	36.57
ShiftNet [5]	CVPR’22	12.7	154.3	36.83
VRT [7]	ECCV’22	18.3	143.2	36.52
Ours	-	22.3	129.9	37.04

Table 3. Performance and complexity analysis of various methods for video deblurring on DVD [10] (256×256 patch).

Method	Publication	Params (M)	GFLOPs	PSNR
RVRT [6]	NeurIPS’22	13.6	88.8	34.30
ShiftNet [5]	CVPR’23	12.7	154.3	34.69
VRT [7]	TIP’24	18.3	143.2	34.27
BSSNet [11]	CVPR’21	13.0	133.0	34.95
Ours	-	22.3	129.9	35.15

4. Analysis with Various Flow Estimation Methods

We demonstrate in Table 4 the performance of the proposed method by using different initial optical flow estimation methods during inference for video super-resolution. It can be seen that the performance using lightweight GMFlow [14] is similar to RAFT [12], illustrating the stability

of the proposed design.

Table 4. Analysis with different flow estimation methods.

Metric	IRR-PWC [3]	GMFlow [14]	RAFT [12]
PSNR / SSIM	31.76 / 0.879	32.36 / 0.912	32.45 / 0.9136
GMac / Params	17.3 / 6.3	2.3 / 4.6	12.9 / 5.2

5. Cost Volume Effectiveness

In this section, we present an ablation study on varying cost volume sizes within the JFFR block. Increasing the cost volume size improves reconstruction quality but also incurs higher computational overhead, resulting in longer runtimes.

Table 5. Ablation study on different cost volume sizes for JFFRA validation. Results demonstrated below on DVD [10]

Cost Volume Size	PSNR	Runtime (ms)
4×4	30.42	245
6×6	32.63	410
8×8	35.15	623
10×10	35.17	934

6. Qualitative Analysis with SOTA

In this section, we present visual evidence of the quality of JFFRA against task-specific SOTA, across video restoration tasks like video denoising, video deblurring, and video super-resolution. We clearly see JFFRA outperforming SOTA in noise removal, texture preservation, extracting sharper results, fine details, and texture recovery.

7. Limitations and Future Scope

Although JFFRA demonstrates strong performance across various video restoration tasks such as deblurring, denoising, and super-resolution, it has certain limitations. In particular, when dealing with extremely noisy inputs (for example, in the Set-8 dataset with noise levels above $\sigma = 40$), its ability to correct and improve results is reduced. This reduction is mainly due to the degradation of initial flow and feature estimates, which makes the iterative refinement less effective. At present, JFFRA does not explicitly handle noise uncertainty or incorporate specialized noise modeling, which limits its robustness under severe degradation. However, the overall design remains flexible and offers opportunities for further improvement. Future work may focus on integrating noise-aware modules, uncertainty modeling, or adaptive noise modulation strategies to enhance robustness in highly challenging scenarios. Additionally, exploring the incorporation of generative priors and advanced regularization techniques could further improve generalization and performance in real-world applications.

References

- [1] Pablo Arias and Jean-Michel Morel. Video denoising via empirical bayesian estimation of space-time patches. *Journal of Mathematical Imaging and Vision*, 60:70–93, 2018. 6
- [2] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2
- [3] 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*, pages 5754–5763, 2019. 2
- [4] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. 2, 4
- [5] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9822–9832, 2023. 1, 2, 5
- [6] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhong Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022. 1, 2, 4, 5, 6
- [7] Jingyun Liang, Jiezhong Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 2024. 1, 2, 4, 6
- [8] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring, 2018. 6
- [9] Mingyang Song, Yang Zhang, and Tunç O Aydın. Tempformer: Temporally consistent transformer for video denoising. In *European conference on computer vision*, pages 481–496. Springer, 2022. 2, 4
- [10] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1279–1288, 2017. 2, 3, 5
- [11] Matias Tassano, Julie Delon, and Thomas Veit. Dvdnet: A fast network for deep video denoising. In *IEEE International Conference on Image Processing*, 2019. 1, 2
- [12] 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
- [13] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 1, 2
- [14] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezaatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022. 2

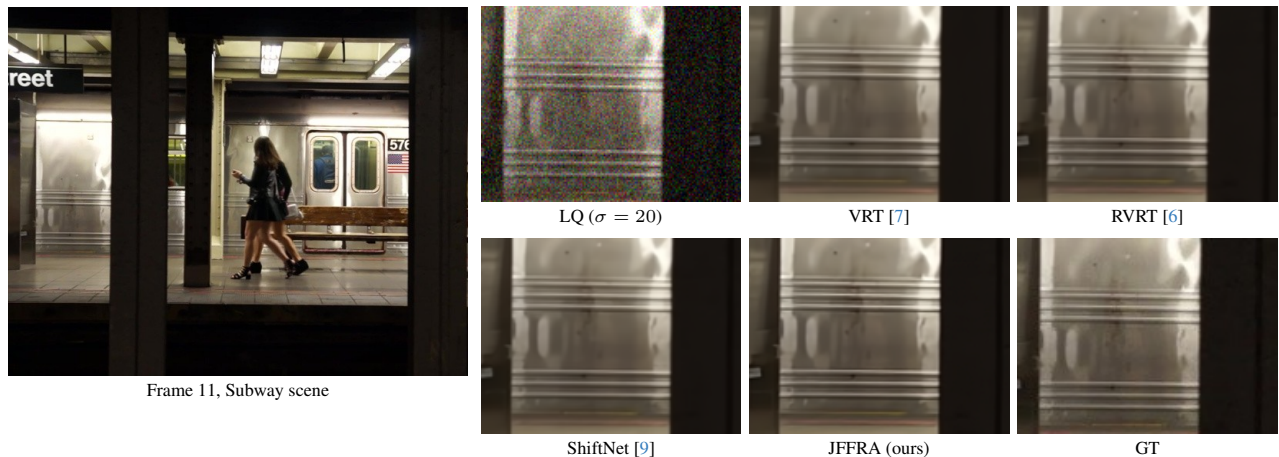


Figure 2. Qualitative analysis of various video denoising methods on Davis dataset [4].

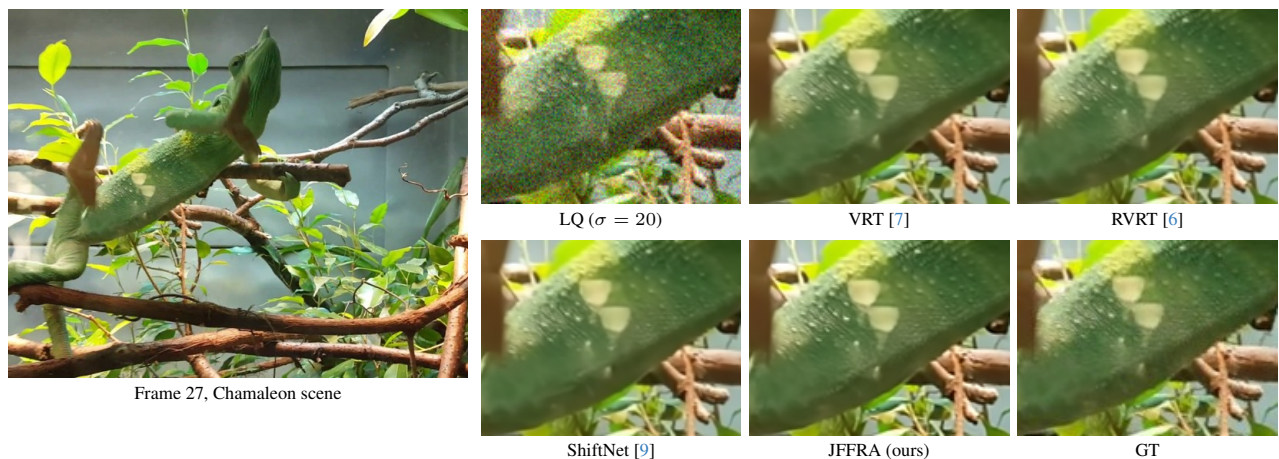


Figure 3. Qualitative analysis of various video denoising methods on Davis dataset [4].

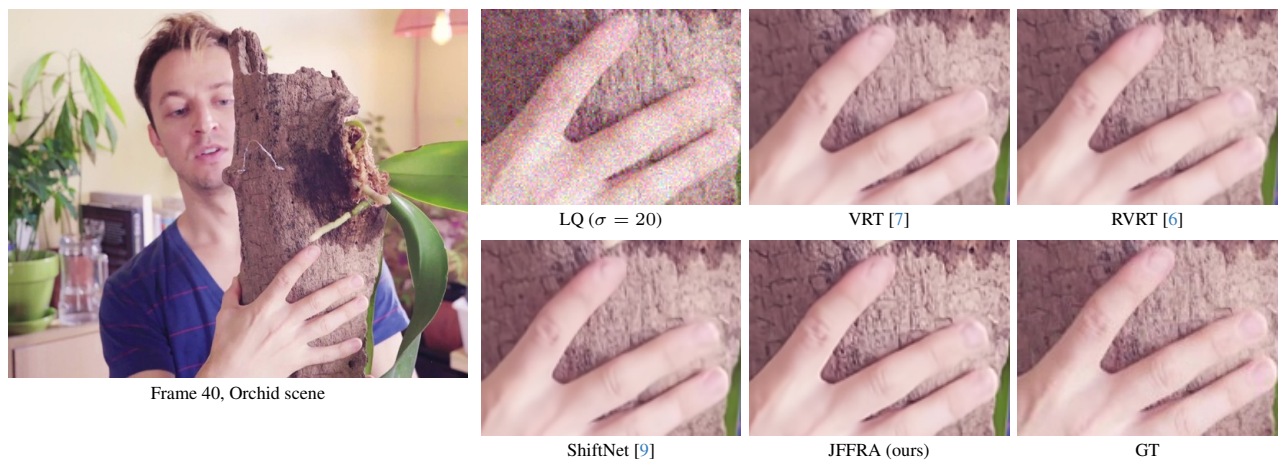


Figure 4. Qualitative analysis of various video denoising methods on Davis dataset [4].

- [15] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2019. 1, 2

- [16] Geunhyuk Youk, Jihyong Oh, and Munchurl Kim. Fmanet: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring. In *Proceedings of the IEEE/CVF Conference*



Figure 5. Visual comparison of various video deblurring methods on DVD dataset [10].



Figure 6. Visual comparison of various video deblurring methods on DVD dataset [10].

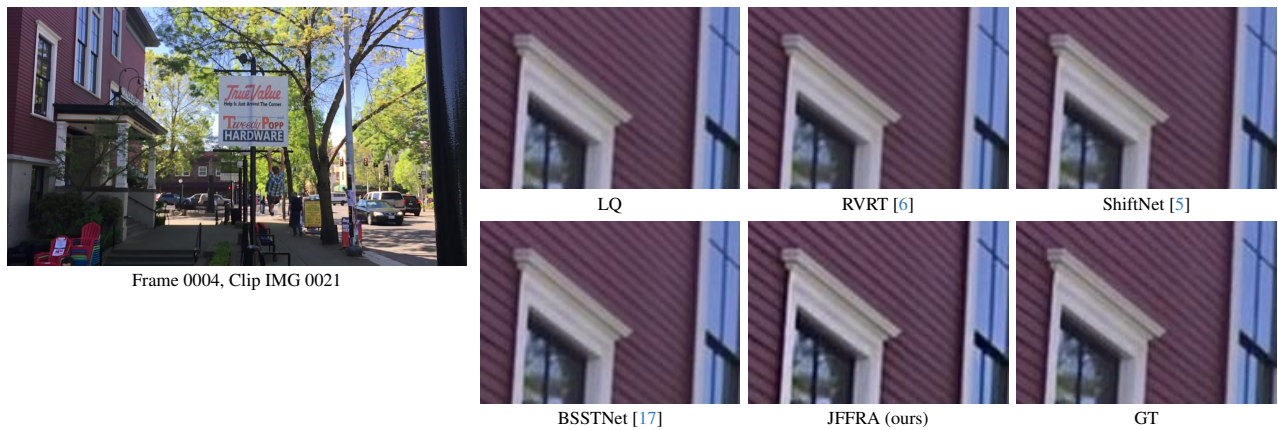


Figure 7. Visual comparison of various video deblurring methods on DVD dataset [10].

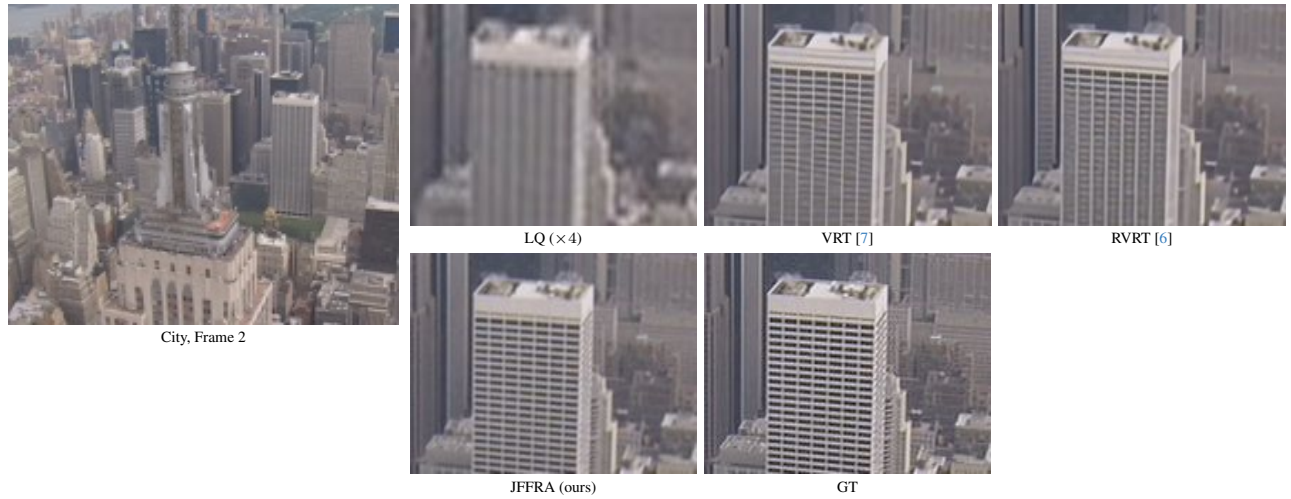


Figure 8. Qualitative analysis of various video Super-resolution methods on Vid4 dataset [1].

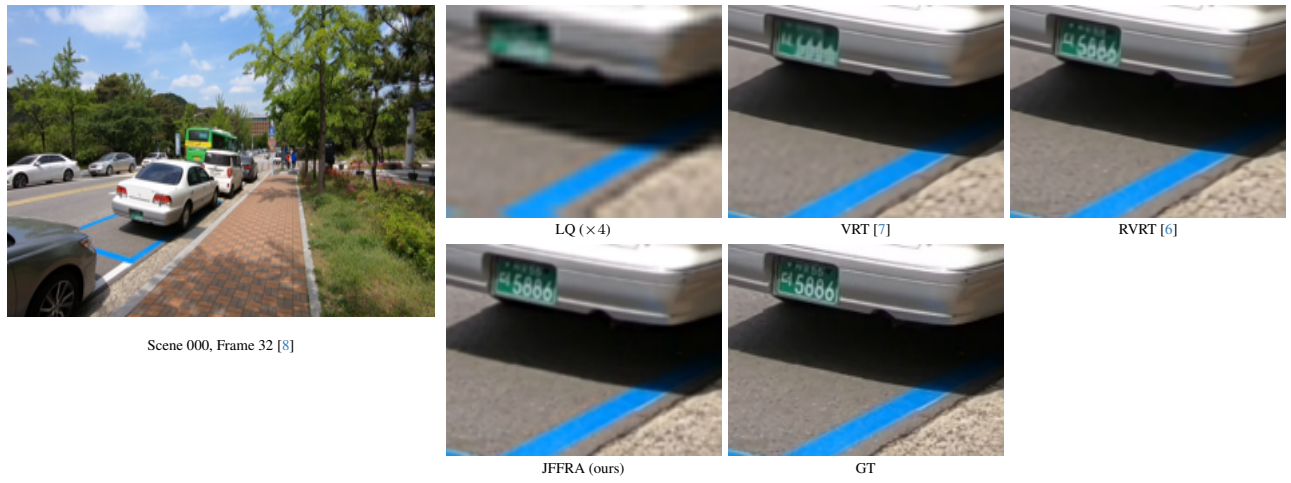


Figure 9. Qualitative analysis of various video Super-resolution methods on Reds dataset [8].

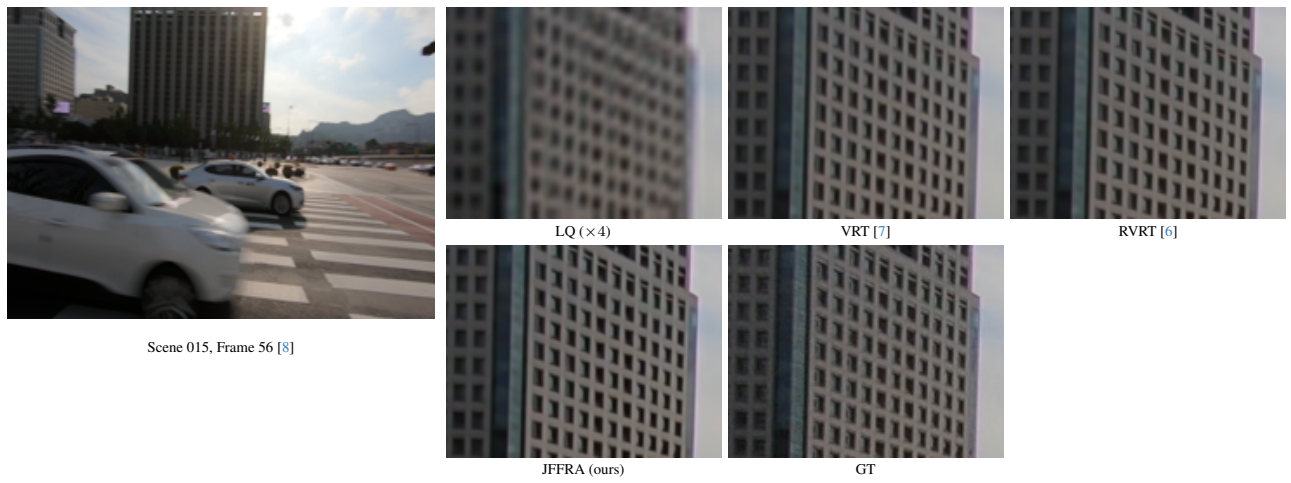


Figure 10. Qualitative analysis of various video Super-resolution methods on Reds dataset [8].

on Computer Vision and Pattern Recognition, pages 44–55, 2024. [1](#), [2](#)

- [17] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Blur-aware spatio-temporal sparse transformer for video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2673–2681, 2024. [5](#)