Supplemental Materials on

Asymmetric Bilateral Motion Estimation for Video Frame Interpolation

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

Chul Lee
Dongguk University
chullee@dongguk.edu

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

S-1. Network Details

S-1.1. ABMR-Net

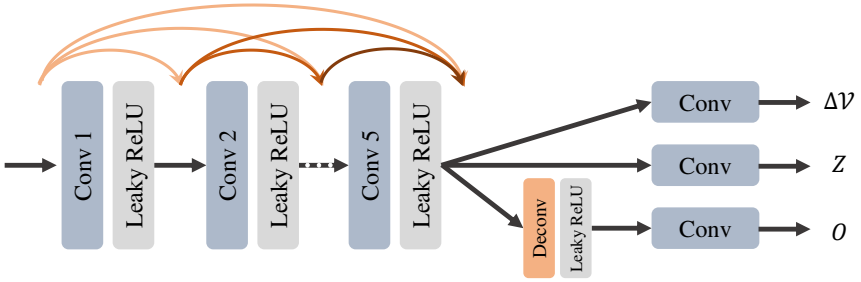


Figure S-1: Architecture of the estimator in ABMR-Net.

Figure S-1 shows a multi-layer CNN, which is denoted as *Net* in ABMR-Net in Figure 5 in the main paper. It consists of five convolution layers with dense connections. All convolution layers have 3×3 kernels. The numbers of feature channels at the five convolution layers are 64, 64, 48, 32, and 16. It takes the cost volume, feature maps of the source frame, and up-sampled motion field as input. At the final stage, the residual motion field $\Delta\mathcal{V}$, reliability mask Z, and offset map O are estimated by the corresponding convolution layers. Especially, different from $\Delta\mathcal{V}$ and Z, the feature map is up-sampled before the convolution for estimating O. Different levels have their own parameters, instead of sharing the same parameters.

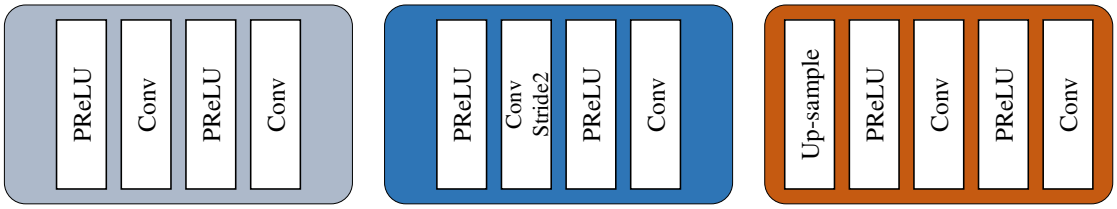
S-1.2. Feature Extractor in Frame Synthesis Network

Table S-1: Architecture of the feature extractor in the frame synthesis network.

Layer	# of channels	Kernel size	Stride	Padding size
Conv1_0	32	3×3	1×1	1×1
PReLU1_0	-	-	-	-
Conv1_1	32	3×3	1×1	1×1
PReLU1_1	-	-	-	-
Conv2_0	64	3×3	2×2	1×1
PReLU2_0	-	-	-	-
Conv2_1	64	3×3	1×1	1×1
PReLU2_1	-	-	-	-
Conv3_0	96	3×3	2×2	1×1
PReLU3_0	-	-	-	-
Conv3_1	96	3×3	1×1	1×1
PReLU3_1	-	-	-	-

Table S-1 shows the architecture of the feature extractor in the frame synthesis network in Figure 4 in the main paper. All layers are connected in series. Conv1_0 takes a single frame as input. The outputs of PReLU1_1, PReLU2_1, and PReLU3_1 are C^3, C^2, and C^1, respectively.

S-1.3. Modified GridNet



(a) Lateral block (b) Down block (c) Up block

Figure S-2: Detailed architectures of the lateral, down, and up blocks in the modified GridNet.

Figure S-2 shows the architectures of the blocks in the modified version of GridNet in Figure 6 in the main paper, which is used as the backbone of FilterNet and RefineNet. All convolution layers in Figure S-2 have 3×3 kernels. As shown in Figure 6 in the main paper, the modified GridNet consists of 19 lateral blocks, 6 down blocks, and 6 up blocks. Figure S-2(a) is the architecture of the lateral block. The first lateral block at each level does not contain the first PReLU layer. At level 3, the second convolution layer of the last lateral block has $5 \times 5 \times 4$ and 3 feature channels for FilterNet and RefineNet, respectively. The number of feature channels at convolution layers of lateral blocks are 96, 64, and 32 at level 1, 2, and 3, respectively. In Figures S-2(b) and (c), the convolution layer with stride 2 and bilinear up-sampling are employed in the down block and up block. The numbers of feature channels at convolution layers of down blocks and up blocks between levels 2 and 3 are 64 and 32, respectively, while those between levels 1 and 2 are 96 and 64, respectively.

S-2. Anchor Frame Interpolation

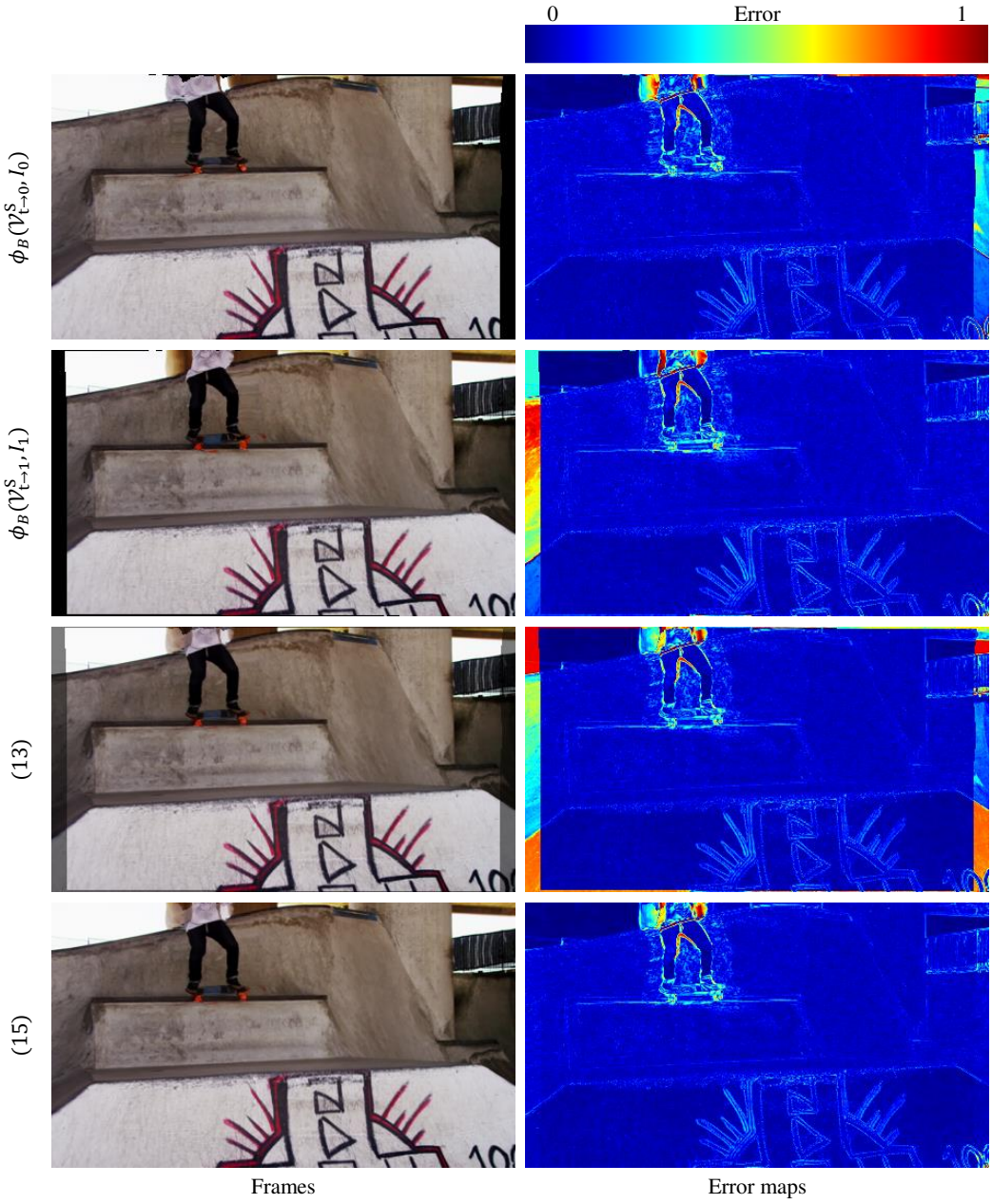


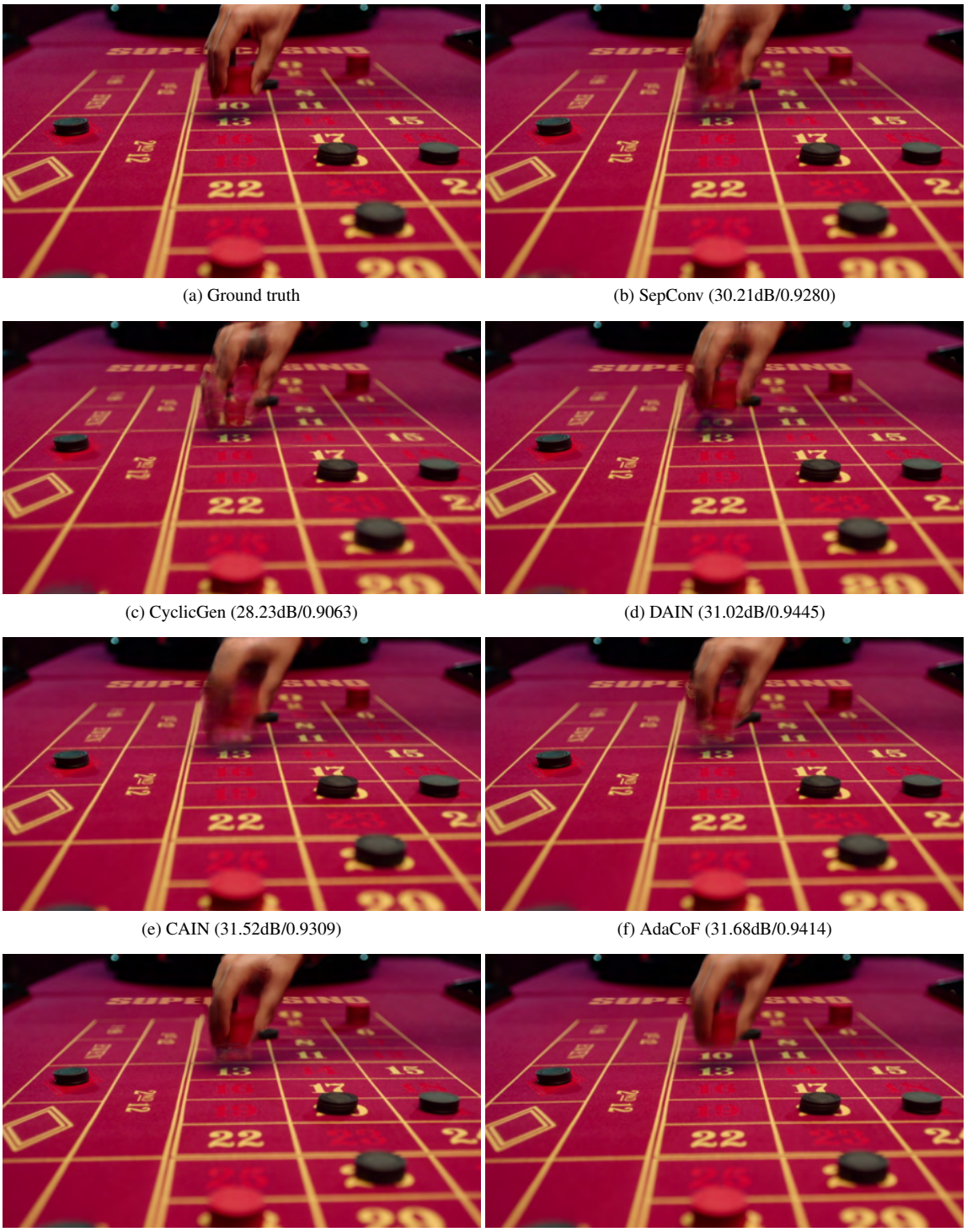
Figure S-3: Comparison of different anchor frames and the corresponding error maps.

Figure S-3 shows two backwardly warped frames using the symmetric bilateral motion fields $\mathcal{V}^S_{t\to 0}$ and $\mathcal{V}^S_{t\to 1}$, two reconstructed anchor frames using Eqs. (13) and (15), respectively, and their error maps. Due to camera panning, the two warped frames using the symmetric fields in the top two rows have missing parts at frame boundaries. Thus, the simple blending in Eq. (13) also causes errors in those regions. On the contrary, the proposed anchor frame interpolation in Eq. (15) reconstructs the anchor frame more faithfully by exploiting the occlusion-aware masks.

S-3. More Experimental Results

Let us provide more comparative results on the Vimeo90K [3], SNU-FILM [1], and Xiph [2] datasets.

S-3.1. Vimeo90K



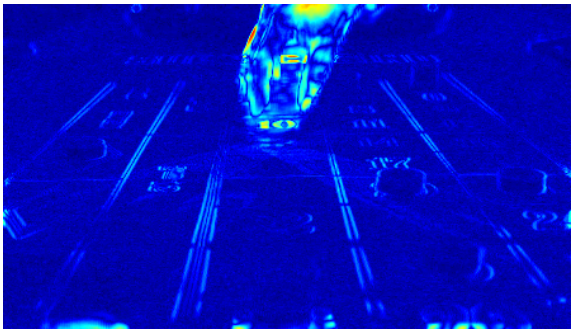
(a) Ground truth (b) SepConv (30.21dB/0.9280)

(c) CyclicGen (28.23dB/0.9063) (d) DAIN (31.02dB/0.9445)

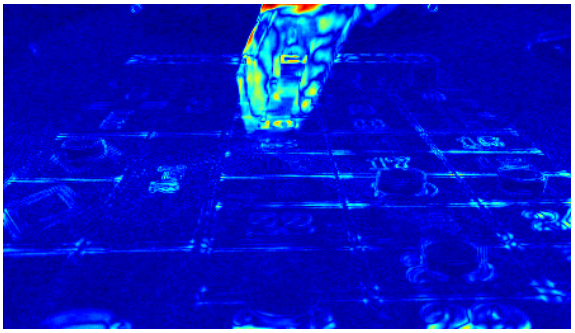
(e) CAIN (31.52dB/0.9309) (f) AdaCoF (31.68dB/0.9414)

(g) BMBC (32.00/0.9461) (h) ABME (33.99dB/0.9588)

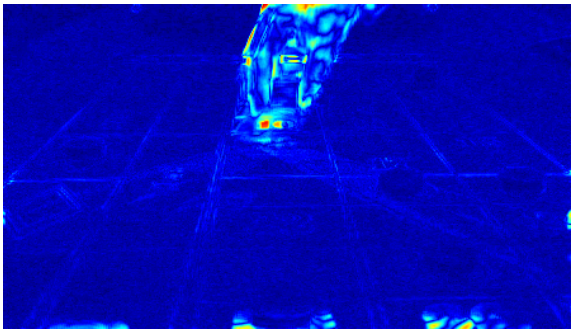
Figure S-4: Qualitative comparison of interpolated frames in the Vimeo90K dataset.



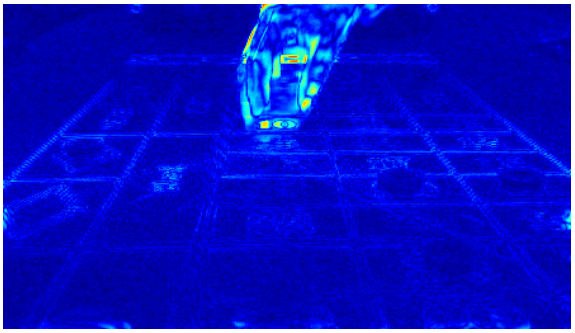
(b) SepConv (30.21dB/0.9280)



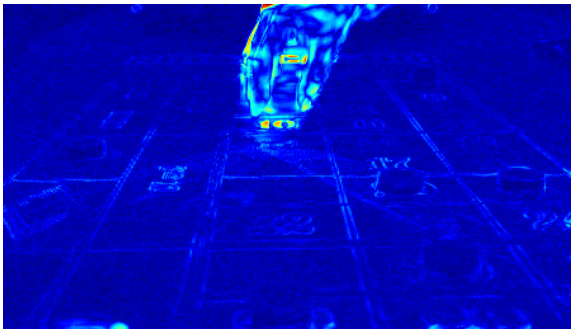
(c) CyclicGen (28.23dB/0.9063)



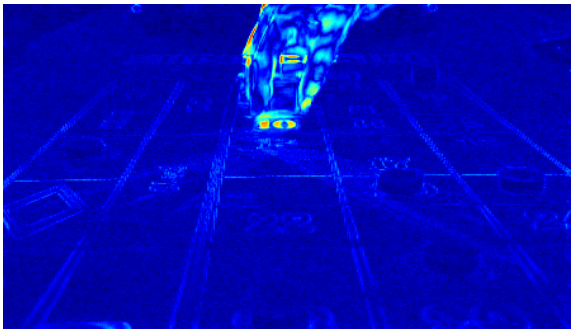
(d) DAIN (31.02dB/0.9445)



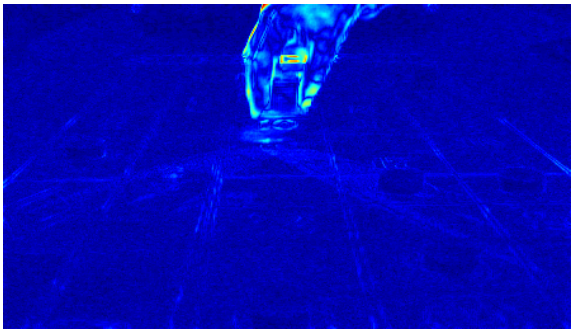
(e) CAIN (31.52dB/0.9309)



(f) AdaCoF (31.68dB/0.9414)



(g) BMBC (32.00/0.9461)



(h) ABME (33.99dB/0.9588)

Figure S-5: Error maps of the interpolated frames in Figure S-4.

(a) Ground truth

(b) SepConv (28.07dB/0.9303)

(c) CyclicGen (24.58dB/0.8382)

(d) DAIN (29.41dB/0.9486)

(e) CAIN (28.33dB/0.9272)

(f) AdaCoF (28.12dB/0.9239)

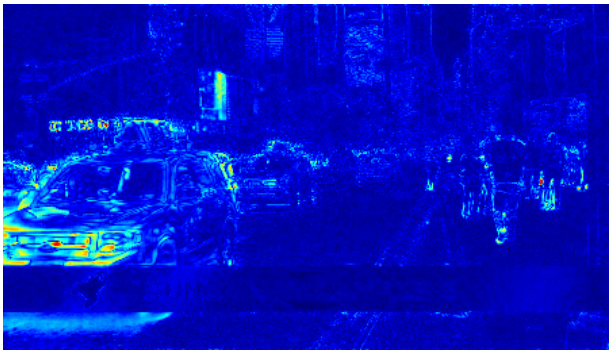
(g) BMBC (29.97dB/0.9525)

(h) ABME (31.67dB/0.9657)

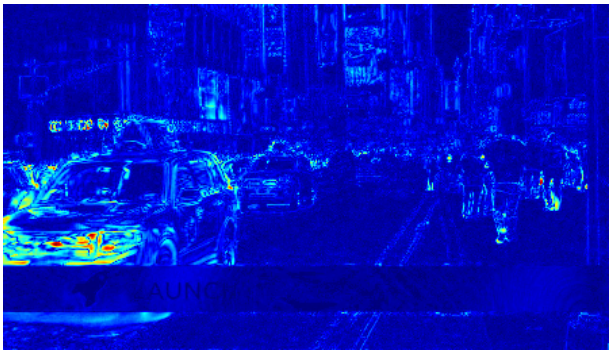
Figure S-6: Qualitative comparison of interpolated frames in the Vimeo90K dataset.

(c) CyclicGen (24.58dB/0.8382)

(d) DAIN (29.41dB/0.9486)



(e) CAIN (28.33dB/0.9272)



(f) AdaCoF (28.12dB/0.9239)

(g) BMBC (29.97dB/0.9525)

(h) ABME (31.67dB/0.9657)

Figure S-7: Error maps of the interpolated frames in Figure S-6.

(a) Ground truth

(b) SepConv (33.26dB/0.9554)

(c) CyclicGen (31.19dB/0.9302)

(d) DAIN (34.56dB/0.9660)

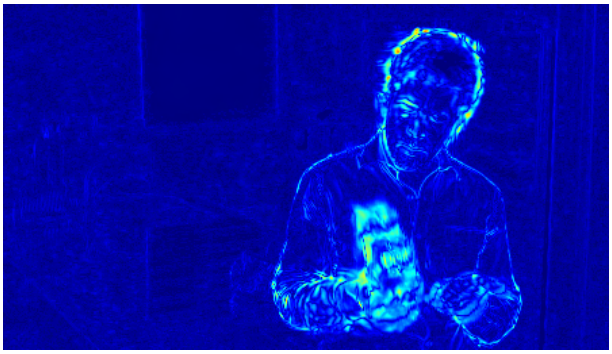
(e) CAIN (33.51dB/0.9576)

(f) AdaCoF (34.12dB/0.9598)

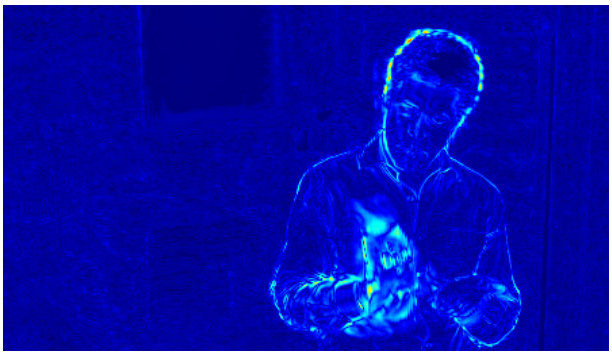
(g) BMBC (33.98dB/0.9603)

(h) ABME (36.18dB/0.9741)

Figure S-8: Qualitative comparison of interpolated frames in the Vimeo90K dataset.

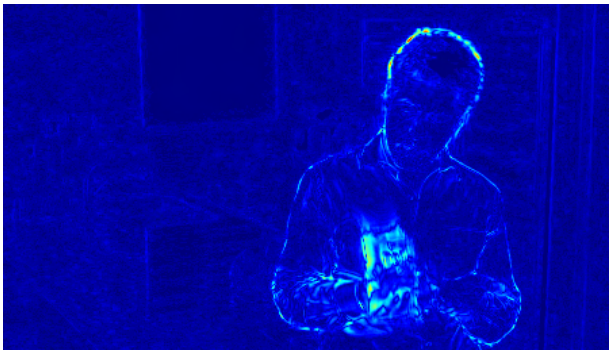


(a) Ground truth

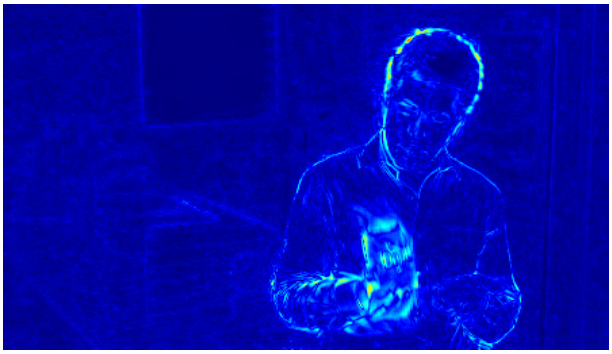


(b) SepConv (33.26dB/0.9554)

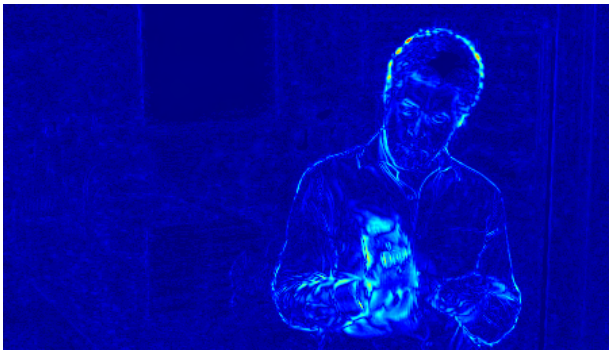
(c) CyclicGen (31.19dB/0.9302)



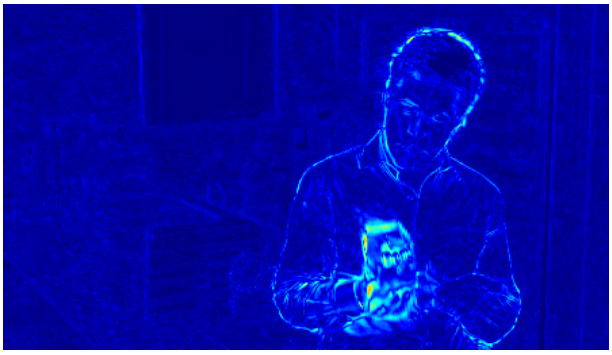
(d) DAIN (34.56dB/0.9660)



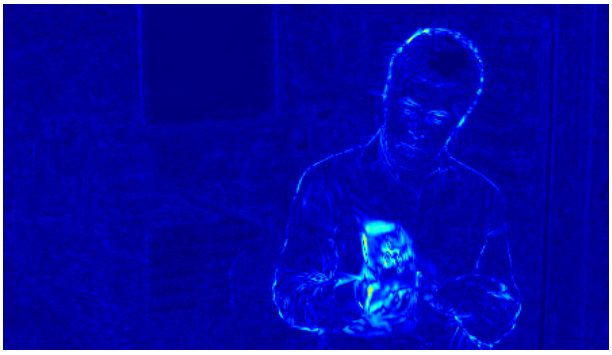
(e) CAIN (33.51dB/0.9576)



(f) AdaCoF (34.12dB/0.9598)



(g) BMBC (33.98dB/0.9603)



(h) ABME (36.18dB/0.9741)

Figure S-9: Error maps of the interpolated frames in Figure S-6.

S-3.2. SNU-FILM

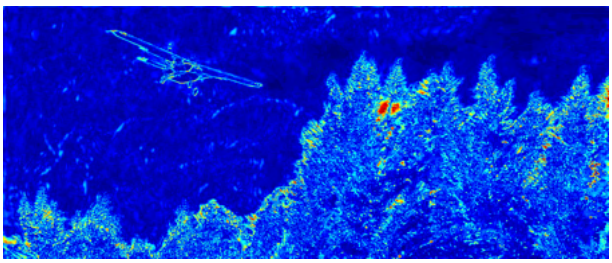
(a) Ground truth (b) SepConv (23.77dB/0.6641)

(c) CyclicGen (18.94dB/0.5484) (d) DAIN (28.21dB/0.8748)

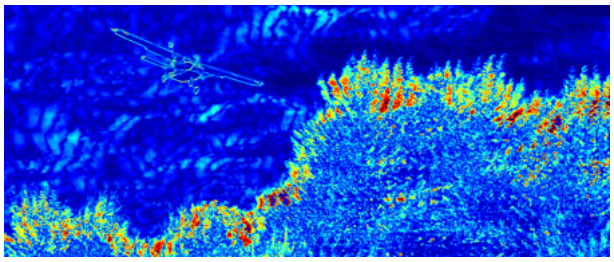
(e) CAIN (26.94dB/0.7693) (f) AdaCoF (19.95dB/0.6139)

(g) BMBC (23.95dB/0.7711) (h) ABME (28.94dB/0.8859)

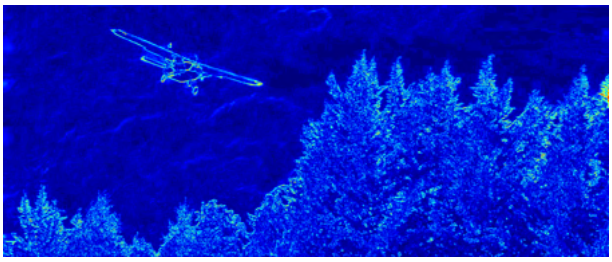
Figure S-10: Qualitative comparison of interpolated frames in the SNU-FILM Extreme dataset.



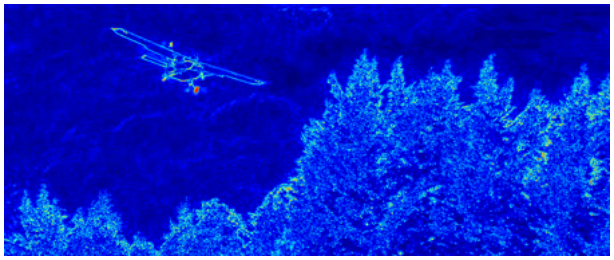
(b) SepConv (23.77dB/0.6641)



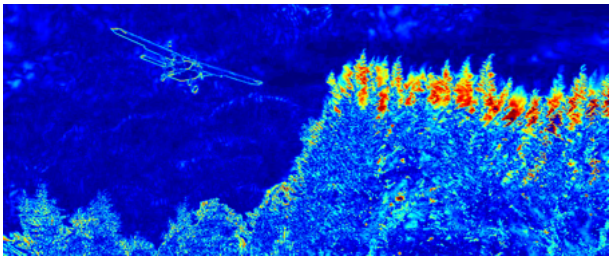
(c) CyclicGen (18.94dB/0.5484)



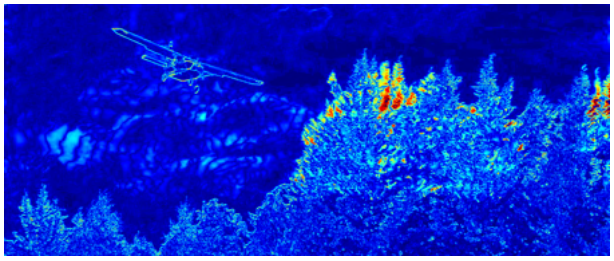
(d) DAIN (28.21dB/0.8748)



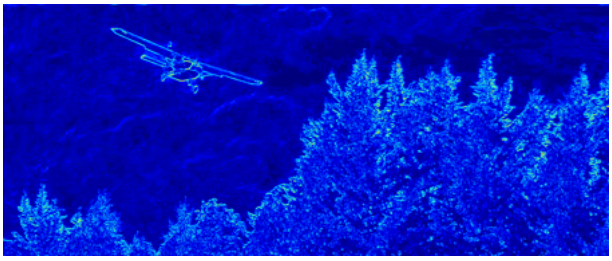
(e) CAIN (26.94dB/0.7693)



(f) AdaCoF (19.95dB/0.6139)



(g) BMBC (23.95dB/0.7711)



(h) ABME (28.94dB/0.8859)

Figure S-11: Error maps of the interpolated frames in Figure S-10.

(a) Ground truth

(b) SepConv (23.74dB/0.9198)

(c) CyclicGen (21.27dB/0.7540)

(d) DAIN (26.46dB/0.9456)

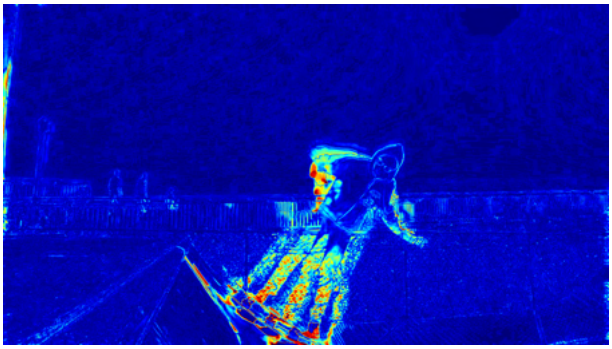
(e) CAIN (24.27dB/0.9319)

(f) AdaCoF (23.17dB/0.8961)

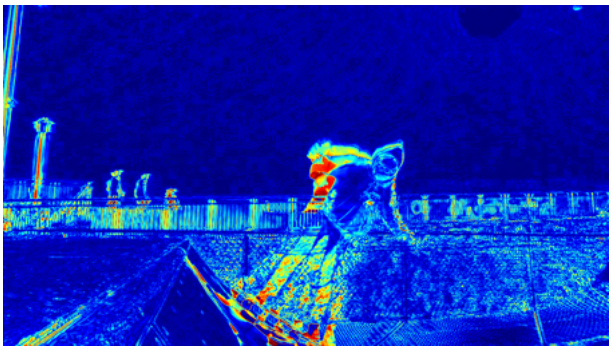
(g) BMBC (22.10dB/0.8737)

(h) ABME (28.89dB/0.9480)

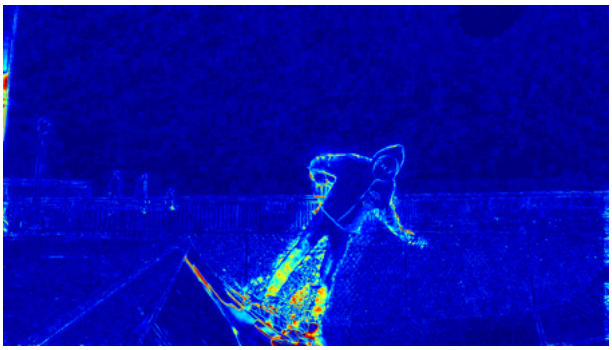
Figure S-12: Qualitative comparison of interpolated frames in the SNU-FILM Extreme dataset.



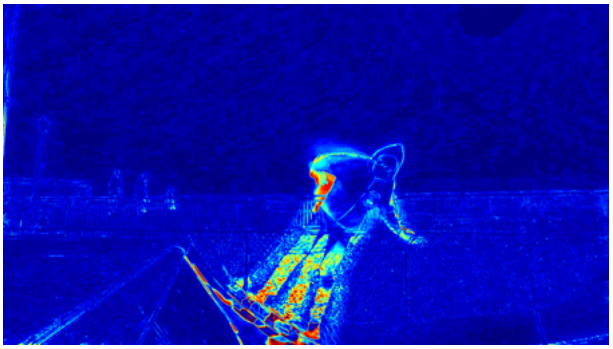
(b) SepConv (23.74dB/0.9198)



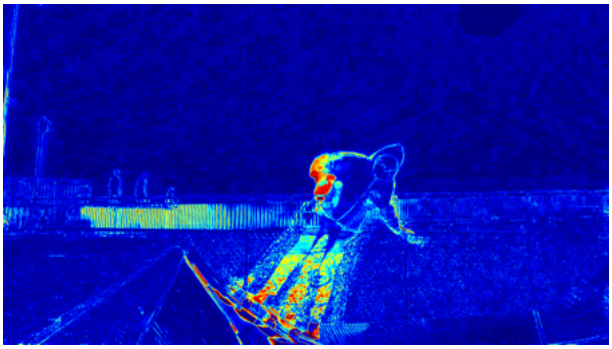
(c) CyclicGen (21.27dB/0.7540)



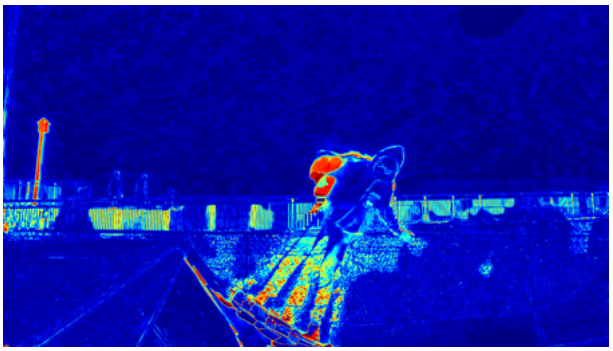
(d) DAIN (26.46dB/0.9456)



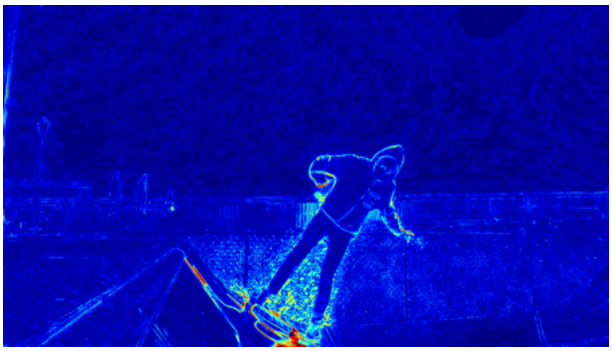
(e) CAIN (24.27dB/0.9319)



(f) AdaCoF (23.17dB/0.8961)



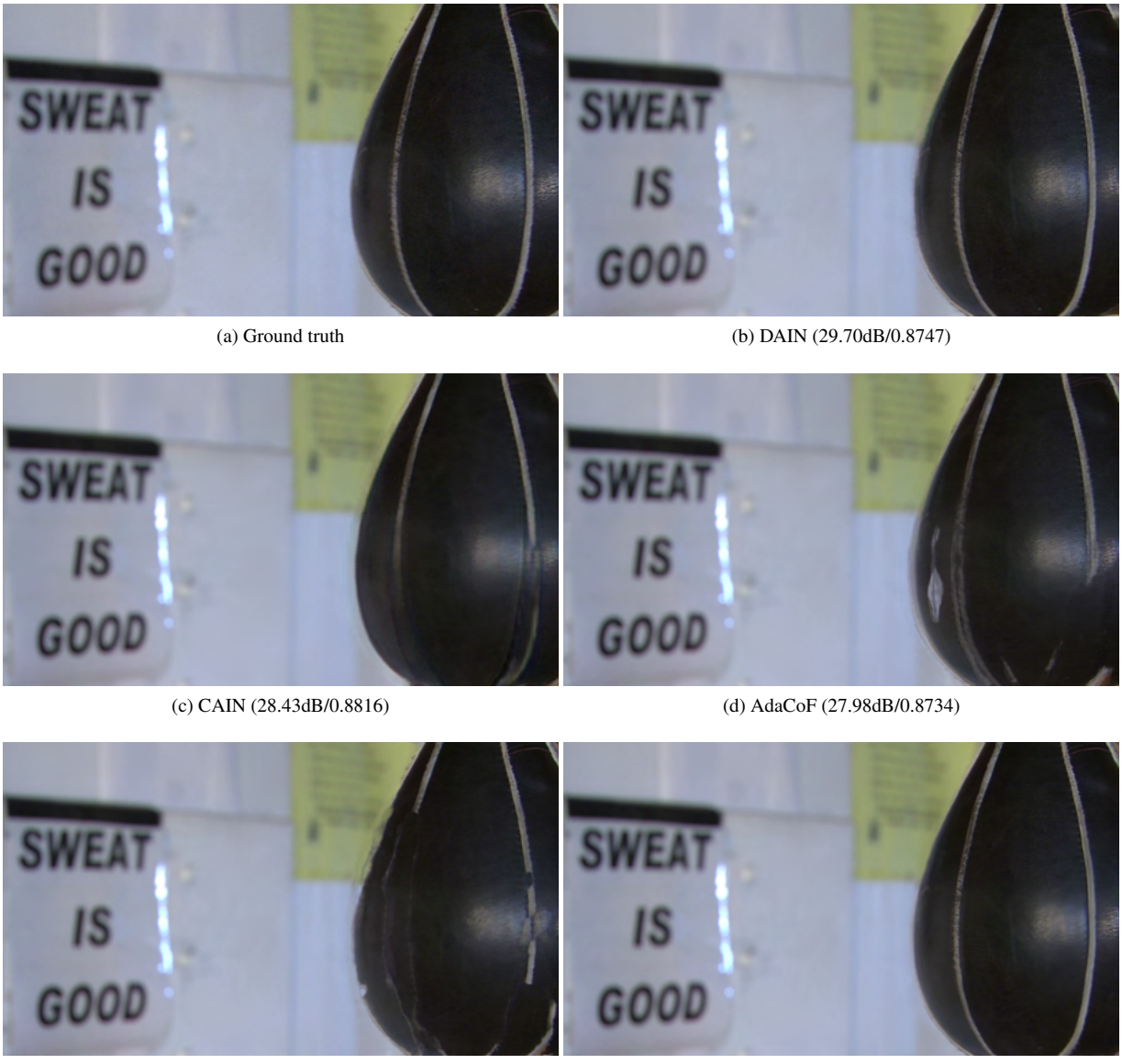
(g) BMBC (22.10dB/0.8737)



(h) ABME (28.89dB/0.9480)

Figure S-13: Error maps of the interpolated frames in Figure S-12.

S-3.3. Xiph

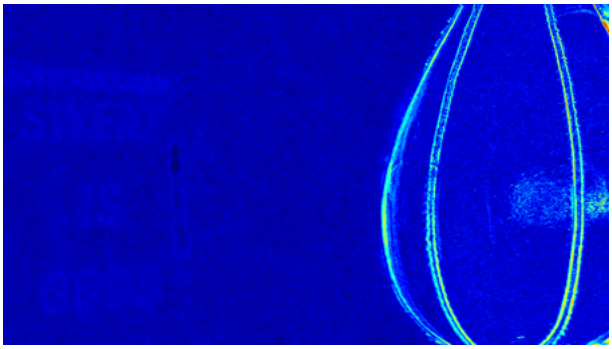


(a) Ground truth (b) DAIN (29.70dB/0.8747)

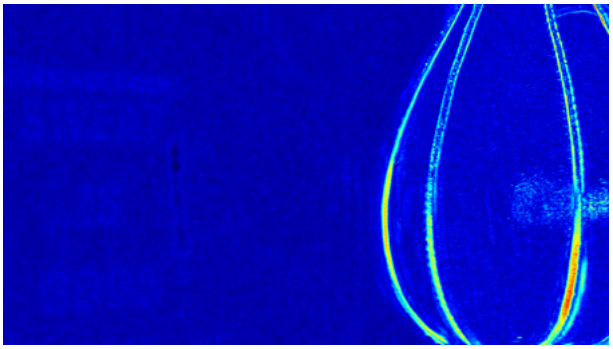
(c) CAIN (28.43dB/0.8816) (d) AdaCoF (27.98dB/0.8734)

(e) BMBC (27.62dB/0.8765) (f) ABME (31.61dB/0.8954)

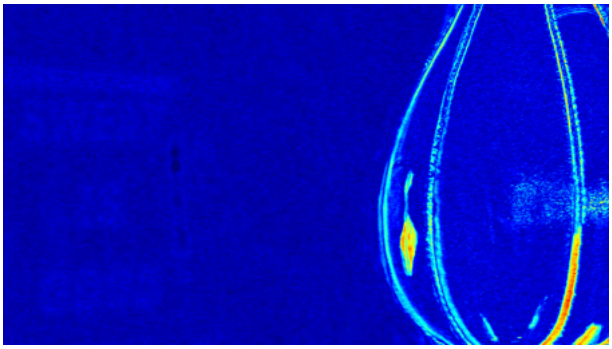
Figure S-14: Qualitative comparison of interpolated frames in the Xiph D2 dataset.



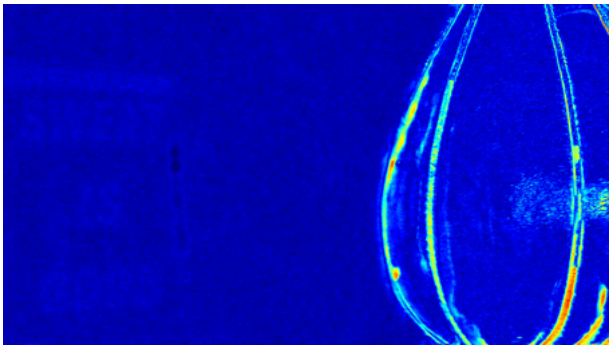
(b) DAIN (29.70dB/0.8747)



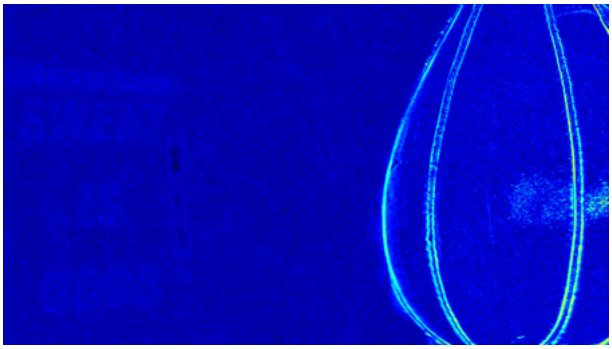
(c) CAIN (28.43dB/0.8816)



(d) AdaCoF (27.98dB/0.8734)

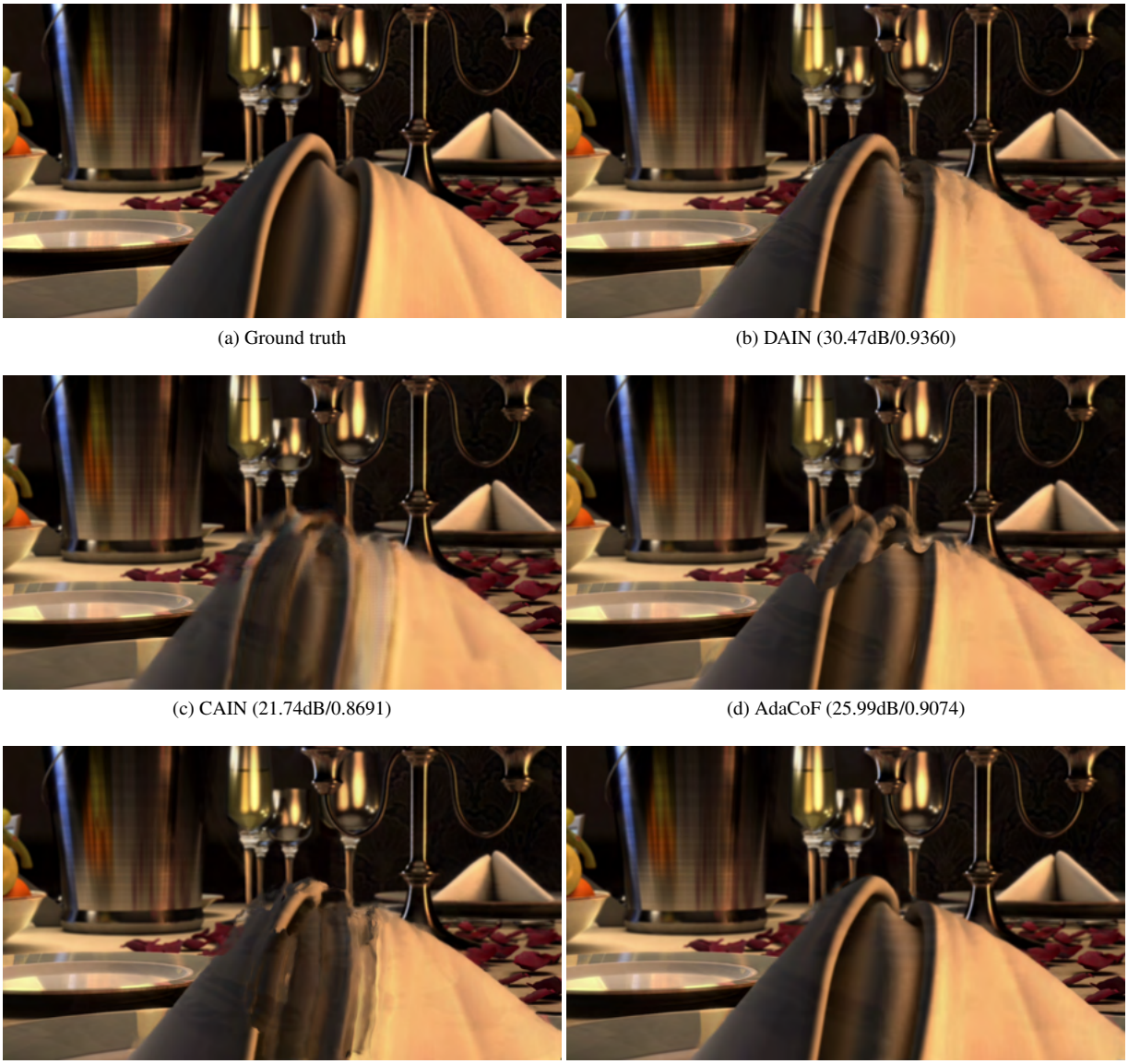


(e) BMBC (27.62dB/0.8765)



(f) ABME (31.61dB/0.8954)

Figure S-15: Error maps of the interpolated frames in Figure S-14.

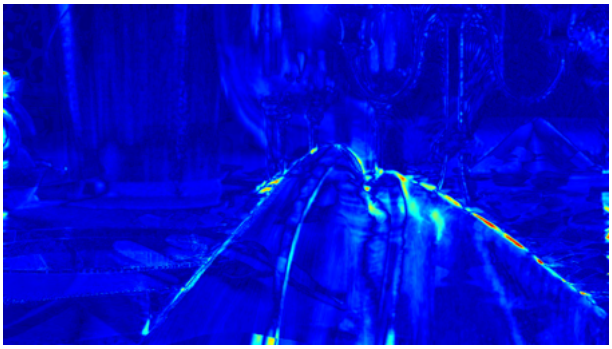


(a) Ground truth (b) DAIN (30.47dB/0.9360)

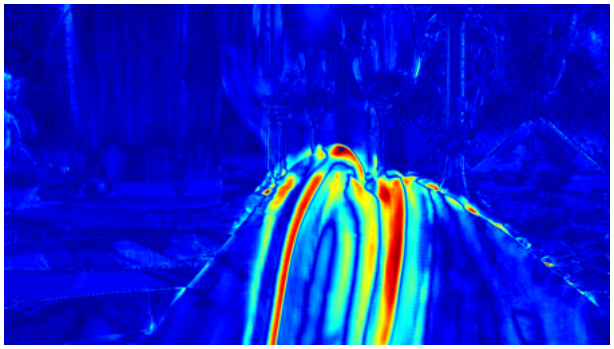
(c) CAIN (21.74dB/0.8691) (d) AdaCoF (25.99dB/0.9074)

(e) BMBC (23.17dB/0.8828) (f) ABME (33.91dB/0.9505)

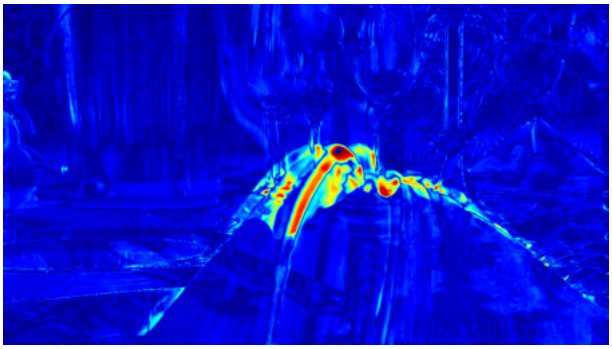
Figure S-16: Qualitative comparison of interpolated frames in the Xiph D3 dataset.



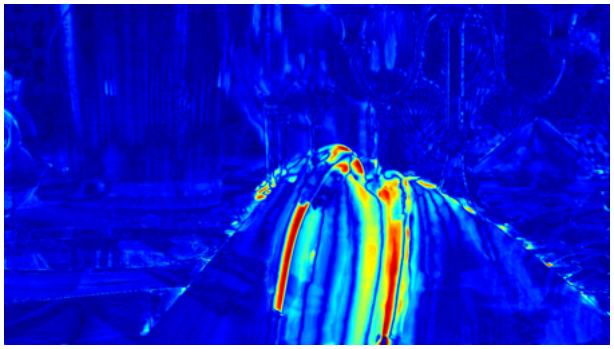
(b) DAIN (30.47dB/0.9360)



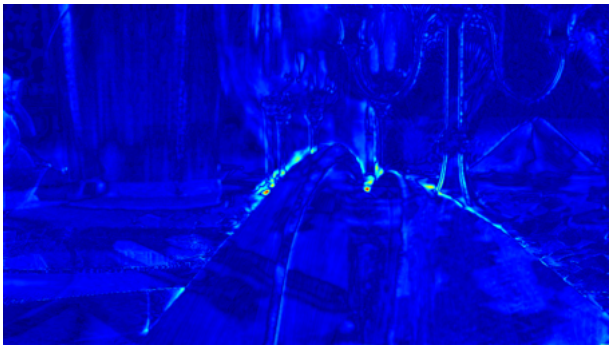
(c) CAIN (21.74dB/0.8691)



(d) AdaCoF (25.99dB/0.9074)



(e) BMBC (23.17dB/0.8828)



(f) ABME (33.91dB/0.9505)

Figure S-17: Error maps of the interpolated frames in Figure S-16.

References

[1] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *AAAI*, pages 10663–10671, Feb. 2020.

[2] Christopher Montgomery. Xiph. org video test media (derf's collection), the xiph open source community, 1994. *Online, https://media. xiph. org/video/derf*.

[3] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *Int. J. Comput. Vis.*, 127(8):1106–1125, Feb. 2019.