

Table 2: Quantitative comparison with SOTA methods on the low-resolution datasets, in terms of PSNR/SSIM. In “Training Dataset”, “V” stands for the triplet of Vimeo-90K and “X” stands for X-TRAIN. In “Runtime”, we evaluate the inference speed of each method on 1024×1024 resolution inputs by a 2080Ti GPU. In “Average”, we calculate the average performance of each method in terms of PSNR and SSIM. The best results and the second best results are **boldfaced** and underlined respectively. * indicates the results copied from EMA-VFI and AMT; the rest of the results are obtained through the same procedure.

	Training Dataset	Vimeo-90K	UCF101	SNU-FILM				Average	FLOPs (T)	Runtime (ms)
				easy	medium	hard	extreme			
DAIN*	V	34.71/0.9756	34.99/0.9683	39.73/0.9902	35.46/0.9780	30.17/0.9335	25.09/0.8584	33.36/0.9507	5.51	897.8
AdaCof*	V	34.47/0.9730	34.90/0.9680	39.80/0.9900	35.05/0.9754	29.46/0.9244	24.31/0.8439	33.00/0.9458	0.36	85.1
CAIN*	V	34.65/0.9730	34.91/0.9690	39.89/0.9900	35.61/0.9776	29.90/0.9292	24.78/0.8507	33.29/0.9483	1.29	102.4
Softsplat	V	36.13/0.9805	35.39/0.9697	<u>40.26/0.9911</u>	36.07/0.9798	30.53/0.9365	25.16/0.8604	33.92/0.9530	0.94	266.4
M2M-VFI	V	35.47/0.9778	35.28/0.9694	39.66/0.9904	35.74/0.9794	30.30/0.9360	25.08/0.8604	33.59/0.9522	0.26	60.9
RIFE	V	35.61/0.9779	35.28/0.9690	39.80/0.9903	35.76/0.9787	30.36/0.9351	25.27/0.8601	33.68/0.9519	0.20	35.2
XVFI	V	35.09/0.9759	35.17/0.9685	39.93/0.9907	35.37/0.9782	29.58/0.9276	24.17/0.8450	33.22/0.9477	0.37	165.2
EMA-VFI-S	V	36.07/0.9797	35.34/0.9696	39.81/0.9906	35.88/0.9795	30.69/0.9375	25.47/0.8632	33.88/0.9534	0.20	76.4
EMA-VFI	V	36.64/0.9819	35.48/0.9701	39.98/0.9910	36.09/ <u>0.9801</u>	<u>30.94/0.9392</u>	25.69/0.8661	<u>34.14/0.9547</u>	0.91	239.6
IFRNet-L	V	36.20/0.9810	35.42/0.9700	40.10/0.9910	36.12/0.9800	30.63/0.9371	25.26/0.8611	33.96/0.9534	0.79	115.3
AMT-L	V	36.35/0.9815	35.39/0.9698	39.95/ 0.9913	36.09/ 0.9805	30.75/0.9384	25.41/0.8638	33.99/0.9542	0.58	183.42
AMT-G	V	<u>36.53/0.9817</u>	35.41/0.9699	39.88/ 0.9913	36.12/ 0.9805	30.78/0.9385	25.43/0.8644	34.03/0.9544	2.07	403.7
SGM-VFI	V+X	35.81/0.9793	35.34/0.9693	40.14/0.9907	36.06/0.9795	30.81/0.9375	<u>25.59/0.8646</u>	33.96/0.9535	1.78	942.9
VFIMamba-S	V+X	36.09/0.9800	35.36/0.9696	40.21/0.9909	<u>36.17/0.9800</u>	30.80/0.9381	<u>25.59/0.8655</u>	34.04/0.9540	0.24	128.0
VFIMamba	V+X	36.64/0.9819	<u>35.45/0.9702</u>	40.51/0.9912	36.40/0.9805	30.99/0.9401	25.79/0.8682	34.30/0.9554	0.94	310.9

Table 3: Quantitative comparison with SOTA methods on high-resolution datasets.

	Training Dataset	X-TEST		X-TEST-L		Xiph		Average
		2K	4K	2K	4K	2K	4K	
XVFI	X	31.15/0.9144	30.12/0.9045	29.82/0.8951	29.02/0.8866	34.76/0.9258	32.84/0.8810	31.29/0.9012
M2M-VFI	V	32.13/0.9258	30.89/0.9138	30.90/0.9092	29.73/0.9001	36.44/0.9427	33.92/0.8992	32.34/0.9151
RIFE	V	31.10/0.8972	30.13/0.8927	29.87/0.8805	28.98/0.8756	36.19/0.9380	33.76/0.8940	31.67/0.8963
FILM	V	31.61/0.9174	OOM	30.18/0.8960	OOM	36.32/0.9343	33.27/0.8760	/
IFRNet-L	V	31.78/0.9147	30.66/0.9050	30.76/0.8963	29.74/0.8884	36.21/0.9374	34.25/0.8946	32.23/0.9061
FLDR	X	31.12/0.9092	30.46/0.9041	29.90/0.8906	29.30/0.8879	34.80/0.9280	33.00/0.8862	31.43/0.9010
BiFormer	V+X	31.32/0.9200	31.32/0.9215	30.36/0.9068	30.14/0.9069	34.20/0.9246	33.49/0.8953	31.81/0.9125
EMA-VFI-S	V	30.91/0.9000	29.91/0.8951	29.51/0.8775	28.60/0.8733	36.55/0.9421	34.25/0.9020	31.62/0.8983
AMT-L	V	32.08/0.9277	30.96/0.9147	31.09/0.9103	30.12/0.9019	36.27/0.9402	34.49/0.9030	32.50/0.9163
AMT-G	V	32.35/0.9300	31.12/0.9157	31.35/0.9125	30.33/0.9036	36.38/0.9410	34.63/0.9039	32.69/0.9178
SGM-VFI	V+X	32.38/0.9272	31.35/0.9179	30.99/0.9072	29.91/0.8972	36.57/0.9424	34.23/0.9021	32.57/0.9157
VFIMamba-S	V+X	<u>32.84/0.9328</u>	<u>31.73/0.9238</u>	<u>31.58/0.9169</u>	<u>30.50/0.9077</u>	<u>36.72/0.9428</u>	34.32/0.9034	<u>32.95/0.9212</u>
VFIMamba	V+X	33.34/0.9361	32.15/0.9246	32.22/0.9259	31.05/0.9159	<u>37.13/0.9451</u>	<u>34.62/0.9059</u>	33.42/0.9256

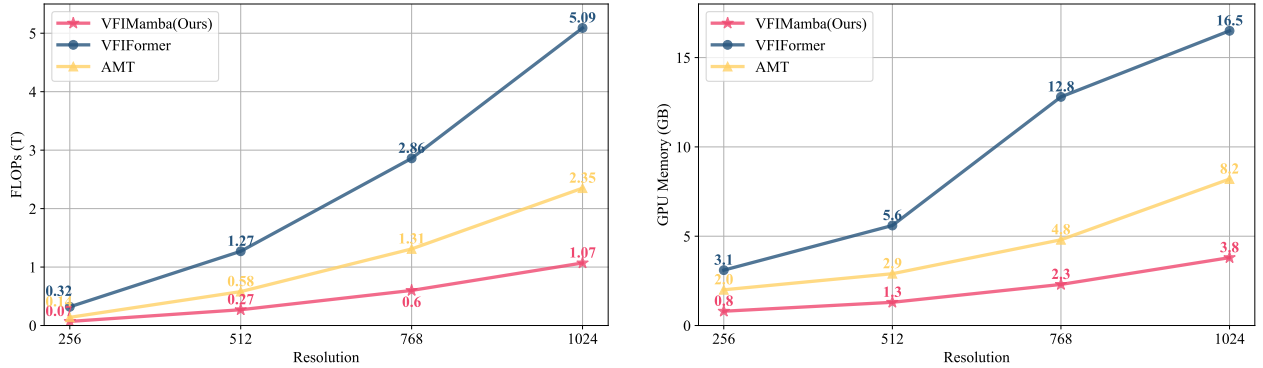


Figure 10: Comparisons of FLOPs and GPU memory usage with increasing resolution. All results are tested on an NVIDIA A5000 GPU.

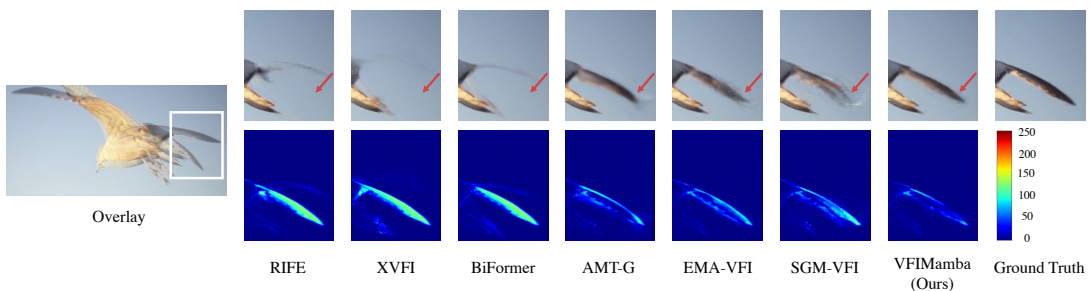


Figure 11: Error map visualization by calculating the absolute difference between interpolation results and ground truth.