Video Infilling with Rich Motion Prior Supplementary Materials

Xinyu Hou[1]
xinyu.hou@ntu.edu.sg
Liming Jiang[1]
liming002@ntu.edu.sg
Rui Shao[2]
shaorui@hit.edu.cn
Chen Change Loy[1]
ccloy@ntu.edu.sg

[1] S-Lab
Nanyang Technological University
Singapore

[2] School of Computer Science and Technology
Harbin Institute of Technology (Shenzhen)
Shenzhen, China

1 Additions to Section 3.3 Bi-directional Fusion

Given the consistent bi-directional warping results $\hat{I}^c_{fwd}$ and $\hat{I}^c_{bwd} \in \mathbb{R}^{T\times H\times W\times 3}$, the raw warping results without consistency check $\hat{I}^r_{fwd}$ and and $\hat{I}^r_{bwd} \in \mathbb{R}^{T\times H\times W\times 3}$, and the current mask regions μ_{fwd} and $\mu_{bwd} \in \mathbb{R}^{T\times H\times W\times 1}$, pixel value at pixel position p is defined as:

$$\hat{I}_{t\in T}(p) = \begin{cases} \hat{I}^c_{bwd}(p) & \text{if } p \in \mu_{fwd} \text{ and } p \notin \mu_{bwd} \\ \hat{I}^c_{fwd}(p) & \text{if } p \notin \mu_{fwd} \text{ and } p \in \mu_{bwd} \\ \hat{I}^r_{fwd}(p) & \text{if } p \in \mu_{fwd} \text{ and } p \in \mu_{bwd} \text{ and } t \leq (T/2) \\ \hat{I}^r_{bwd}(p) & \text{if } p \in \mu_{fwd} \text{ and } p \in \mu_{bwd} \text{ and } t > (T/2) \\ t\hat{I}^c_{fwd}(p) + (1-t)\hat{I}^c_{bwd}(p) & \text{if } p \notin \mu_{fwd} \text{ and } p \notin \mu_{bwd} \end{cases} \tag{1}$$

Equ. 1 performs three operations: 1) if p is consistent in only one direction, it keeps its value in that direction; 2) if p is inconsistent in both directions, its value is approximated by the raw warping value from its nearest context frame; 3) if p is consistent in both directions, it takes the weighted sum of its consistent warping values from both directions. We use naïve fusion because our primary focus of this work is learning effective motion priors via masked modeling. The non-learnable operation can better illustrate the validity of our predicted motions.

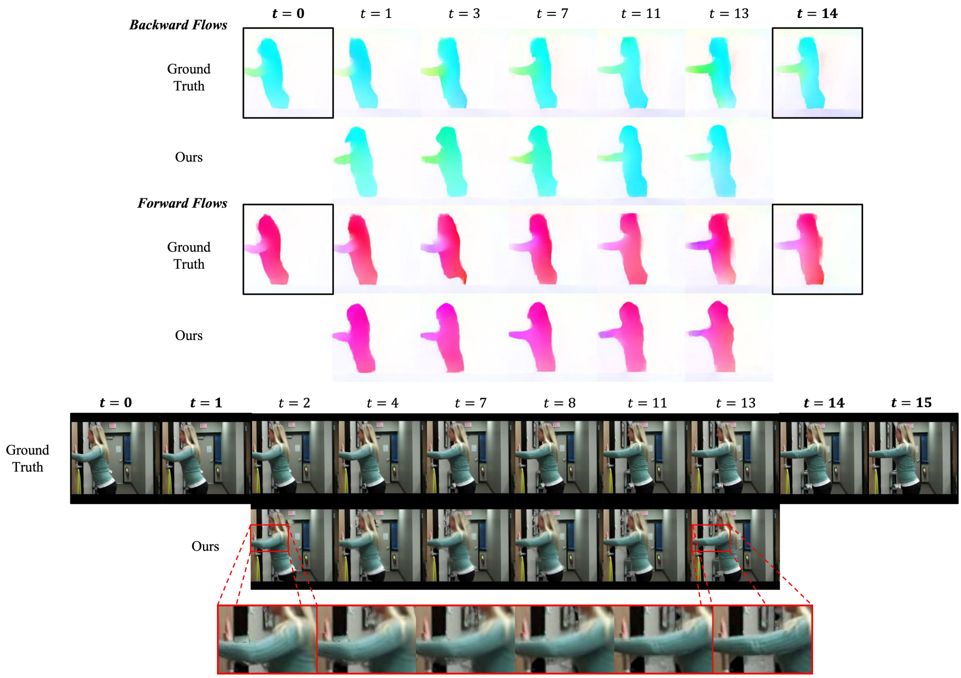

Figure 1: Additional qualitative evaluation of our method. Example 1. Zoom in for a better view.

2 Additions to Section 4.2: Qualitative Results

2.1 Additional Qualitative Results

More qualitative results of our method are shown. Same as Fig.2 in the main paper, all shown examples predict bi-directional optical flows by iterative decoding for 20 iterations. Pixel-level infilling results are obtained from a non-learnable naïve fusion. Context flows and frames are marked with black outlines. At time t, forward flow denotes $f_{t \to t+1}$ and backward flow denotes $f_{t+1 \to t}$.

3 Additions to Section 4.3: Comparison with Recent Approaches

3.1 Input Setting

Though briefly explained in the main paper, we would like to further clarify how we align the input settings of two previous approaches [1, 2] with ours. An illustration of the input setting alignment is shown in Fig.3. As mentioned in the main paper Sec.3, our method takes in context frames $I_0, I_1, I_{T-2}, I_{T-1}$, and synthesizes intermediate frames $\hat{I}_t$, $t \in [2, (T-3)]$ under the same frame rate. In our experiments, $T = 16$ is applied. Since QVI [2] employs the same input setting of 4 context frames (2 before the interpolation time step and 2 after the interpolation time step) and is able to predict several intermediate frames at the same time, we follow the same input setting as ours to conduct experiments on QVI. As for FILM [1], it takes two context frames and predicts the central frame only. To make sure the ground truths of the predicted frames are available and the interpolation motion range is comparable to our

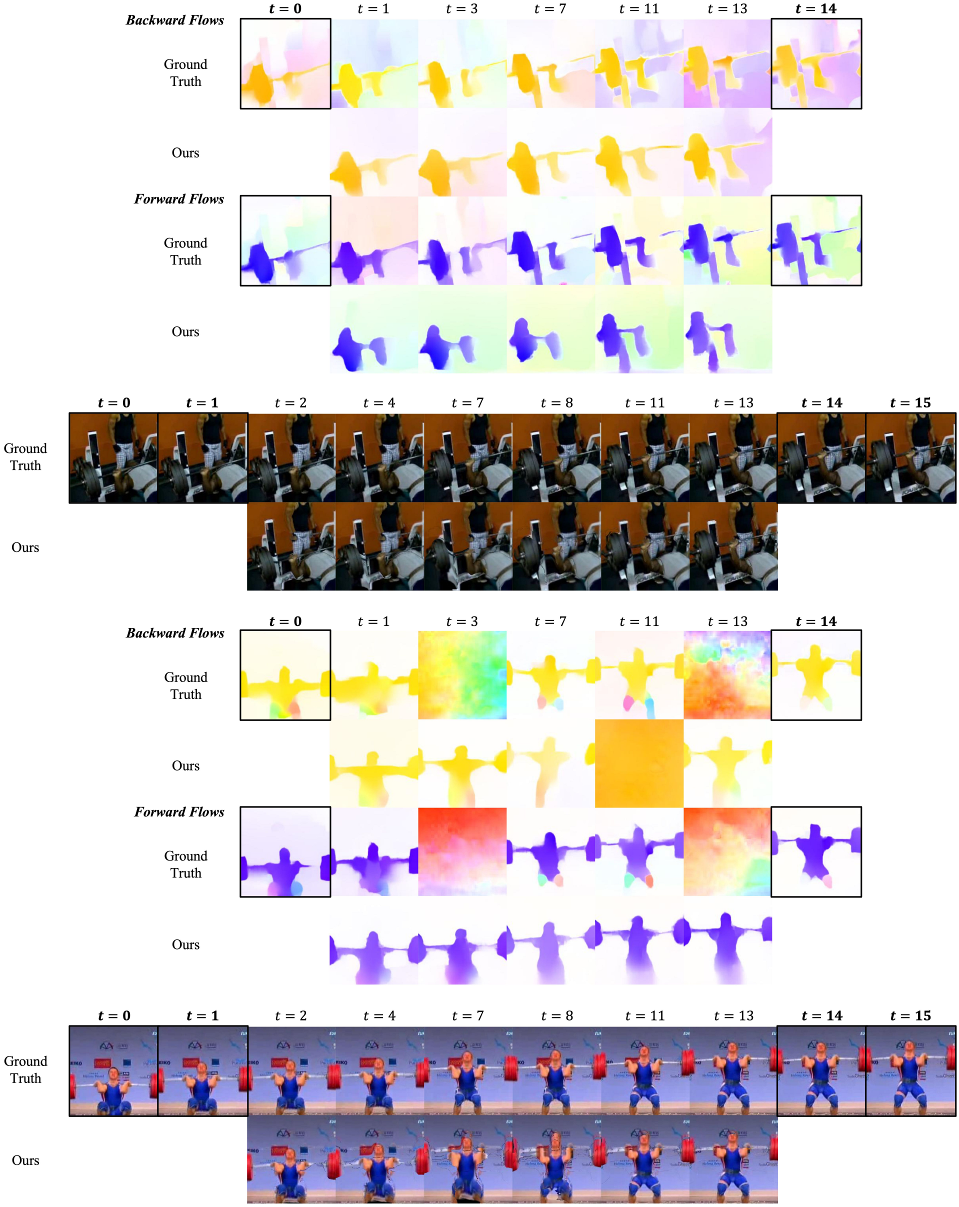

Figure 2: Additional qualitative evaluation of our method. Example 2 and 3. Zoom in for a better view.

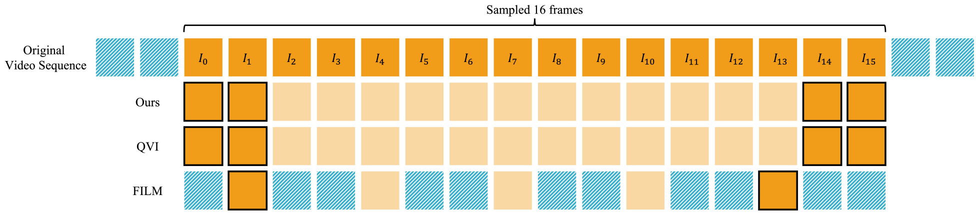

Figure 3: Illustration of input setting alignment. Blue-shaded squares indicate unused frames in the original video sequence; Orange squares are sampled frames, and black outlines indicate context frames; light orange squares are frames to be predicted.

setting, we choose to feed I_1 and I_{13} to FILM to predict $\hat{I}_7$. In addition, with the predicted $\hat{I}_7$, we can further run FILM on $(I_1, \hat{I}_7)$ and $(\hat{I}_7, I_{13})$ pairs to get $\hat{I}_4$ and $\hat{I}_{10}$, respectively. To compare the quantitative performances of the above methods, we report the PSNR and SSIM metrics averaged over predicted frames.

References

[1] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision (ECCV)*, 2022.

[2] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *NeurIPS*, 2019.