Supplementary Materials: Flow-Guided Video Inpainting with Scene Templates

Dong Lao Peihao Zhu Peter Wonka Ganesh Sundaramoorthi
King Abdullah University of Science and Technology (KAUST), Saudi Arabia
{dong.lao, peihao.zhu, peter.wonka, ganesh.sundaramoorthi}@kaust.edu.sa

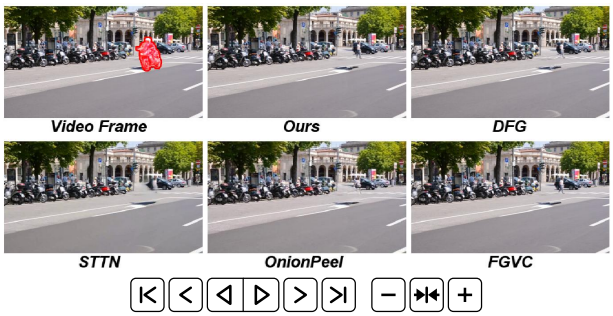

Figure 1: **Comparison with state-of-the-art.**

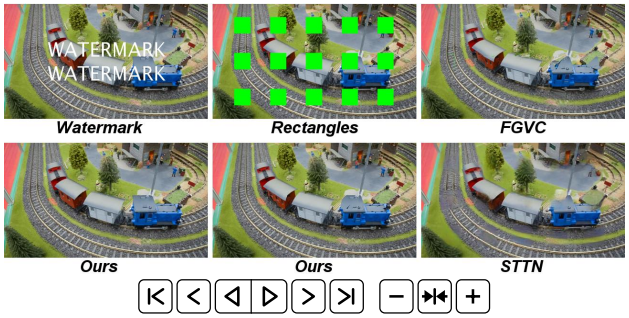

Figure 2: **Fixed region removal.**

Visual Results. We provide animated images for Figure 1 and Figure 7 in the main paper (best viewed in Adobe Acrobat Reader).

L_1 regularization corrects artifacts caused by occlusion. In principle, our method does not pose any restriction on the background motion, as in general the motion is handled by the w_i's that can be any mapping between the scene template to the images. However, in practice, limited by inaccurate optical flow at the occlusion boundary, the method could mistakenly aggregate occluded information, which leads to artifacts. Such artifacts at occlusion boundaries are also reported by [1, 2]. In Figure 3 we show such an example where inaccurate flow leads to artifacts in the L_2 template. However, our L_1 regularization successfully corrects the artifacts, resulting in plausible appearance. This shows the robustness to dynamic scenes of our L_1 regularization.

Figure 3: **Robustness to dynamic scenes.** Inaccurate flow leads to artifacts in the L_2 template. Our L_1 regularization successfully corrects the artifacts.

Occlusion reasoning for fixed region removal and dynamic scenes. Nevertheless, we provide a solution specifically modeling occlusion in dynamic scenes. In the main paper, we have shown the optimizer of the scene template is given by

$$f^*(p) = \frac{\sum_{i=1}^{T} I_i(w_i(p)) \mathbb{1}_i(w_i(p)) J_i(p)}{\sum_{i=1}^{T} \mathbb{1}_i(w_i(p)) J_i(p)}, \quad p \in \Omega \tag{1}$$

To further address occlusion, we can add an occlusion term to the optimizer:

$$f^*(p) = \frac{\sum_{i=1}^{T} I_i(w_i(p)) \mathbb{1}_i(w_i(p)) J_i(p) v_i(p)}{\sum_{i=1}^{T} \mathbb{1}_i(w_i(p)) J_i(p) v_i(p)}, \quad p \in \Omega \tag{2}$$

where $v_i(\cdot)$ is the indicator function of portion in the scene template that visible in I_i.

To estimate $v_i(\cdot)$, we can perform forward-backward optical flow consistency check between the scene template and I_i by

$$v_i(p) = \begin{cases} 0 & \text{if } |w_i^{-1} \circ w_i(p)|_2 > T \\ 1 & \text{else} \end{cases} \tag{3}$$

where T is a threshold that controls the sensitivity of occlusion estimation. In such a way, $v_i(\cdot)$ measures the visibility of content in the scene template. Figure 4 shows an example. In this example, multiple consecutive frames are aggregated into a template. Without occlusion reasoning, there are artifacts at the occlusion boundary. Occlusion reasoning successfully removes such artifacts. Note that the combination with L_1 provides even more rigid result.

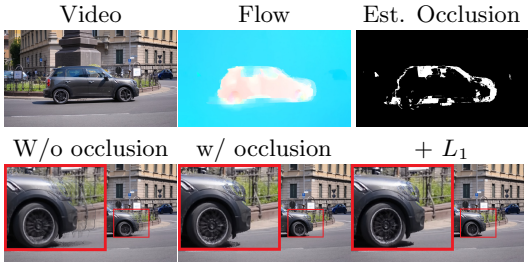

Figure 4: **Occlusion reasoning improves the model's ability to handle dynamics.** By further occlusion reasoning, our method successfully corrects artifacts at the occlusion boundary.

Specifically, when running fixed region removal experiments, enabling this occlusion reasoning improves the inpainting performance, since the masks are likely to pass an occlusion boundary. We tested fixed region removal following the same settings as the current state-of-the-art method [1] on all sequences on DAVIS 2016. As stated in the main paper, the results are on par with [1]. In Figure 5 we also showcase the scene template that is constructed during fixed region inpainting. The train in this scene is moving, and our model is able to obtain a reasonable template and perform inpainting.

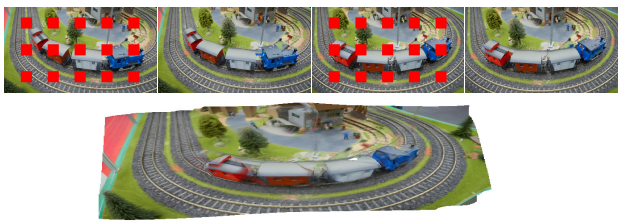

Figure 5: **Fixed region inpainting and corresponding scene template.** Our model obtains a reasonable template and performs inpainting.

Speed: Our approach (Algorithm 2 in the main paper) runs at 3 seconds per frame on DAVIS, which is comparable to state-of-the-art flow-guided inpainting methods [2, 1]. The bottleneck is optical flow, which we expect to improve.

Failure Cases. Typical failure cases are caused by illumination change and optical flow failure. Figure 6 shows an example where all existing methods including ours fail. The smoke from the drifting car changes the illumination of the background. Our scene template was constructed by pixel values from frames without the smoke. Therefore the model still maps clear instead of smokey scene parts to the masked region. Note that the change in illumination also leads to failure in optical flow estimation for both FlowNet2 and Sobolev Flow. In such cases, ALL methods fail as the flow cannot be accurately estimated, and rapid illumination change makes it infeasible to propagate image content across frames.

Figure 6: **A failure case.** Typical failure cases are caused by illumination change and optical flow failure.

References

[1] C. Gao, A. Saraf, J.-B. Huang, and J. Kopf. Flow-edge guided video completion. In *ECCV*, 2020. 1, 2

[2] R. Xu, X. Li, B. Zhou, and C. C. Loy. Deep flow-guided video inpainting. In *CVPR*, 2019. 1, 2