– Supplemental Materials –

BiFormer: Learning Bilateral Motion Estimation via Bilateral Transformer for 4K Video Frame Interpolation

Junheum Park
Korea University
`jhpark@mcl.korea.ac.kr`

Jintae Kim
Korea University
`jtkim@mcl.korea.ac.kr`

Chang-Su Kim*
Korea University
`changsukim@korea.ac.kr`

A. Implementation Details

A.1. Global Motion Estimation: BiFormer

Table S-1. Network details of BiFormer.

	Layer					
	Type	Input	Output	Kernel	Stride	Padding
Motion prediction module	Conv2D	256 + 225	64	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Res_Conv2D	64	64	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Res_Conv2D	64	64	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Conv2D	64	64	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Conv2D	64	64	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	DeConv2D	64	64	3×3	2×2	1×1
	Conv2D	64	2	3×3	1×1	1×1

As shown in Figure 3 in the main paper, BiFormer contains the global feature extraction and motion prediction modules.

Global feature extraction: We adopt the 'Twins-SVT-L' architecture in [49] as the transformer encoder (or global feature extractor) of BiFormer. We use blocks at the first and second scales only and exclude those at the third and fourth scales. Hence, we use the output at the second scale as a global feature map. Since it has a channel size of 256 and a multi-head size of 8, we also design the transformer blocks in the bilateral attention module to have the same hyper-parameters. As the effectiveness of transformer initialization was demonstrated in optical flow estimation [37], we also initialize the transformer encoder with pre-trained parameters for ImageNet-1K.

Motion prediction: Table S-1 presents the details of the motion prediction module in BiFormer. It consists of four convolution layers, one residual block, and one deconvolution layer. The first convolution layer takes the concatenation of a bilateral cost volume for 15×15 search windows and a bilateral feature map from the bilateral attention module. Thus, the number of input channels is $256 + 15 \times 15$. The deconvolution layer doubles the spatial resolution of features, and its last convolution layer yields the global bilateral motion field $\mathcal{V}^{\mathrm{G}}_{t\rightarrow 1}$. According to (13), we have $\mathcal{V}^{\mathrm{G}}_{t\rightarrow 0} = -\mathcal{V}^{\mathrm{G}}_{t\rightarrow 1}$.

Training details: As done in [48,49,50], we employ the AdamW optimizer with a learning rate of $\eta = 1.25 \times 10^{-4}$ and a weight decay of 10^{-4}. We use a batch size of 16 for 0.35M iterations. We augment training data by random rotation, order reversing, random flipping, and random cropping of 256×256 patches.

A.2. Local Motion Refinement: Upsampler

Table S-2. Network details of the proposed upsampler.

	Layer					
	Type	Input	Output	Kernel	Stride	Padding
Shallow encoder	Conv2D	3	32	3×3	2×2	1×1
	LeakyReLU	-	-	-	-	-
	Conv2D	32	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Conv2D	32	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
Block embedding	Conv2D	32	32	1×1	1×1	0×0
	LeakyReLU	-	-	-	-	-
	Conv2D	32	64	2×2	2×2	0×0
	LeakyReLU	-	-	-	-	-
	Conv2D	32	128	4×4	4×4	1×1
	LeakyReLU	-	-	-	-	-
Cost embedding	Conv2D	9 + 2	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Conv2D	9 + 2	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Conv2D	9 + 2	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Conv2D	96	32	1×1	1×1	1×1
	LeakyReLU	-	-	-	-	-
Motion decoder	Conv2D	64 + 32	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Res_Conv2D	32	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Res_Conv2D	32	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Res_Conv2D	32	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Res_Conv2D	32	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Res_Conv2D	32	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Res_Conv2D	32	32	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	Conv2D	32	64	3×3	1×1	1×1
	LeakyReLU	-	-	-	-	-
	DeConv2D	64	32	3×3	2×2	1×1
	Conv2D	32	2	3×3	1×1	1×1

Table S-2 lists the details of the proposed upsampler, whose block diagram is in Figure 6 in the main paper. The shallow encoder consists of three convolution layers, and the first has stride 2. In block embedding, 1×1, 2×2, and 4×4 block embedding layers have output channel sizes 32, 64, and 128, respectively. In cost embedding, there are three 3×3 convolution layers for three BBCVs $\{\mathcal{B}_t^0, \mathcal{B}_t^1, \mathcal{B}_t^2\}$ and one 1×1 convolution layer. Each 3×3 convolution layer takes its corresponding BBCV and the bilateral motion field to generate a cost feature map. Then, the three cost feature maps are aggregated via a 1×1 convolution layer. Next, the motion decoder consists of three convolution layers, three residual blocks, and one deconvolution layer. The first convolution layer takes the concatenation of the two warped local feature maps and the aggregated cost feature map. The deconvolution layer doubles the spatial resolution of features. Then, the last convolution layer generates the residual bilateral motion field. Moreover, we employ the context network of PWC-Net [31], composed of 7 dilated convolution layers. It takes the concatenation of the refined bilateral motion field and the output of the deconvolution layer in the motion decoder as input and further refines the motion field.

A.3. Frame Synthesis

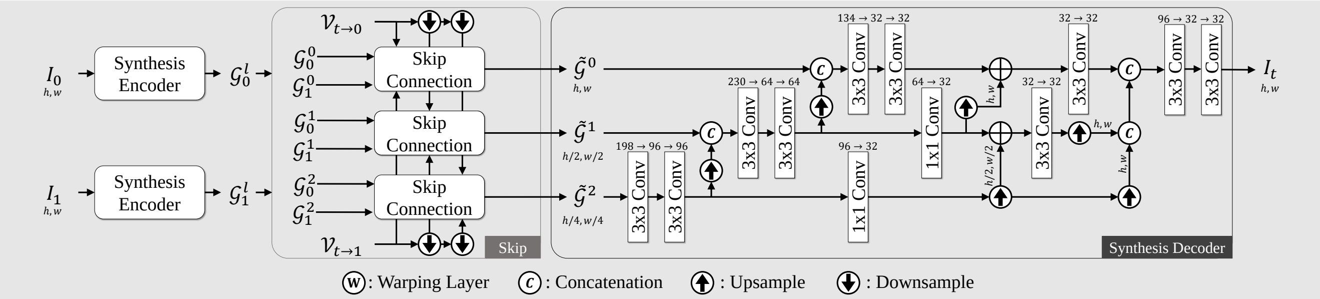
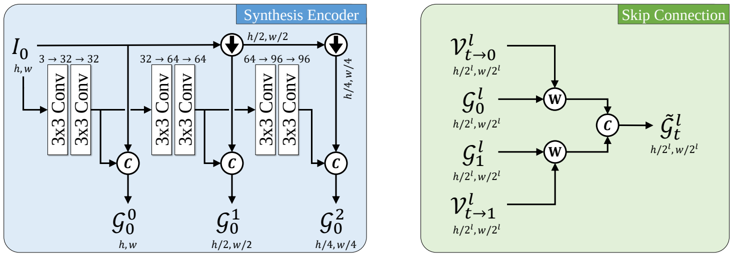


Figure S-1. The network structure of the frame synthesis network.



(a)

(b)

Figure S-2. (a) The synthesis encoder and (b) the skip connection layer.

Figure S-1 shows the architecture of the frame synthesis network. Note that the refined motion fields are at a $\frac{1}{2}$ scale. We bilinearly interpolate them to the full 4K resolution and denote the full-scale fields also as $\mathcal{V}_{t\to0}$ and $\mathcal{V}_{t\to1}$ here. Given input frames I_0 and I_1, the synthesis encoder extracts multi-scale feature maps $\mathcal{G}_0^l$ and $\mathcal{G}_1^l$, where $l \in \{0, 1, 2\}$ is a scale index. In the synthesis encoder, as shown in Figure S-2(a), a frame is processed by two convolution layers at each scale, six convolution layers in total, and the results are concatenated with downsampled input frames to yield multi-scale feature maps. Then, at level l, the skip connection layer in Figure S-2(b) warps the synthesis feature maps $\mathcal{G}_0^l$ and $\mathcal{G}_1^l$ with the downsampled bilateral motion fields $\mathcal{V}_{t\to0}^l$ and $\mathcal{V}_{t\to1}^l$ to yield the warped maps

$$\phi_B(\mathcal{V}_{t\to0}^l, \mathcal{G}_0^l) \quad \text{and} \quad \phi_B(\mathcal{V}_{t\to1}^l, \mathcal{G}_1^l). \tag{1}$$

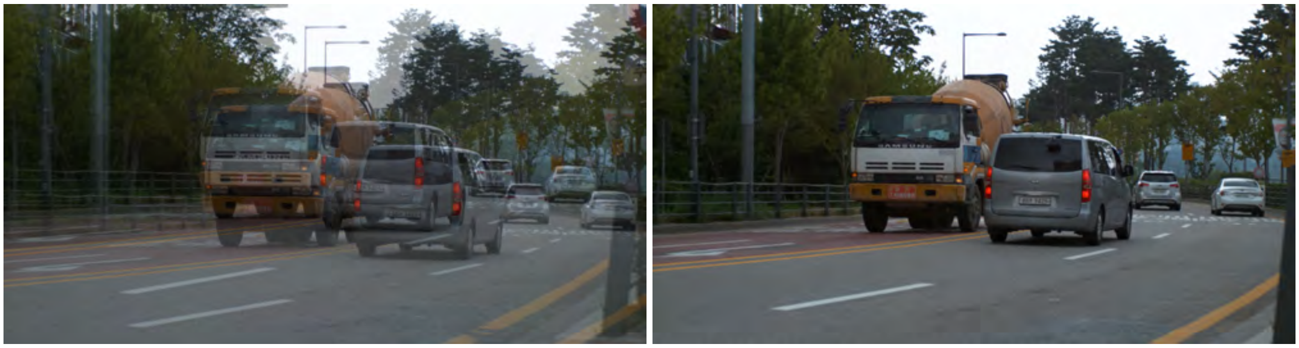
These two maps are then concatenated to yield a skipped feature map $\tilde{\mathcal{G}}_t^l$. The synthesis decoder processes the three skipped feature maps at $l \in \{0, 1, 2\}$ to reconstruct the intermediate frame I_t finally.

In Figure S-1 and Figure S-2(a), $N \times N$ Conv represents a convolution layer with kernel size $N \times N$. The numbers of input and output channels are specified above each convolution block. The third and fifth convolution layers in the synthesis encoder have stride 2, while the others have stride 1.

B. More Experimental Results

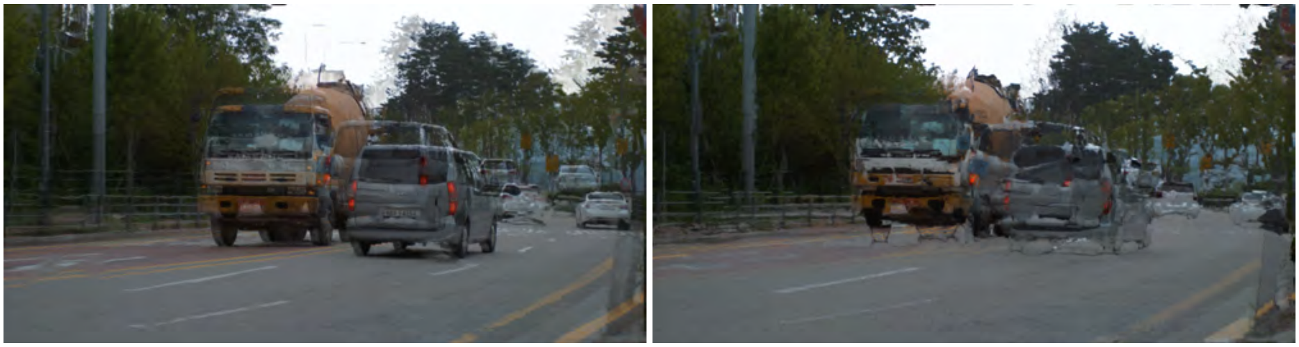
We provide more comparative results on the X4K1000FPS [2], Xiph-4K [41], and BVI-DVC-4K [42] datasets.

B.1. X4K1000FPS



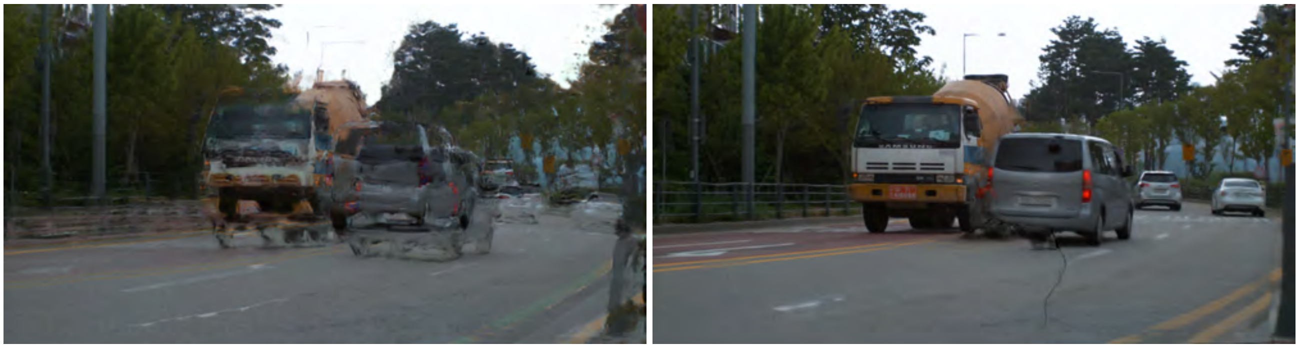
(a) Blended inputs

(b) Ground-Truth



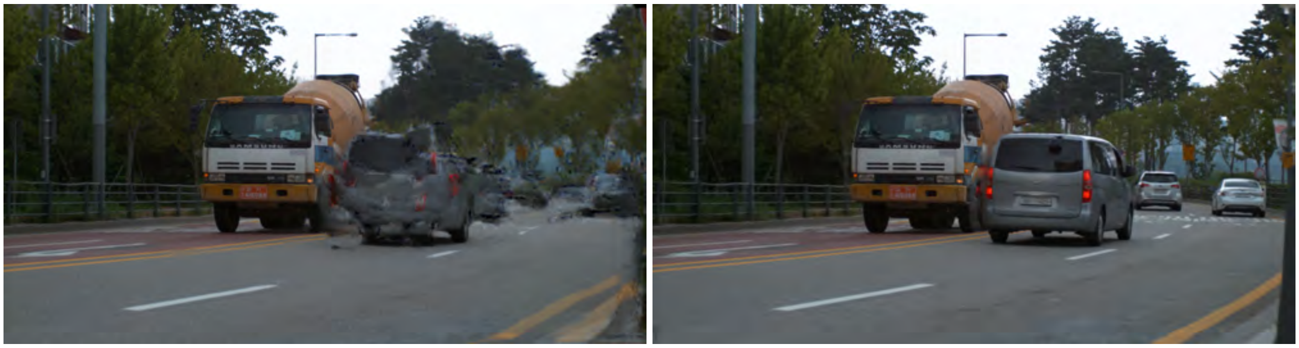
(c) AdaCoF (16.68dB/0.6340)

(d) BMBC (16.65dB/0.6591)



(e) CDFI (17.25dB/0.6501)

(f) XVFI (24.32dB/0.8259)

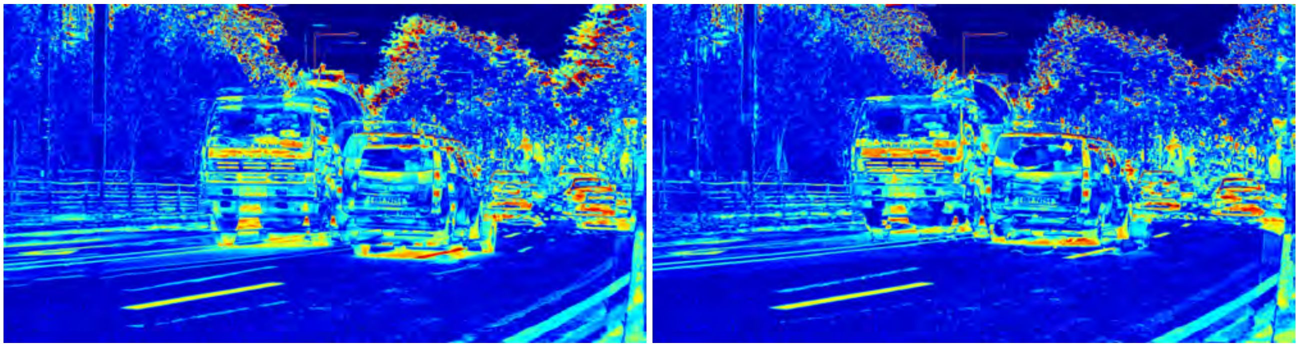


(g) ABME (20.49dB/0.7806)

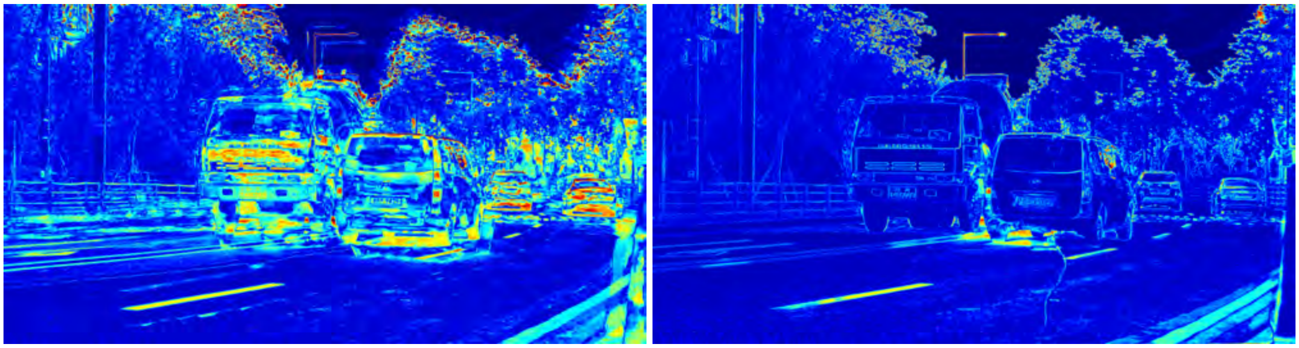
(h) BiFormer (26.07dB/0.8733)

Figure S-3. Qualitative comparison of interpolated frames in the X4K1000FPS dataset.

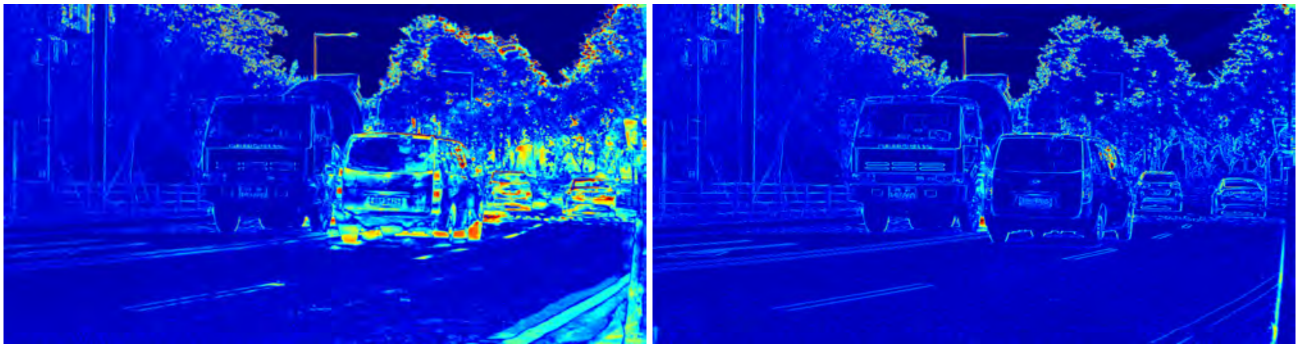
(b) Ground-Truth



(c) AdaCoF (16.68dB/0.6340) (d) BMBC (16.65dB/0.6591)



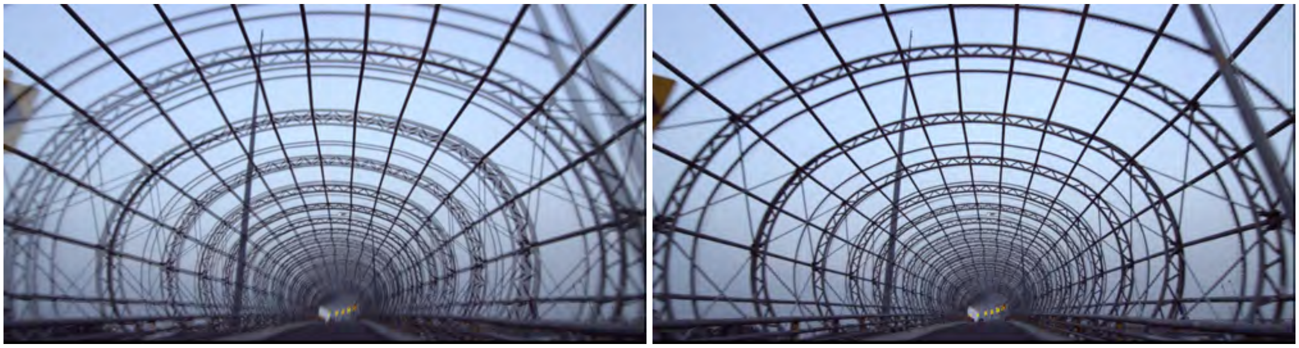
(e) CDFI (17.25dB/0.6501) (f) XVFI (24.32dB/0.8259)



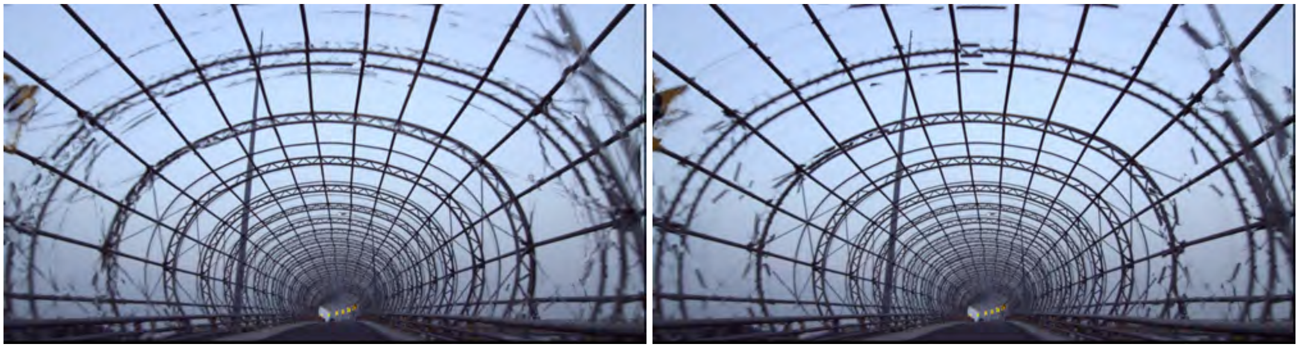
(g) ABME (20.49dB/0.7806) (h) BiFormer (26.07dB/0.8733)

Figure S-4. Error maps of the interpolated frames in Figure S-3.

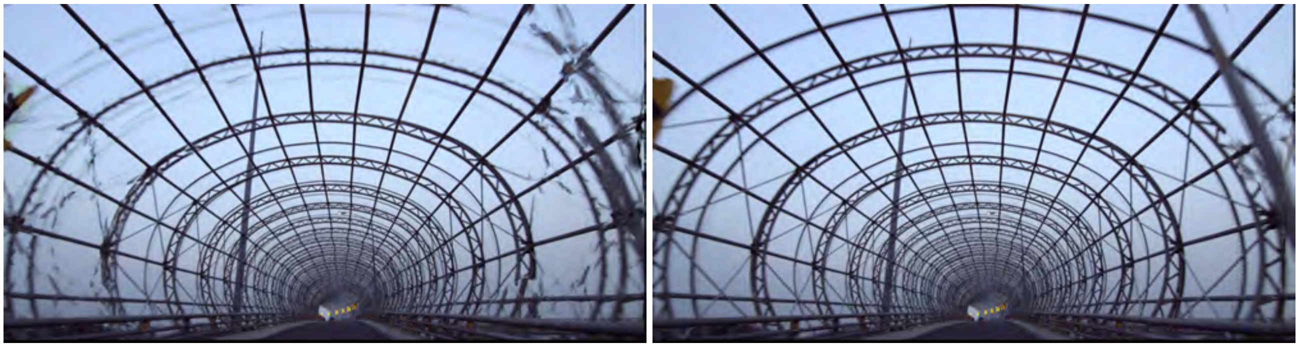
B.2. Xiph-4K



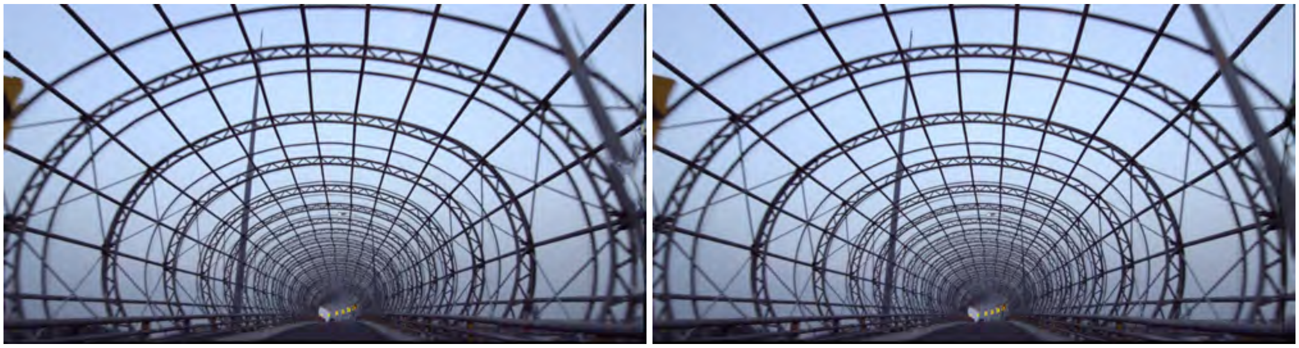
(a) Blended inputs (b) Ground-Truth



(c) AdaCoF (14.99dB/0.8377) (d) BMBC (14.38dB/0.8356)



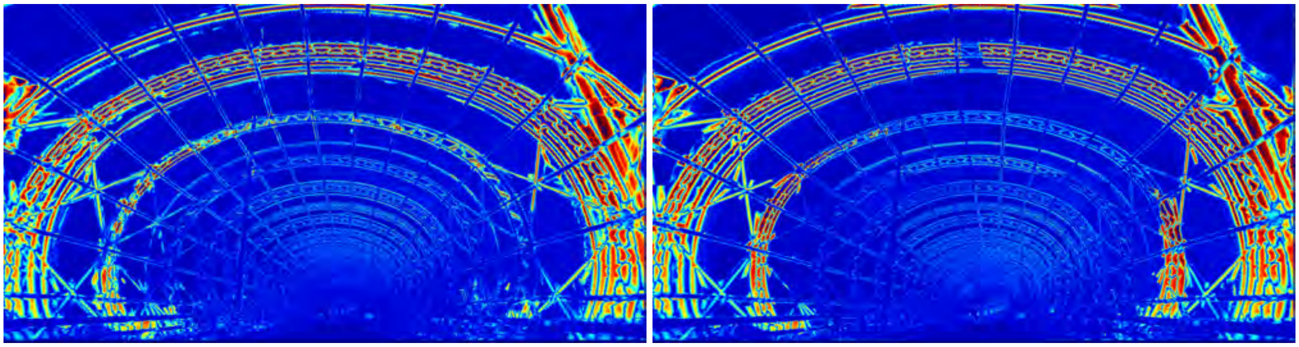
(e) CDFI (15.14dB/0.8481) (f) XVFI (24.95dB/0.9419)



(g) ABME (25.82dB/0.9452) (h) BiFormer (26.07dB/0.9469)

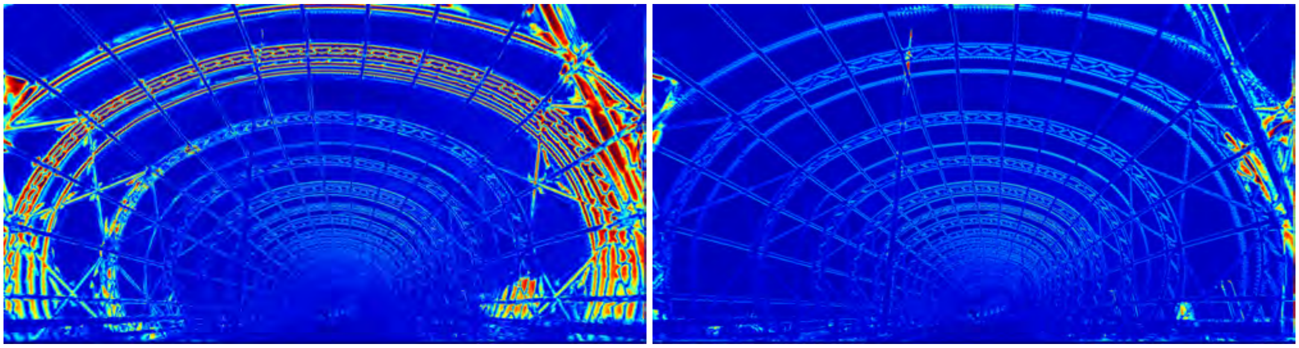
Figure S-5. Qualitative comparison of interpolated frames in the Xiph-4K dataset.

(b) Ground-Truth



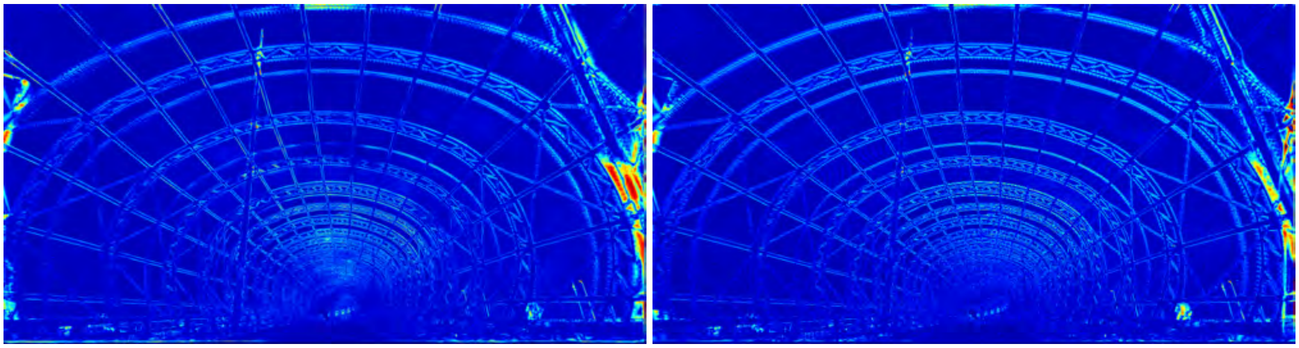
(c) AdaCoF (14.99dB/0.8377)

(d) BMBC (14.38dB/0.8356)



(e) CDFI (15.14dB/0.8481)

(f) XVFI (24.95dB/0.9419)

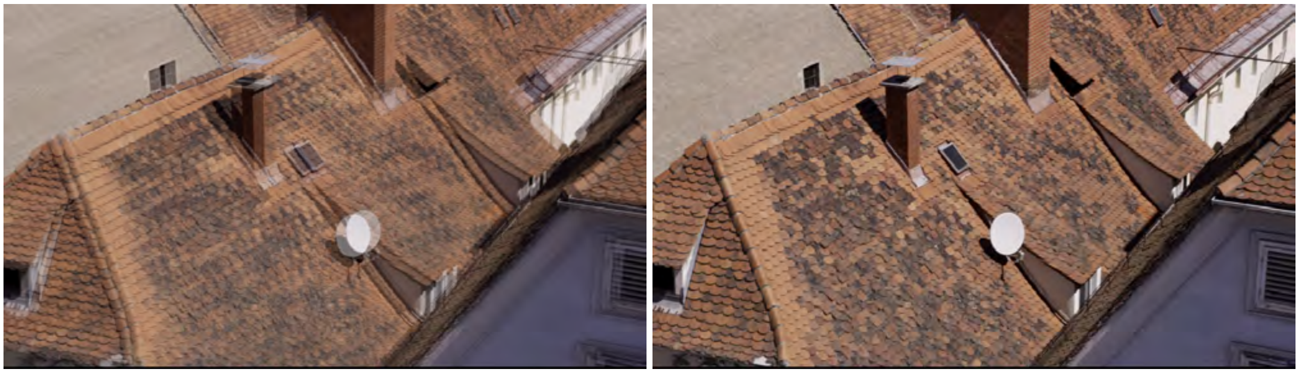


(g) ABME (25.82dB/0.9452)

(h) BiFormer (26.07dB/0.9469)

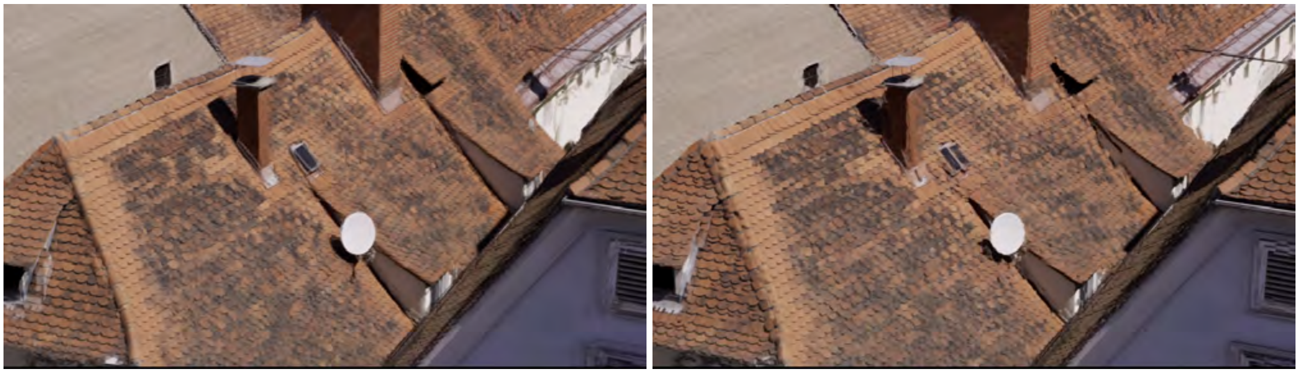
Figure S-6. Error maps of the interpolated frames in Figure S-5.

B.3. BVI-DVC-4K



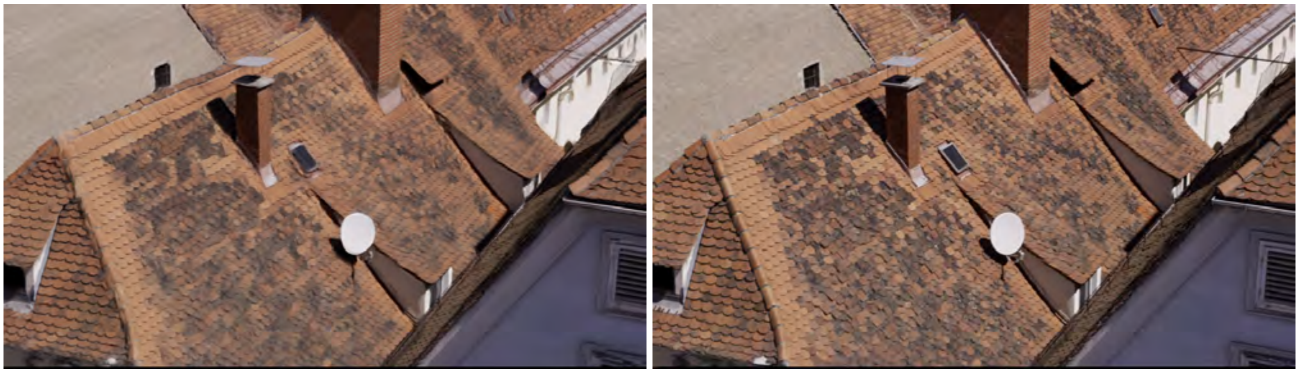
(a) Blended inputs

(b) Ground-Truth



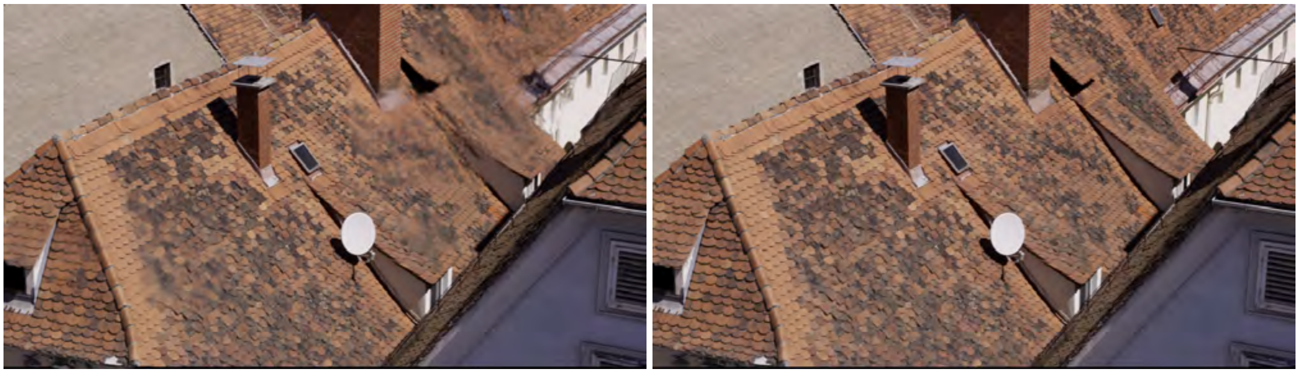
(c) AdaCoF (20.83dB/0.8441)

(d) BMBC (20.15dB/0.8355)



(e) CDFI (21.30dB/0.8457)

(f) XVFI (34.65dB/0.9861)

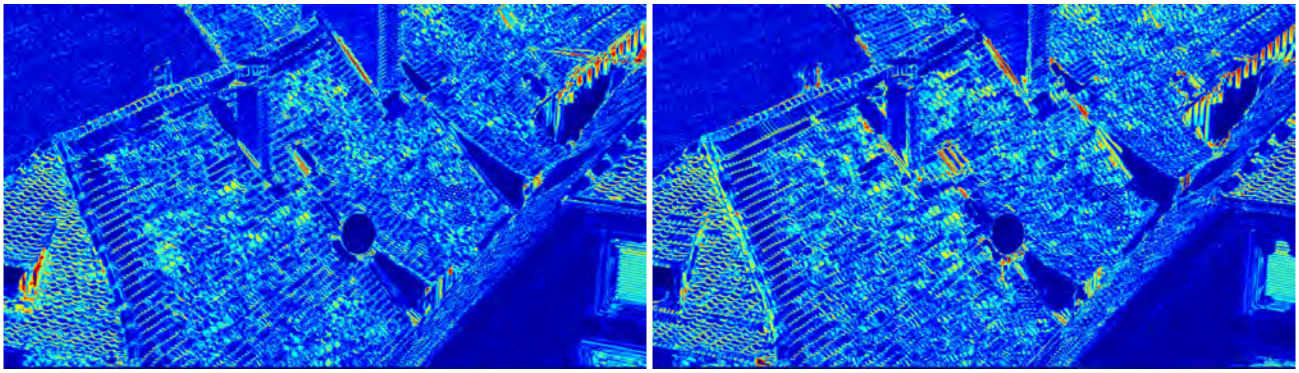


(g) ABME (26.28dB/0.9568)

(h) BiFormer (38.44dB/0.9902)

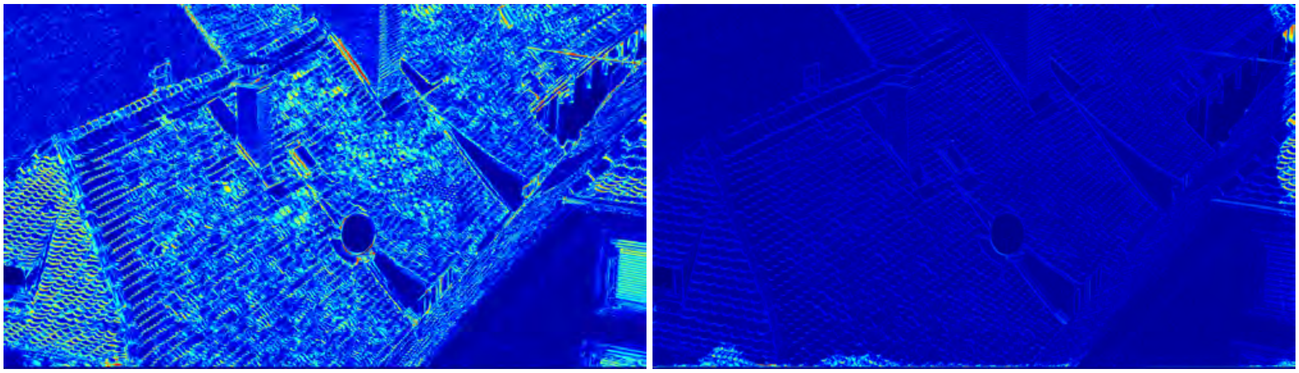
Figure S-7. Qualitative comparison of interpolated frames in the BVI-DVC-4K dataset.

(b) Ground-Truth



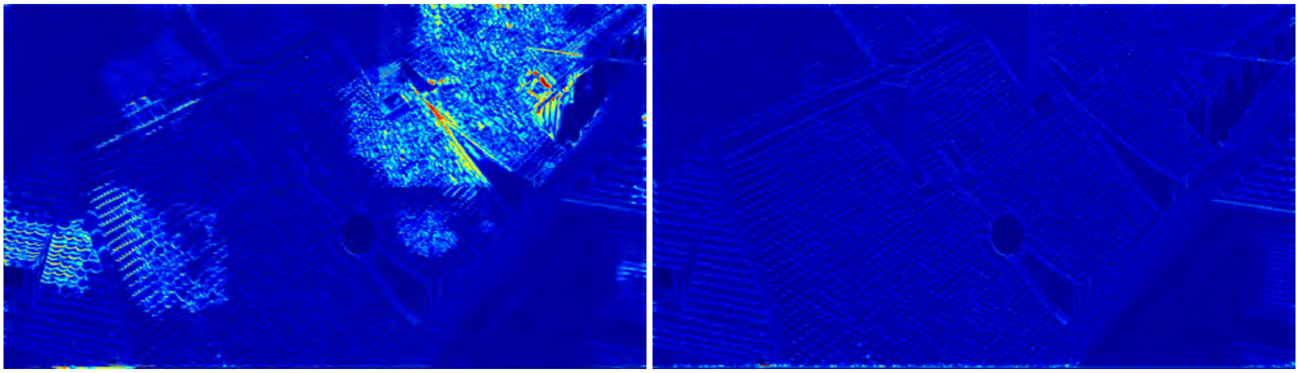
(c) AdaCoF (20.83dB/0.8441)

(d) BMBC (20.15dB/0.8355)



(e) CDFI (21.30dB/0.8457)

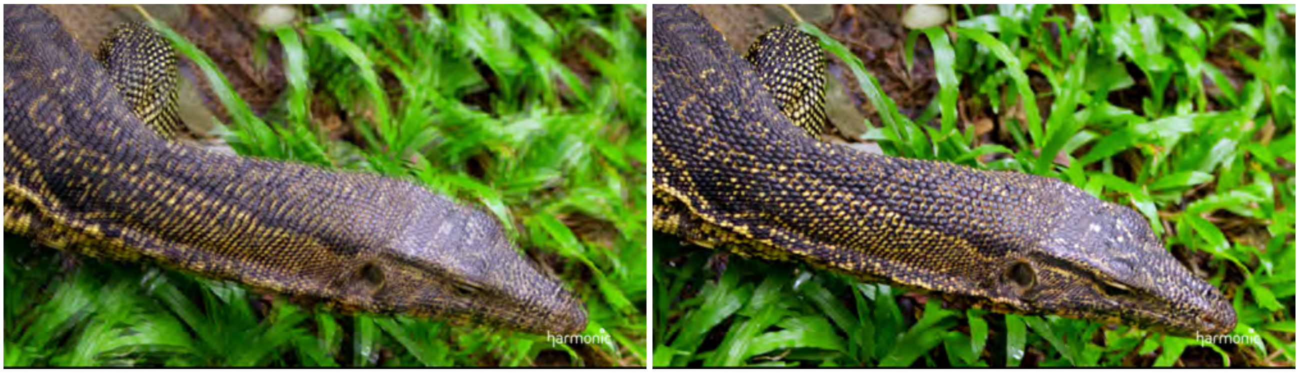
(f) XVFI (34.65dB/0.9861)



(g) ABME (26.28dB/0.9568)

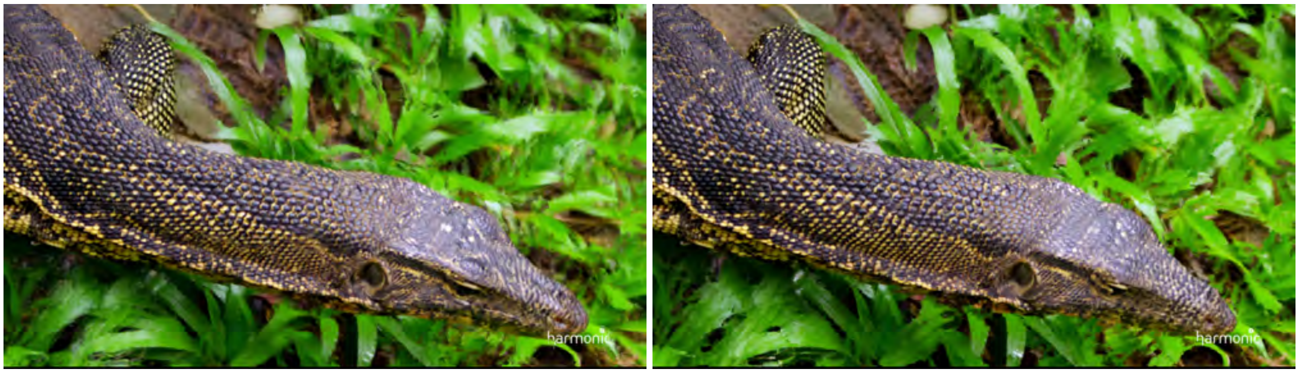
(h) BiFormer (38.44dB/0.9902)

Figure S-8. Error maps of the interpolated frames in Figure S-7.



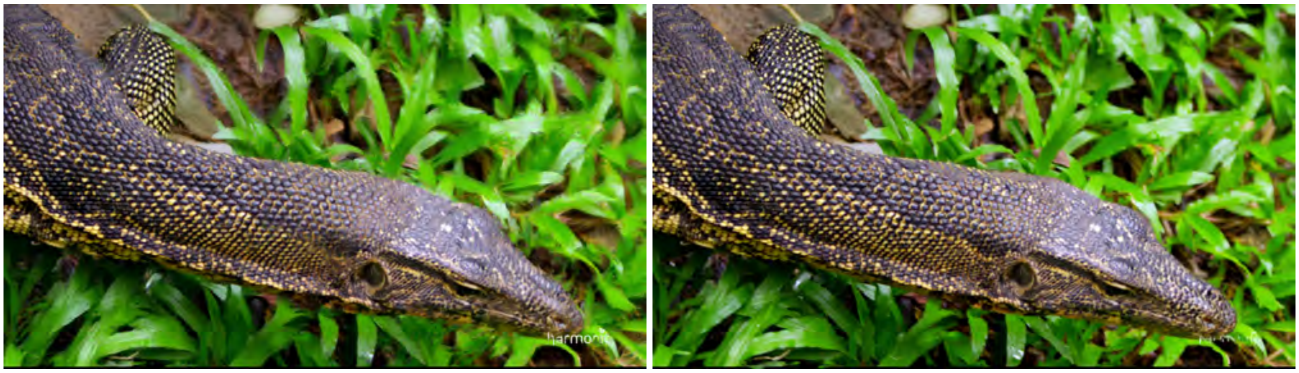
(a) Blended inputs

(b) Ground-Truth



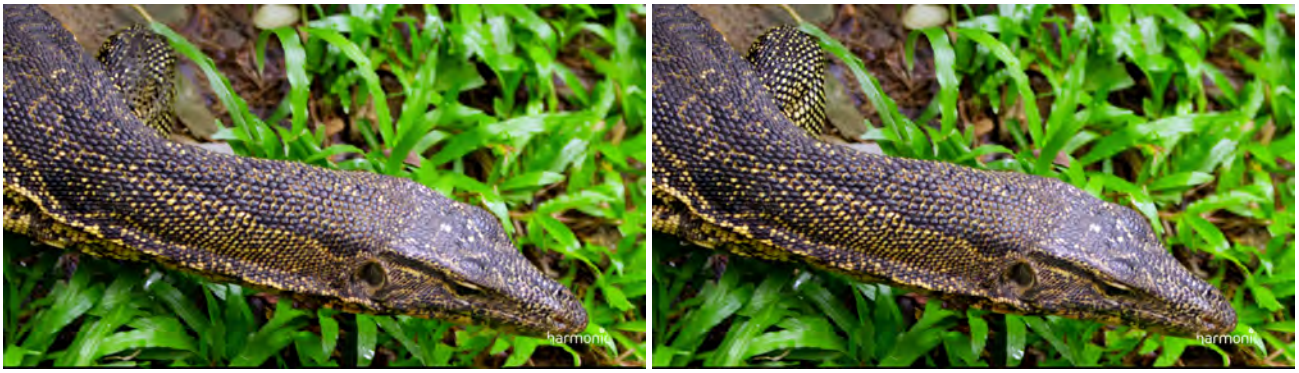
(c) AdaCoF (19.24dB/0.8302)

(d) BMBC (18.19dB/0.7759)



(e) CDFI (19.58dB/0.8085)

(f) XVFI (29.03dB/0.9560)

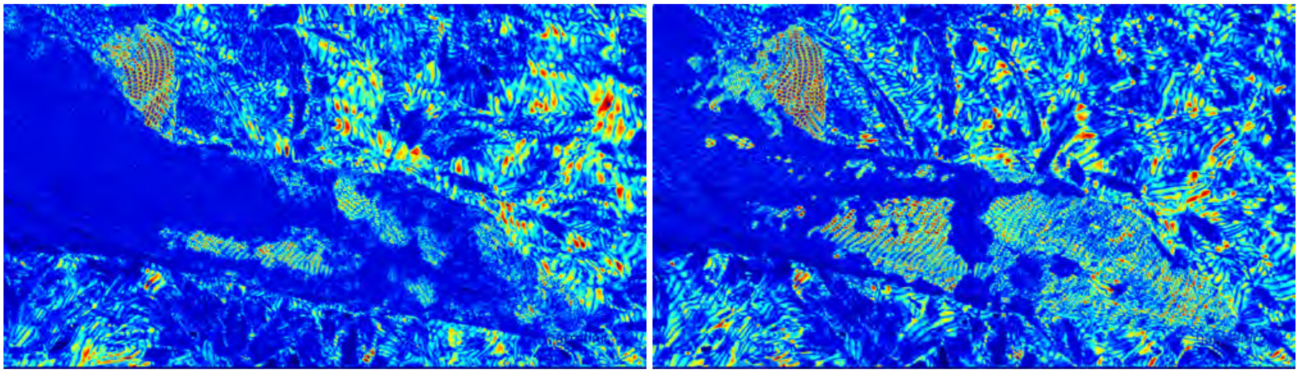


(g) ABME (26.67dB/0.9559)

(h) BiFormer (32.05dB/0.9658)

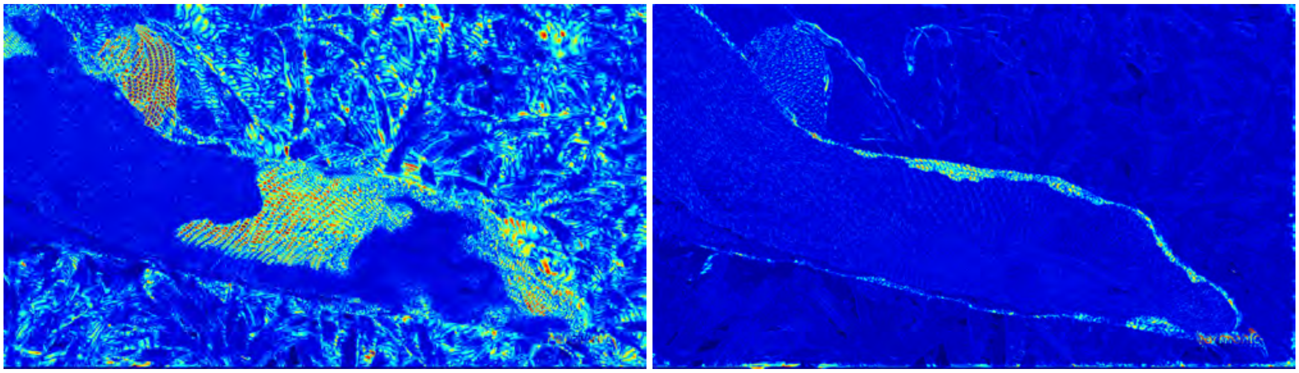
Figure S-9. Qualitative comparison of interpolated frames in the BVI-DVC-4K dataset.

(b) Ground-Truth



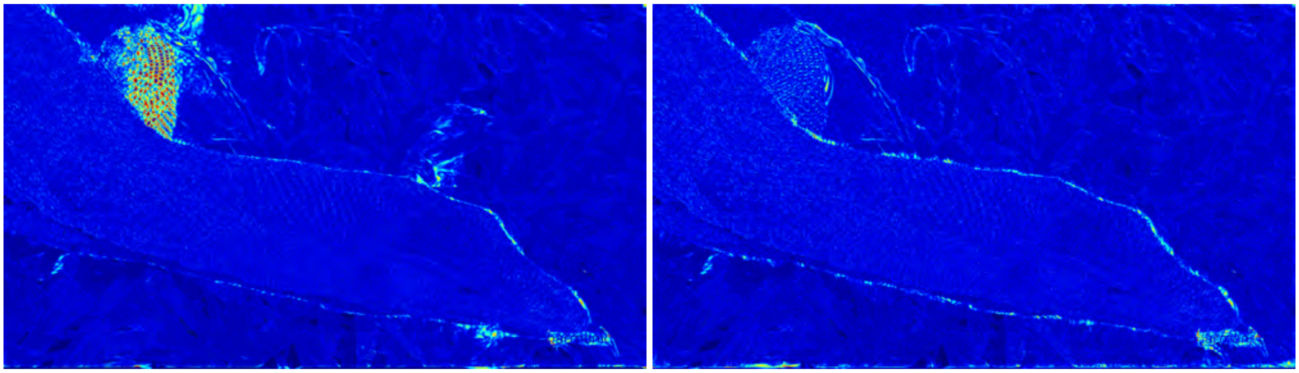
(c) AdaCoF (19.24dB/0.8302)

(d) BMBC (18.19dB/0.7759)



(e) CDFI (19.58dB/0.8085)

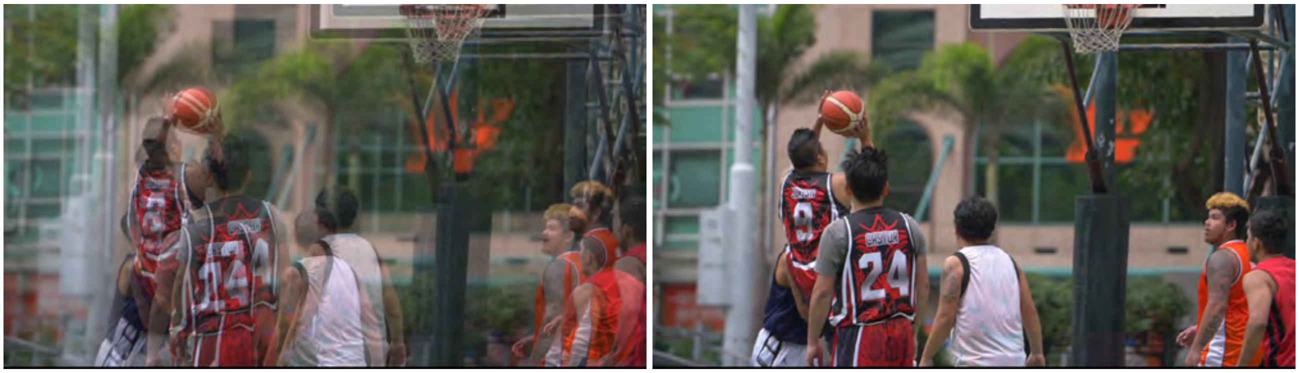
(f) XVFI (29.03dB/0.9560)



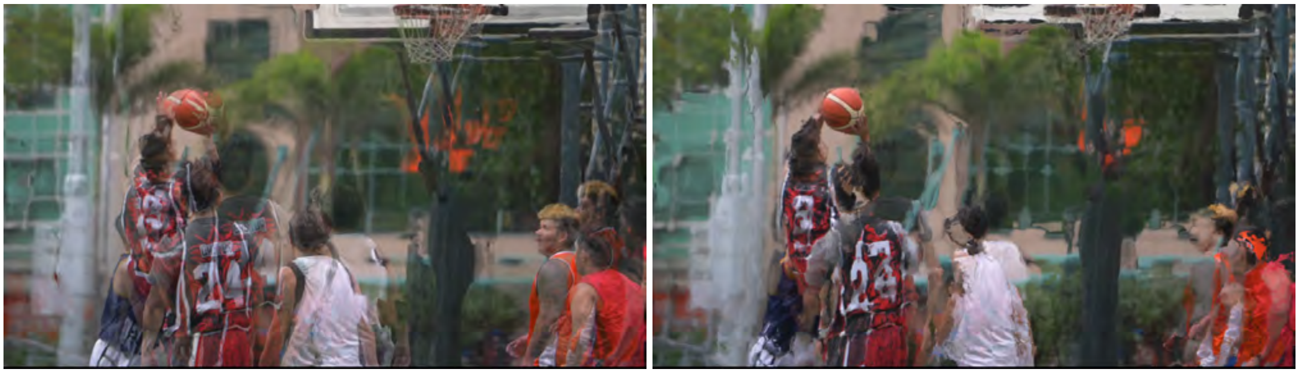
(g) ABME (26.67dB/0.9559)

(h) BiFormer (32.05dB/0.9658)

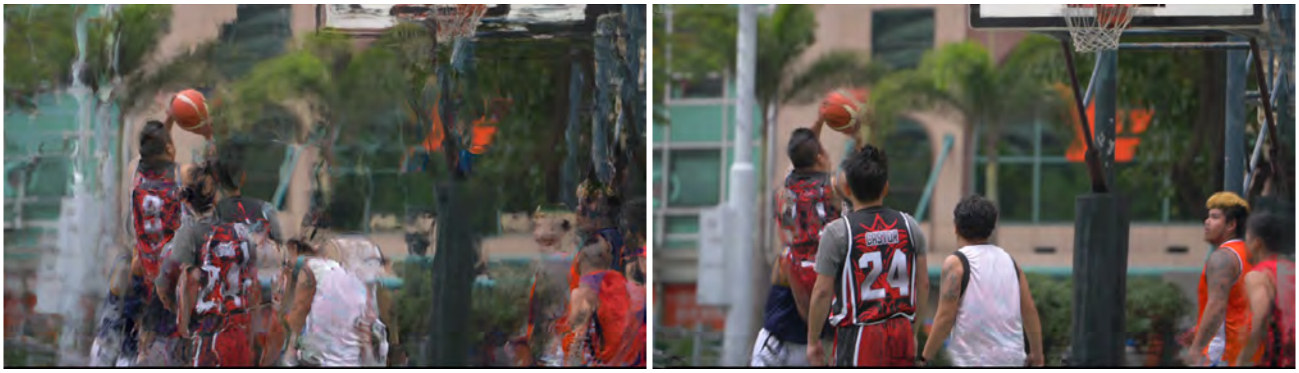
Figure S-10. Error maps of the interpolated frames in Figure S-9.



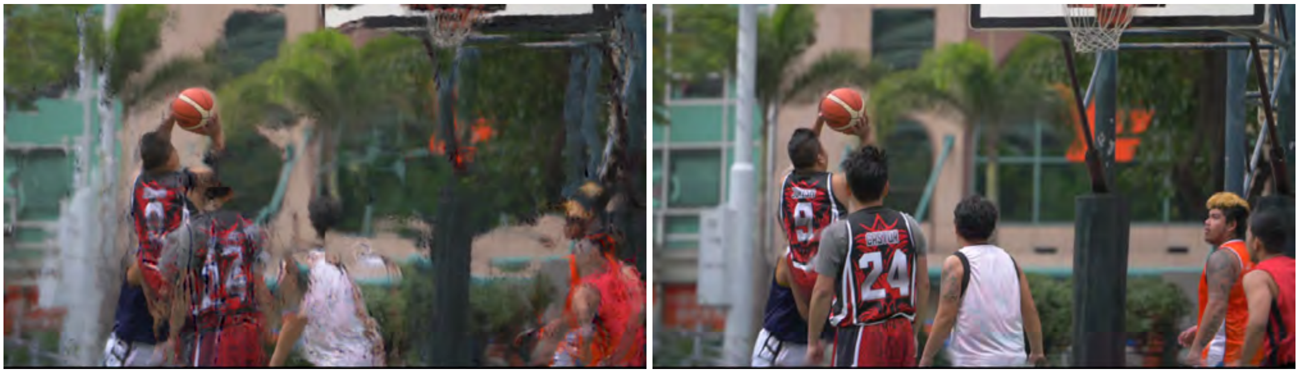
(a) Blended inputs (b) Ground-Truth



(c) AdaCoF (15.36dB/0.5784) (d) BMBC (16.24dB/0.6434)



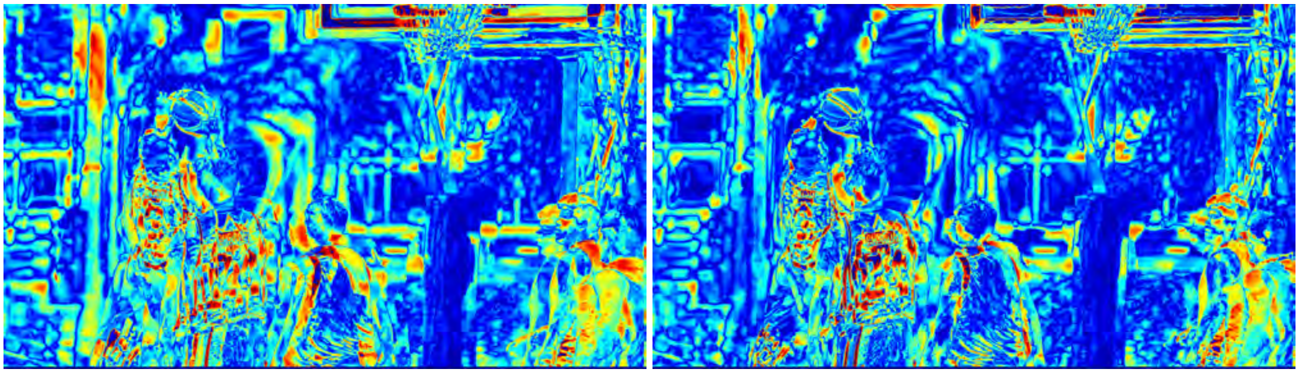
(e) CDFI (15.82dB/0.6019) (f) XVFI (23.78dB/0.9050)



(g) ABME (17.03dB/0.6807) (h) BiFormer (26.53dB/0.9343)

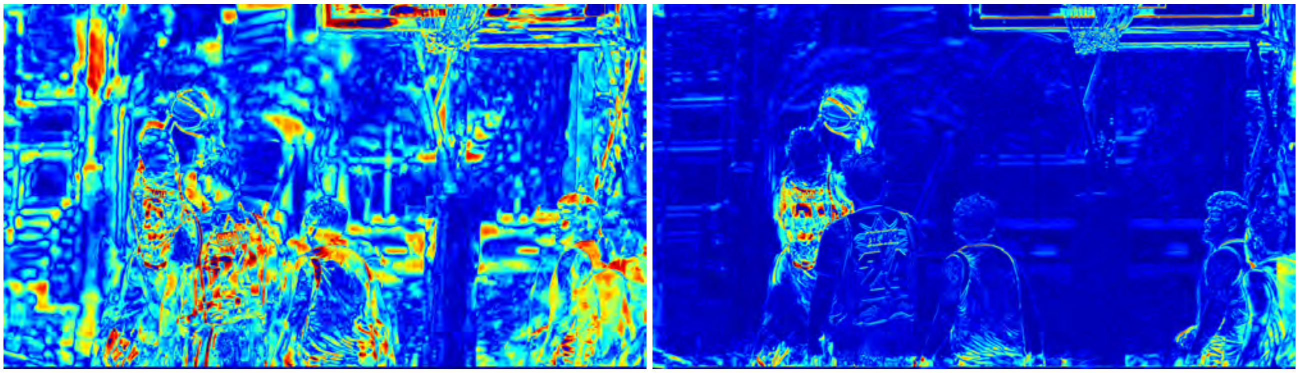
Figure S-11. Qualitative comparison of interpolated frames in the BVI-DVC-4K dataset.

(b) Ground-Truth



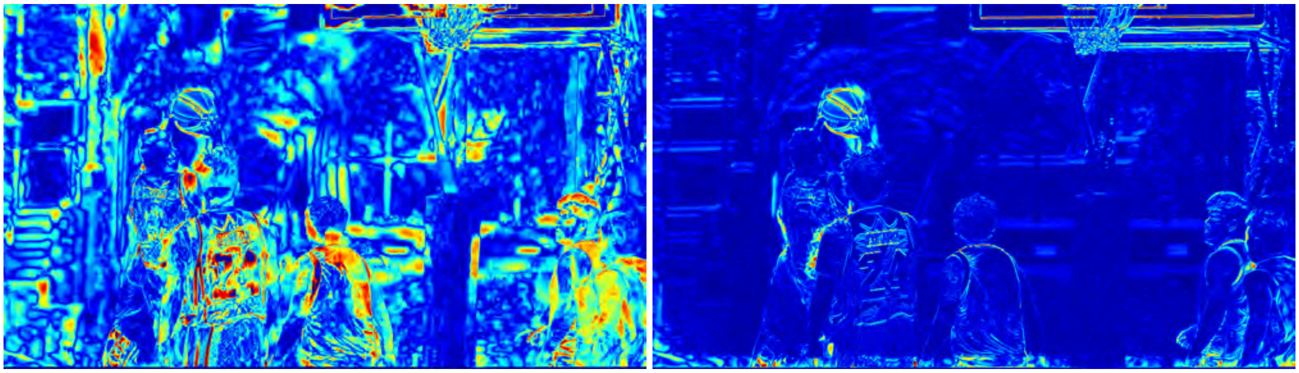
(c) AdaCoF (15.36dB/0.5784)

(d) BMBC (16.24dB/0.6434)



(e) CDFI (15.82dB/0.6019)

(f) XVFI (23.78dB/0.9050)



(g) ABME (17.03dB/0.6807)

(h) BiFormer (26.53dB/0.9343)

Figure S-12. Error maps of the interpolated frames in Figure S-11.