

A IMPLEMENTATION DETAILS OF MADIFF

We summarize the training procedure of VQ-MAGAN and de-noising U-net of MADIFF as Algorithm 1 and Algorithm 2. Following [15], we use the DDIM [58] sampler, which has been shown to achieve sampling quality on par with the full original sampling method, but with fewer steps. And we summarize the MA-SAMPLING and the corresponding DDIM [58] procedure of MADIFF as Algorithm 3 and Algorithm 4. And the differences from the original procedure are highlighted in blue.

A.1 Training procedure of VQ-MAGAN

We update the parameters of the encoder \mathcal{E} and the decoder \mathcal{D} by using the loss function consists of an LPIPS-based [77] perceptual loss, a patch-based adversarial loss [27] and a latent regularization term based on a vector quantization (VQ) layer [66] following [15]. More details please refer to the Algorithm 1.

Algorithm 1 Training procedure of VQ-MAGAN

- 1: **Input:** dataset $\mathcal{D} = \{I_{-1}^s, I_{+1}^s, I_0^s\}_{s=1}^S$ of consecutive frame triplets
 - 2: **Load:** the pre-trained EventGAN [80] $f_{I2E}(\cdot)$
 - 3: **Initialize:** The encoder \mathcal{E} and the decoder \mathcal{D} of VQ-MAGAN
 - 4: **repeat**
 - 5: Sample $(I_{-1}, I_{+1}, I_0) \sim \mathcal{D}$
 - 6: Encode $z_0 = \mathcal{E}(I_0)$, $z_{-1} = \mathcal{E}(I_{-1})$, $z_{+1} = \mathcal{E}(I_{+1})$ and store features ϕ_{-1}, ϕ_{+1} extracted by \mathcal{E}
 - 7: Obtain inter-frame motion hints from the ground-truth target frame and neighboring frames as additional conditions by $f_{I2E}(\cdot)$:
 $m_{-1 \rightarrow 0} = f_{I2E}(I_{-1}, I_0)$
 $m_{0 \rightarrow +1} = f_{I2E}(I_0, I_{+1})$
 - 8: Reconstruct the target frame:
 $\hat{I}_0 = \mathcal{D}(z_0, \phi_{-1}, \phi_{+1}, m_{-1 \rightarrow 0}, m_{0 \rightarrow +1})$
 - 9: Compute loss \mathcal{L} with \hat{I}_0 and I_0
 - 10: Jointly update \mathcal{E} and \mathcal{D} by minimizing \mathcal{L}
 - 11: **until** converged
-

A.2 Training procedure of MADIFF

Algorithm 2 Training procedure of MADIFF

- 1: **Input:** dataset $\mathcal{D} = \{I_{-1}^s, I_{+1}^s, I_0^s\}_{s=1}^S$ of consecutive frame triplets, maximum diffusion step T , noise schedule $\{\beta_t\}_{t=1}^T$, learnable de-noising U-net $\epsilon_\theta(\cdot)$
 - 2: **Load:** the encoder \mathcal{E} of the pre-trained VQ-MAGAN, the pre-trained EventGAN $f_{I2E}(\cdot)$
 - 3: Compute $\{\bar{\alpha}_t\}_{t=1}^T$ from $\{\beta_t\}_{t=1}^T$
 - 4: **repeat**
 - 5: Sample $(I_{-1}, I_{+1}, I_0) \sim \mathcal{D}$
 - 6: Encode $z_{-1} = \mathcal{E}(I_{-1})$, $z_{+1} = \mathcal{E}(I_{+1})$, $z_0 = \mathcal{E}(I_0)$
 - 7: Sample $t \sim \mathcal{U}(1, T)$
 - 8: Sample $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 - 9: $z_t^n = \sqrt{\bar{\alpha}_t} z^n + \sqrt{1 - \bar{\alpha}_t} \epsilon$
 - 10: Obtain inter-frame motion hints from the ground-truth target frame and neighboring frames as additional conditions by $f_{I2E}(\cdot)$:
 $m_{-1 \rightarrow 0} = f_{I2E}(I_{-1}, I_0)$
 $m_{0 \rightarrow +1} = f_{I2E}(I_0, I_{+1})$
 - 11: Take a gradient descent step on:
 $\nabla_\theta \|\epsilon - \epsilon_\theta(z_t^n, t, z^0, z^1, m_{-1 \rightarrow 0}, m_{0 \rightarrow +1})\|^2$
 - 12: **until** converged
-

A.3 MA-SAMPLING procedure of MADIFF

Algorithm 3 MA-SAMPLING procedure of MADIFF

- 1: **Input:** original frames I^0, I^1 , noise schedule $\{\beta_t\}_{t=1}^T$, maximum diffusion step T
 - 2: **Load:** pre-trained de-noising U-net ϵ_θ , the encoder \mathcal{E} and the decoder \mathcal{D} of VQ-MAGAN, the pre-trained EventGAN $f_{I2E}(\cdot)$
 - 3: Compute $\{\bar{\alpha}_t\}_{t=1}^T$ from $\{\beta_t\}_{t=1}^T$
 - 4: Sample $\hat{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 - 5: Encode $z_{-1} = \mathcal{E}(I_{-1})$, $z_{+1} = \mathcal{E}(I_{+1})$ and store features ϕ_{-1}, ϕ_{+1} extracted by \mathcal{E}
 - 6: **Let** $\hat{m}_{-1 \rightarrow 0|T+1} = \mathbf{O}$ and $\hat{m}_{0 \rightarrow +1|T+1} = \mathbf{O}$
 - 7: **for** $t = T, \dots, 1$ **do**
 - 8: Predict noise:
 $\hat{\epsilon} = \epsilon_\theta(\hat{z}_t, t, z_{-1}, z_{+1}, \hat{m}_{-1 \rightarrow 0|t+1}, \hat{m}_{0 \rightarrow +1|t+1})$
 - 9: Predict $\hat{z}_{0|t}$ from noise:
 $\hat{z}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\hat{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon})$
 - 10: Reconstruct the interpolated frame
 $\hat{I}_{0|t} = \mathcal{D}(\hat{z}_{0|t}, \phi_{-1}, \phi_{+1}, \hat{m}_{-1 \rightarrow 0|t+1}, \hat{m}_{0 \rightarrow +1|t+1})$
 - 11: Update extracted inter-frame motion hints
 $\hat{m}_{-1 \rightarrow 0|t} = f_{I2E}(I_{-1}, \hat{I}_{0|t})$
 $\hat{m}_{0 \rightarrow +1|t} = f_{I2E}(\hat{I}_{0|t}, I_{+1})$
 - 12: $\sigma_t^2 = \bar{\beta}_t$
 - 13: Sample $\zeta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 - 14: $\hat{z}_{t-1} = \hat{z}_{0|t} + \sigma_t \zeta$
 - 15: **end for**
 - 16: **return** $\hat{I}_{0|1} = \mathcal{D}(\hat{z}_{0|1}, \phi_{-1}, \phi_{+1}, \hat{m}_{-1 \rightarrow 0|1}, \hat{m}_{0 \rightarrow +1|1})$ as the final interpolated frame
-

Algorithm 4 DDIM Sampler of MA-SAMPLING

- 1: **Input:** original frames I_{-1}, I_{+1} , noise schedule $\{\beta_t\}_{t=1}^T$, maximum DDIM step \mathcal{T}
 - 2: **Load:** pre-trained de-noising U-net ϵ_θ , the encoder \mathcal{E} and the decoder \mathcal{D} , the pre-trained EventGAN $f_{I2E}(\cdot)$
 - 3: Compute $\{\bar{\alpha}_t\}_{t=1}^T$ from $\{\beta_t\}_{t=1}^T$
 - 4: Sample $\hat{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 - 5: Encode $z_{-1} = \mathcal{E}(I_{-1})$, $z_{+1} = \mathcal{E}(I_{+1})$ and store features ϕ_{-1}, ϕ_{+1} extracted by \mathcal{E}
 - 6: **Let** $\hat{m}_{-1 \rightarrow 0|T+1} = \mathbf{O}$ and $\hat{m}_{0 \rightarrow +1|T+1} = \mathbf{O}$
 - 7: **for** $t = T, \dots, 1$ **do**
 - 8: Predict noise:
 $\hat{\epsilon} = \epsilon_\theta(\hat{z}_t, t, z_{-1}, z_{+1}, \hat{m}_{-1 \rightarrow 0|t+1}, \hat{m}_{0 \rightarrow +1|t+1})$
 - 9: Predict $\hat{z}_{0|t}$ from noise:
 $\hat{z}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\hat{z}_t - \sqrt{1 - \bar{\alpha}_t}\hat{\epsilon})$
 - 10: Predict $\hat{z}_{0|t}$ from noise:
 $\hat{z}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\hat{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon})$
 - 11: Reconstruct the interpolated frame
 $\hat{I}_{0|t} = \mathcal{D}(\hat{z}_{0|t}, \phi_{-1}, \phi_{+1}, \hat{m}_{-1 \rightarrow 0|t+1}, \hat{m}_{0 \rightarrow +1|t+1})$
 - 12: Update extracted inter-frame motion hints
 $\hat{m}_{-1 \rightarrow 0|t} = f_{I2E}(I_{-1}, \hat{I}_{0|t})$
 $\hat{m}_{0 \rightarrow +1|t} = f_{I2E}(\hat{I}_{0|t}, I_{+1})$
 - 13: $\hat{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{z}_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1}}\hat{\epsilon}$
 - 14: **end for**
 - 15: **return** $\hat{I}_{0|1} = \mathcal{D}(\hat{z}_{0|1}, \phi_{-1}, \phi_{+1}, \hat{m}_{-1 \rightarrow 0|1}, \hat{m}_{0 \rightarrow +1|1})$ as the final interpolated frame
-

	Middlebury					UCF-101					DAVIS				
	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓	FID↓
BMBC	36.368	0.982	0.023	0.037	12.974	32.576	0.968	0.034	0.045	33.171	26.835	0.869	0.125	0.185	15.354
AdaCoF	35.256	0.975	0.031	0.052	15.633	32.488	0.968	0.034	0.046	32.783	26.234	0.850	0.148	0.198	17.194
CDFI	36.205	0.981	0.022	0.043	12.224	32.541	0.968	0.036	0.049	33.742	26.471	0.857	0.157	0.211	18.098
XVFI	34.724	0.975	0.036	0.070	16.959	32.224	0.966	0.038	0.050	33.868	26.475	0.861	0.129	0.185	16.163
ABME	37.639	0.986	0.027	0.040	11.393	32.055	0.967	0.058	0.069	37.066	26.861	0.865	0.151	0.209	16.931
IFRNet	36.368	0.983	0.020	0.039	12.256	32.716	0.969	0.032	0.044	28.803	27.313	0.877	0.114	0.170	14.227
VFIformer	35.566	0.977	0.031	0.065	15.634	32.745	0.968	0.039	0.051	34.112	26.241	0.850	0.191	0.242	21.702
ST-MFNet	N/A	N/A	N/A	N/A	N/A	33.384	0.970	0.036	0.049	34.475	28.287	0.895	0.125	0.181	15.626
FLAVR	N/A	N/A	N/A	N/A	N/A	33.224	0.969	0.035	0.046	31.449	27.104	0.862	0.209	0.248	22.663
MCVD	20.539	0.820	0.123	0.138	41.053	18.775	0.710	0.155	0.169	102.054	18.946	0.705	0.247	0.293	28.002
LDMVFI	34.033	0.971	0.019	0.044	16.167	32.186	0.963	0.026	0.035	26.301	25.541	0.833	0.107	0.153	12.554
MADIFF w/o MS	34.002	0.973	0.016	0.034	13.649	32.141	0.966	0.024	0.032	24.677	25.952	0.849	0.098	0.143	11.764
MADIFF	34.170	0.974	0.016	0.034	11.678	32.159	0.966	0.024	0.033	24.289	26.069	0.853	0.096	0.142	11.089

Table 6: Quantitative comparison of MADIFF ($f = 32$) and 11 tested methods on Middlebury, UCF-101 and DAVIS. Note ST-MFNet and FLAVR require four input frames so cannot be evaluated on Middlebury dataset which contains frame triplets. For each column, we highlight the best result in **red and the second best in **blue**.**

	SNU-FILM-Easy					SNU-FILM-Medium					SNU-FILM-Hard					SNU-FILM-Extreme				
	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓	FID↓
BMBC	39.809	0.990	0.020	0.031	6.162	35.437	0.978	0.034	0.059	12.272	29.942	0.933	0.068	0.118	25.773	24.715	0.856	0.145	0.237	49.519
AdaCoF	39.632	0.990	0.021	0.033	6.587	34.919	0.975	0.039	0.066	14.173	29.477	0.925	0.080	0.131	27.982	24.650	0.851	0.152	0.234	52.848
CDFI	39.881	0.990	0.019	0.031	6.133	35.224	0.977	0.036	0.066	12.906	29.660	0.929	0.081	0.141	29.087	24.645	0.854	0.163	0.255	53.916
XVFI	38.903	0.989	0.022	0.037	7.401	34.552	0.975	0.039	0.072	16.000	29.364	0.928	0.075	0.138	29.483	24.545	0.853	0.142	0.233	54.449
ABME	39.697	0.990	0.022	0.034	6.363	35.280	0.977	0.042	0.076	15.159	29.643	0.929	0.092	0.168	34.236	24.541	0.853	0.182	0.300	63.561
IFRNet	39.881	0.990	0.019	0.030	5.939	35.668	0.979	0.033	0.058	12.084	30.143	0.935	0.065	0.122	25.436	24.954	0.859	0.136	0.229	50.047
ST-MFNet	40.775	0.992	0.019	0.031	5.973	37.111	0.984	0.036	0.061	11.716	31.698	0.951	0.073	0.123	25.512	25.810	0.874	0.148	0.238	53.563
FLAVR	40.161	0.990	0.022	0.034	6.320	36.020	0.979	0.049	0.077	15.006	30.577	0.938	0.112	0.169	34.746	25.206	0.861	0.217	0.303	72.673
MCVD	22.201	0.828	0.199	0.230	32.246	21.488	0.812	0.213	0.243	37.474	20.314	0.766	0.250	0.292	51.529	18.464	0.694	0.320	0.385	83.156
LDMVFI	38.674	0.987	0.014	0.024	5.752	33.996	0.970	0.028	0.053	12.485	28.547	0.917	0.060	0.114	26.520	23.934	0.837	0.123	0.204	47.042
MADIFF w/o MS	38.690	0.988	0.013	0.021	5.157	34.183	0.974	0.025	0.048	10.919	28.774	0.923	0.058	0.110	23.143	23.861	0.841	0.125	0.210	49.435
MADIFF	38.644	0.988	0.013	0.021	5.334	34.255	0.973	0.027	0.049	11.022	28.961	0.923	0.058	0.107	22.707	24.150	0.847	0.118	0.198	44.923

Table 7: Quantitative comparison results on SNU-FILM (note VFIformer is not included because the GPU goes out of memory). For each column, we highlight the best result in **red and the second best in **blue**.**

A.4 Architecture of De-noising U-net

Following [15], we employ the time-conditioned U-Net as in [55] for ϵ_θ and replace all the vanilla self-attention blocks [67] with the MaxViT blocks [64] for computational efficiency. And each encoder layer consists of 2 ResNetBlock, 1 Max Cross-Attention Block, each decoder layer consists of 2 ResNetBlock, 1 Max Cross-Attention Block and a Up-sampling Layer. And the number of attention head is set to 32.

B MORE RESULTS

B.1 Quantitative Comparisons

The full evaluation results of MADIFF and the compared VFI methods on all test sets (Middlebury [1], UCF-101 [62], DAVIS [52] and SNU-FILM [10]) in terms of all metrics (PSNR, SSIM [70], LPIPS [77], FloLPIPS [12] and FID [19]) are summarized in Table 6 and Table 7. We observe that, on perceptual-related metrics such as LPIPS, FloLPIPS, and FID, our MADIFF attains state-of-the-art performance when compared to existing VFI methods, encompassing both non-diffusion and diffusion-based approaches. Furthermore, in terms of PSNR and SSIM, our MADIFF demonstrates superior performance compared to existing diffusion-based methods.

B.2 Qualitative Comparisons

We provide more qualitative comparison results as shown in Figure 4.



Figure 4: More visual examples of frames interpolated by the state-of-the-art methods and the proposed MADIFF. Under large and complex motions, our method preserves the most high-frequency details, delivering superior perceptual quality.