

# Supplementary Materials: Learning to Handle Large Obstructions in Video Frame Interpolation

Anonymous Authors

## 1 OVERVIEW

We present additional results and analysis in this supplemental material along with video results. We also give a description of our Real World Obstruction dataset in Section 4.

## 2 LAYER SELECTION FOR FEATURE REPAIR

Our method uses feature repair with the Region Similarity Map (RSM) as discussed in Sec. 3.4 of the main paper. We select a layer for our repair module in each network: RIFE [1], IFRNet [2], VFIFormer [3], and EMA [5]. In this section, we will report in Table 1 an ablation study that shows how we chose the layer for each network. For an efficient study, we use a smaller version of the network (with about 10% parameters of the original network) and a smaller training schedule for this ablation study. According to the results of the study, we find that feature repair is most effective in Level 2 in RIFE, VFIFormer and EMA and Level 1 for IFRNet as the performance of the repair leads to the highest results compared to other layers. However, we also note that the use of RSM and COA in neighboring layers achieve similar performance, showing the robustness of our approach.

## 3 ANALYSIS OF FEATURE REPAIR MODULE

To analyze the effectiveness of our feature repair module, we visualize further results and the corresponding RSM in this supplemental material. Figure 1 shows the results of our EMA-ORF in comparison to EMA. EMA fails in the region indicated by the yellow circle, and shows our method to be successful. We can see that the corresponding RSM highlights the same region. This shows the effectiveness of our feature repair module can handle the large occlusions in these examples and we believe it gives a good indication of the effectiveness of our method.

## 4 REAL-WORLD OBSTRUCTION DATASET

To our best knowledge, there is no existing dataset focused on obstruction handling in video frame interpolation. There are datasets for occlusion removal [4, 6]. However, we find some of the video frames are discontinuous which means the data cannot directly be used in video frame interpolation. Thus, we collect our Real-World Obstruction (RWO) Dataset in 11 different scenes with two different cameras, a cell phone (iPhone 12 pro) and consumer camera (Canon EOS 5D Mark IV). For each scene, we split 6-33 sequences frames for evaluation. We also select continuous frames from 6 different scenes in [4, 6]. Figure 2 shows some examples of triples in our RWO data. Table 2 reports the statistic of our dataset. We will make our RWO dataset public.

Table 1: Ablation study of layer selection.

Method	Vimeo90K		SNU-FILM EX		RWO	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RIFE [1]	34.31	0.9693	24.36	0.8464	26.82	0.8782
RIFE-l1	34.34	0.9691	24.42	0.8466	27.02	0.8791
RIFE-l2	<b>34.42</b>	<b>0.9698</b>	<b>24.52</b>	<b>0.8481</b>	<b>27.09</b>	<b>0.8798</b>
RIFE-l3	34.28	0.9694	24.41	0.8465	26.84	0.8781
RIFE-l4	34.24	0.9688	24.22	0.8457	26.75	0.8777
IFRNet [2]	34.43	0.9705	24.69	0.8498	27.12	0.8801
IFRNet-l1	<b>34.48</b>	<b>0.9708</b>	<b>24.83</b>	<b>0.8503</b>	<b>27.26</b>	<b>0.8813</b>
IFRNet-l2	34.44	0.9702	24.78	0.8500	27.22	0.8809
IFRNet-l3	34.39	0.9698	24.65	0.8491	27.16	0.8802
IFRNet-l4	34.36	0.9699	24.58	0.8484	27.07	0.8797
VFIFormer [3]	34.92	0.9737	24.83	0.8524	27.54	0.8844
VFIFormer-l1	<b>35.15</b>	<b>0.9743</b>	24.97	0.8526	27.62	<b>0.8849</b>
VFIFormer-l2	35.13	0.9741	<b>24.98</b>	<b>0.8531</b>	<b>27.66</b>	0.8848
VFIFormer-l3	34.88	0.9735	24.85	0.8522	27.57	0.8841
VFIFormer-l4	34.83	0.9732	24.75	0.8517	27.46	0.8835
EMA [5]	35.04	0.9744	25.03	0.8546	27.71	0.8862
EMA-l1	<b>35.13</b>	0.9748	25.08	0.8549	27.88	0.8879
EMA-l2	35.11	<b>0.9749</b>	<b>25.12</b>	<b>0.8551</b>	<b>27.96</b>	<b>0.8885</b>
EMA-l3	34.98	0.9741	24.94	0.8539	27.73	0.8864
EMA-l4	34.90	0.9734	24.98	0.8538	27.62	0.8854

Table 2: Statistics for images of different scenes in our Real World Obstructions (RWO) dataset.

Scene #	# of Frames	Type of obstruction	Indoor
1	15	screen window	✓
2	12	wire fence	✗
3	15	glass bottle	✓
4	15	reflection	✓
5	9	box substation	✗
6	15	bucket	✓
7	12	wall painting	✗
8	9	pillar	✗
9	9	stone sculptures	✗
10	12	reflection	✓
11	12	reflection	✓
12	9	fence	✗
13	3	fence	✗
14	9	raindrops	✗
15	9	reflection	✗
16	9	reflection	✗
17	9	statue	✗

Figure 1: Comparison of Results of EMA to our EMA-ORF and its corresponding RSM.

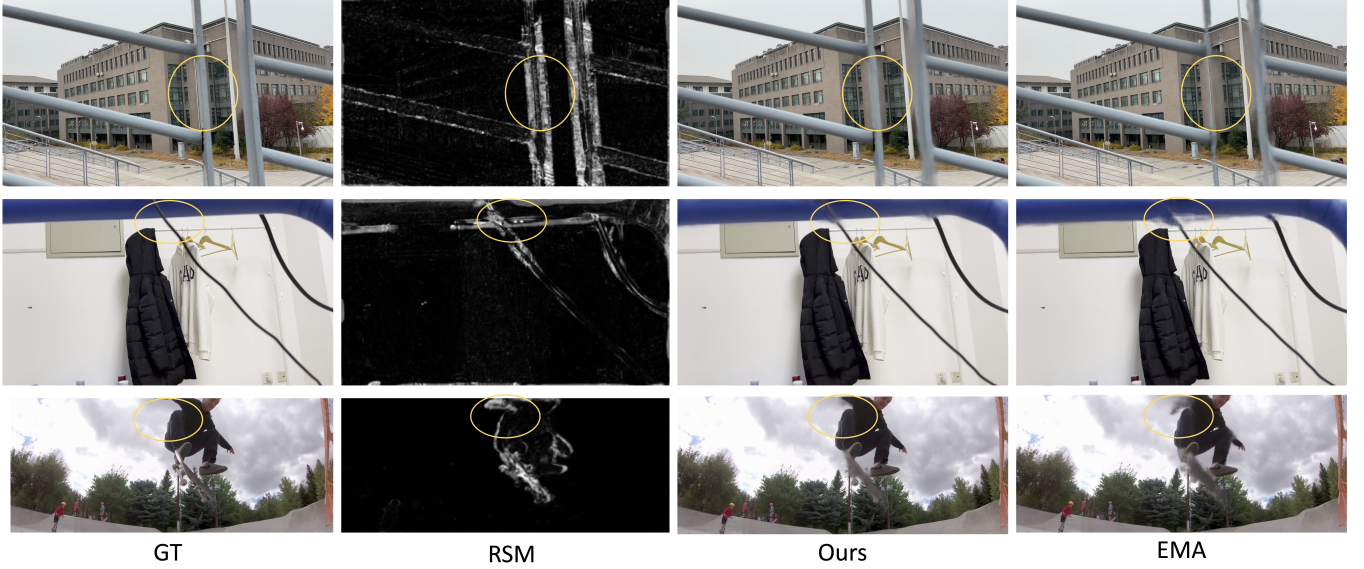


Figure 2: Example image triplets in our RWO dataset.

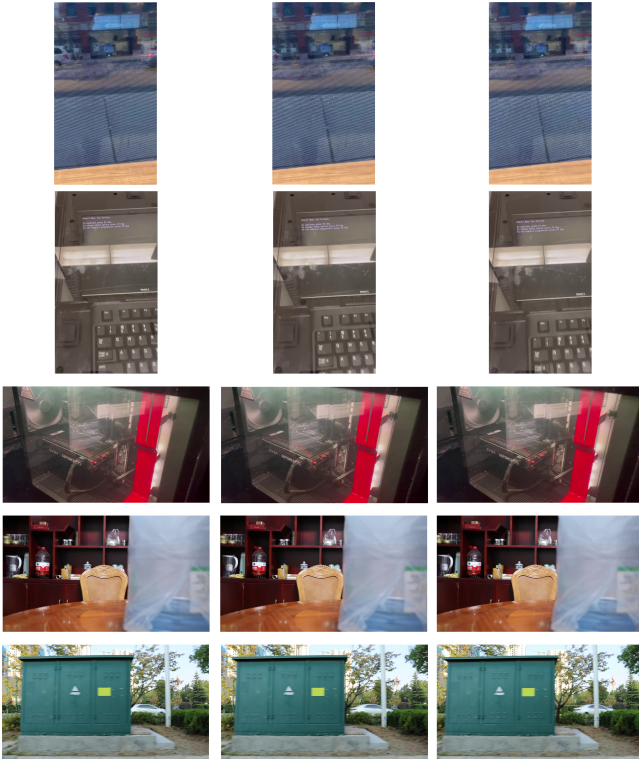


Table 3: Comparison of model complexity.

	VFIformer	EMA
Size(M)	24.1/25.4 (+5.1%)	65.66/66.12 (+ 0.7%)

## 5 DETAILS OF DYNAMIC MASK DISTRACTIONS

During the training, we generate a binary mask with a randomly chosen shape: Circle, Square, Oval, or Grid. The shape size ranges from 10% to 80% of the image resolution. Then, we generate two subsequent binary masks, each with one of three types of motion: translation, deformation, or discontinuous motion. We randomly select an image from another scene and multiply it by each binary mask, respectively. The new triples will be fused with the training triples according to Equation 8 in the main paper.

## 6 NETWORK COMPLEXITY

We report the complexity of our methods in Table 3. Compared with original methods, our VFIformer-ORF and EMA-ORF slightly increase the number of network parameters but significantly enhance the robustness to deal with large obstructions. (According to different architectures, the parameters of VFIformer-ORF and EMA-ORF vary due to the dimensions of repaired features.)

## 7 VIDEO RESULTS

To show the quality of our methods in real-world video with obstructions, we compare our methods with previous state-of-the-art methods. We put them in the folder **video**. We compare our EMA-ORF with the original EMA [5] and our VFIformer-DRC with the original VFIformer [3]. For both samples, we use the respective video interpolation method to increase the fps from 6 to 20. We

show our methods can make videos smoother and more continuous despite large obstructions in comparison to their respective base models. The base models EMA and VFIfomer generate large errors and artifacts. We will release our code for further evaluations.

REFERENCES

[1] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. 2022. Real-time intermediate flow estimation for video frame interpolation. In *ECCV*. Part XIV: 624–642.

[2] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. 2022. IFRNet: Intermediate Feature Refine

Network for Efficient Frame Interpolation. In *CVPR*. 1959–1968.

[3] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. 2022. Video Frame Interpolation with Transformer. In *CVPR*. 3522–3532.

[4] Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Qiang Fu, Hadi Amata, Xiao Li, Praneeth Chakravarthula, Wolfgang Heidrich, and Felix Heide. 2022. Seeing through Obstructions with Diffractive Cloaking. *ACM TOG* 41, 4 (2022). <https://doi.org/10.1145/3528223.3530185>

[5] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. 2023. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *CVPR*. 5682–5692.

[6] Chengxuan Zhu, Renjie Wan, Yunkai Tang, and Boxin Shi. 2023. Occlusion-Free Scene Recovery via Neural Radiance Fields. In *CVPR*. 20722–20731.