ST-MFNet: A Spatio-Temporal Multi-Flow Network for Frame Interpolation
Supplementary Material

Duolikun Danier Fan Zhang David Bull
University of Bristol
{duolikun.danier, fan.zhang, dave.bull}@bristol.ac.uk

The Supplementary Material is organized as follows. Section A presents the discriminator architecture used for training ST-MFNet. Section B provides additional ablation study results. The visualization of the multi-scale multi-flows is given in Section C. Section D presents the additional quantitative evaluation results in terms of LPIPS [28]. Results on multi-frame interpolation are provided in Section E. An additional experiment to validate the model design is described in Section F. The experimental configuration of the user study is described in Section G. Section H provides a link to a supplementary demo video. Finally, Section I summarizes the license information for all data and code assets used in this paper.

A. Discriminator for ST-MFNet

The architecture of the discriminator employed in this work is illustrated in Figure 1; this was originally designed to train ST-GAN [27] for texture synthesis. It contains a temporal and a spatial branch. The former takes the differences between the interpolated output \( I_{\text{out}}^t \) (where \( t = 1.5 \)) of ST-MFNet and its two adjacent original frames \( I_1, I_2 \) as input. The differences here represent the high-frequency temporal information within these three frames. The spatial branch in this network processes the ST-MFNet output \( I_{\text{out}}^t \) to generate spatial features. Finally, the temporal and spatial features generated in these two branches are concatenated before fed into the final fully connected layers.

B. Additional Ablation Study Results

In the main paper, we presented key ablation study results where the primary contributions in the proposed ST-MFNet are evaluated. Here the effectiveness of the up-sampling scale is further investigated, which has been employed during the multi-flow prediction in the MIFNet branch (see Section 3.1 of the main paper). In addition, we present the quantitative ablation study results for the ST-GAN in terms of a perceptually-oriented metric, the Learned Perceptual Image Patch Similarity (LPIPS) [28].

Up-sampling. To evaluate the contribution of the up-sampling scale during the multi-flow prediction, the version of ST-MFNet (Ours-w/o US) with only two multi-flow estimation heads (at \( l = 0, 1 \) scales) were implemented. It was also trained and evaluated using the same configurations described in the main paper. Its interpolation results are summarized in Table 1 alongside more comprehensive ablation study results for the other four variants of ST-MFNet (described in the main paper). It can be observed that Ours-w/o US was outperformed by the full version of ST-MFNet (Ours) on all test datasets. The performance difference can also be demonstrated through visual comparison as shown in Figure 2. All of these confirm the effectiveness of the up-sampling scale in multi-flow estimation.

ST-GAN. In the main paper, due to space limitations, we only evaluated the effectiveness of the adopted ST-GAN using visual examples. Here we additionally present the quantitative ablation study results for the adopted ST-GAN. For this purpose, we evaluate the same variants of ST-MFNet as described in Section 5.1 (the ST-GAN sub-section) of the main paper, that is, the distortion-oriented version (Our-\( \mathcal{L}_{\text{top}} \)), the version fine-tuned with ST-GAN (Our-\( \mathcal{L}_{\text{top}} \)), the version fine-tuned with FIGAN [12] and the version fine-tuned with TGAN [22]. Table 2 summarizes the performance of these variants on all four test sets in terms of LPIPS. It can be clearly observed from the table that the ST-GAN adopted in our work provides the best overall LPIPS performance, indicating its effectiveness for enhancing perceptual quality of the interpolated results.

C. Visualization of Motion Fields

To better understand the effectiveness of the multi-scale multi-flow estimation in the MIFNet branch, the predicted multi-flows are visualized here in the same manner as done in [12]. That is, the mean flow maps at scale \( l \), \( \bar{C}_{l}^{\text{I}_{t\rightarrow n}} \) (where \( n = 1, 2 \)), are obtained using Equations (1) and (2), and shown in Figure 3. Note that for the purpose of visualization, the flows at the down- and up-sampled scales are rescaled to the original resolution using the nearest neighbor interpolation.
Figure 1. Architecture of the discriminator used for training ST-MFNet.

Table 1. Comprehensive ablation study results on ST-MFNet.

<table>
<thead>
<tr>
<th></th>
<th>UCF101</th>
<th>DAVIS</th>
<th>SNU-FILM</th>
<th>VFITex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Medium</td>
<td>Hard</td>
<td>Extreme</td>
</tr>
<tr>
<td>Ours-w/o BLFNet</td>
<td>33.218</td>
<td>0.970</td>
<td>27.767/0.881</td>
<td>40.655/0.990/0.984</td>
</tr>
<tr>
<td>Ours-w/o MIFNet</td>
<td>33.202</td>
<td>0.969</td>
<td>27.886/0.889</td>
<td>40.331/0.991</td>
</tr>
<tr>
<td>Ours-w/o TENet</td>
<td>32.895</td>
<td>0.970</td>
<td>27.484/0.880</td>
<td>40.275/0.991</td>
</tr>
<tr>
<td>Ours-unet</td>
<td>33.378</td>
<td>0.970</td>
<td>28.096/0.892</td>
<td>40.616/0.991</td>
</tr>
<tr>
<td>Ours-w/o US</td>
<td>33.371</td>
<td>0.970</td>
<td>28.155/0.893</td>
<td>40.248/0.990</td>
</tr>
<tr>
<td>Ours</td>
<td>33.384</td>
<td>0.970</td>
<td>28.287/0.895</td>
<td>40.775/0.992</td>
</tr>
</tbody>
</table>

Table 2. Quantitative ablation study results for ST-GAN, in terms of LPIPS.

<table>
<thead>
<tr>
<th></th>
<th>UCF101</th>
<th>DAVIS</th>
<th>SNU-FILM</th>
<th>VFITex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Medium</td>
<td>Hard</td>
<td>Extreme</td>
</tr>
<tr>
<td>Ours-(L_{lap})</td>
<td>0.036</td>
<td>0.125</td>
<td>0.019</td>
<td>0.036</td>
</tr>
<tr>
<td>TGAN</td>
<td>0.036</td>
<td>0.119</td>
<td>0.020</td>
<td>0.035</td>
</tr>
</tbody>
</table>

Figure 2. Qualitative results interpolated by the ST-MFNet with the up-sampled scale removed (Ours-w/o US) and the full version of ST-MFNet (Ours-w/ US). Here “Overlay” means the overlaid adjacent frames.
filter.

\[
g(x, y, i) = (\alpha(x, y, i), \beta(x, y, i))
\]

\[
\bar{G}_{l \rightarrow n}(x, y) = \sum_{i=1}^{N} w(x, y, i) g(x, y, i)
\]

It can be observed from Figure 3 that compared to the mean flow map at the original scale \((l = 0)\), the flows estimated for the down-sampled scale \((l = 1)\) tend to depict the general motion coarsely in different regions. On the other hand, the flow maps at the up-sampled scale \((l = -1)\) reflect more detailed motion information.

D. Comprehensive Evaluation Results

In the main paper, we presented our quantitative evaluation results of the proposed ST-MFNet and 14 competing methods in terms of PSNR and SSIM. Here, we additionally evaluate these methods in terms of LPIPS. The full results on the test sets UCF101 [24], DAVIS [21] and VFI Tex are summarized in Table 3, and the results on SNU-FILM [6] are shown in Table 4.

E. Results of 4\times and 8\times Interpolation

The performance of the proposed ST-MFNet on multi-frame interpolation task is also evaluated, and compared to three best-performing benchmark algorithms: QVI [25], FLAVR [11] and Softsplat [18]. The algorithms were applied recursively to generate all the intermediate frames. The 11 test sequences at 240 FPS in the GoPro dataset [17] were used as the test set for 4\times and 8\times interpolation. Table 5 summarizes the results, where it can be seen that ST-MFNet shows the best overall performance.

F. Validation of Model Design

The proposed ST-MFNet combines multi-flow and single-flow based warping methods to enhance the interpolation quality of both complex and large motions. A natural question to ask is whether the performance of the model comes from the specific model design or simply from ensembling effect. To address this question, we create an ensemble model as a baseline, which simply combines AdaCoF [12] and Softsplat [18] through arithmetic averaging. This baseline model was trained under the same configurations as ST-MFNet and compared to the latter quantitatively. The results are summarized in Table 6, where it is noted that although ensembling of AdaCoF and Softsplat does provide some benefit, the gain is marginal. This implies that the main source of the performance gain in ST-MFNet is the model design.

G. User Study

The user study was conducted in a darkened, lab-based environment. The test sequences were played on a SONY PVM-X550 display, with screen size 124.2\times71.8cm. The display resolutions were configured to 1920\times1080 (spatial) and 60Hz (temporal), and the viewing distance was 2.15 meters (three times the screen height) [9]. The presentation of video sequences was controlled by a Windows PC running Matlab Psychtoolbox [3]. In each trial, a pair of videos to be compared were played twice, then the participant was asked to select the video with better perceived quality through an interface developed using the Psychtoolbox. This user study and the use of human data have undergone an internal ethics review and has been approved by the Institutional Review Board.

H. Video Demo

A video containing interpolation examples generated by ST-MFNet and more visual comparisons is available via this link: https://drive.google.com/file/d/1zpE3rCQNjI4e8ADNWKbJA5wTvPl1KZSj/view?usp=sharing.
Table 3. Quantitative comparison results for our model and 14 tested methods on UCF101, DAVIS and VFI Tex, in terms of PSNR, SSIM and LPIPS. OOM denotes cases where our GPU runs out of memory for the evaluation. For each column, the best result is colored in red and the second best is colored in blue. Underlined scores denote the performance of pre-trained models rather than our re-trained versions.

Table 4. Quantitative comparison results for our model and 14 tested methods on SNU-FILM dataset, in terms of PSNR, SSIM and LPIPS. For each column, the best result is colored in red and the second best is colored in blue. Underlined scores denote the performance of pre-trained models rather than our re-trained versions.

I. Attribution of Assets

The data and code assets employed in this work and their corresponding license information are summarized in Table 7 and 8 respectively.
Table 5. Quantitative comparison results for 4× and 8× interpolation on GoPro dataset in terms of PSNR, SSIM and LPIPS. For each column, the best result is colored in red and the second best is colored in blue.

<table>
<thead>
<tr>
<th>Method</th>
<th>GoPro-4×</th>
<th></th>
<th>GoPro-8×</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR (↑)</td>
<td>SSIM (↑)</td>
<td>LPIPS (↓)</td>
<td>PSNR (↑)</td>
</tr>
<tr>
<td>QVI [25]</td>
<td>29.324</td>
<td>0.927</td>
<td>0.049</td>
<td>29.280</td>
</tr>
<tr>
<td>FLA VR [11]</td>
<td>28.911</td>
<td>0.914</td>
<td>0.110</td>
<td>29.512</td>
</tr>
<tr>
<td>Softsplat [18]</td>
<td>28.858</td>
<td>0.908</td>
<td>0.072</td>
<td>29.663</td>
</tr>
<tr>
<td>ST-MFNet (Ours-(L_{lap}))</td>
<td>29.892</td>
<td>0.926</td>
<td>0.098</td>
<td>30.568</td>
</tr>
</tbody>
</table>

Table 6. Quantitative evaluation results of the proposed ST-MFNet and a simple baseline that combines AdaCoF and Softsplat. For each row, the best result is colored in red and the second best is colored in blue. Note Softsplat here is trained with Charbonnier loss so that AdaCoF, Softsplat and the baseline only differ in model design.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PSNR (↑)</th>
<th>SSIM (↑)</th>
<th>PSNR (↑)</th>
<th>SSIM (↑)</th>
<th>PSNR (↑)</th>
<th>SSIM (↑)</th>
<th>PSNR (↑)</th>
<th>SSIM (↑)</th>
<th>PSNR (↑)</th>
<th>SSIM (↑)</th>
<th>PSNR (↑)</th>
<th>SSIM (↑)</th>
<th>PSNR (↑)</th>
<th>SSIM (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF101</td>
<td></td>
<td>AdaCoF</td>
<td>Softsplat</td>
<td>AdaCoF+Softsplat</td>
<td>ST-MFNet (ours-(L_{lap}))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PSNR (↑)</td>
<td></td>
<td>SSIM (↑)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>32.488</td>
<td>0.968</td>
<td>32.683</td>
<td>0.969</td>
<td>32.729</td>
<td>0.969</td>
<td>33.384</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AdaCoF</td>
<td>Softsplat</td>
<td>AdaCoF+Softsplat</td>
<td>ST-MFNet (ours-(L_{lap}))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PSNR (↑)</td>
<td></td>
<td>SSIM (↑)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>26.445</td>
<td>0.854</td>
<td>27.359</td>
<td>0.878</td>
<td>27.361</td>
<td>0.878</td>
<td>28.287</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AdaCoF</td>
<td>Softsplat</td>
<td>AdaCoF+Softsplat</td>
<td>ST-MFNet (ours-(L_{lap}))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PSNR (↑)</td>
<td></td>
<td>SSIM (↑)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>35.269</td>
<td>0.977</td>
<td>35.833</td>
<td>0.979</td>
<td>35.841</td>
<td>0.979</td>
<td>37.111</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AdaCoF</td>
<td>Softsplat</td>
<td>AdaCoF+Softsplat</td>
<td>ST-MFNet (ours-(L_{lap}))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PSNR (↑)</td>
<td></td>
<td>SSIM (↑)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>29.723</td>
<td>0.928</td>
<td>30.412</td>
<td>0.936</td>
<td>30.449</td>
<td>0.937</td>
<td>31.698</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AdaCoF</td>
<td>Softsplat</td>
<td>AdaCoF+Softsplat</td>
<td>ST-MFNet (ours-(L_{lap}))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PSNR (↑)</td>
<td></td>
<td>SSIM (↑)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>24.656</td>
<td>0.851</td>
<td>25.242</td>
<td>0.862</td>
<td>25.258</td>
<td>0.864</td>
<td>25.810</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AdaCoF</td>
<td>Softsplat</td>
<td>AdaCoF+Softsplat</td>
<td>ST-MFNet (ours-(L_{lap}))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PSNR (↑)</td>
<td></td>
<td>SSIM (↑)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27.639</td>
<td>0.904</td>
<td>28.620</td>
<td>0.922</td>
<td>28.629</td>
<td>0.923</td>
<td>29.175</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(L_{lap}\): Laplacian loss.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dataset URL</th>
<th>License / Terms of Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVI-DVC [14]</td>
<td><a href="https://fan-aaron-zhang.github.io/BVI-DVC/">https://fan-aaron-zhang.github.io/BVI-DVC/</a></td>
<td>All sequences are allowed for academic research.</td>
</tr>
<tr>
<td>UCF101 [24]</td>
<td><a href="https://www.crcv.ucf.edu/research/data-sets/ucf101/">https://www.crcv.ucf.edu/research/data-sets/ucf101/</a></td>
<td>No explicit license terms, but compiled and made available for research use by the University of Central Florida.</td>
</tr>
<tr>
<td>Xiph [16]</td>
<td><a href="https://media.xiph.org/video/derf">https://media.xiph.org/video/derf</a></td>
<td>Sequences used are available for research use.</td>
</tr>
</tbody>
</table>

Table 7. License information for the datasets used in this work.

<table>
<thead>
<tr>
<th>Method</th>
<th>Source code URL</th>
<th>License / Teams of Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>XVFI [23]</td>
<td><a href="https://github.com/JihyongOh/XVFI">https://github.com/JihyongOh/XVFI</a></td>
<td>Research and education only.</td>
</tr>
</tbody>
</table>

Table 8. License information for the code assets used in this work.
References


