VideoFlow: Supplementary Material

Xiaoyu Shi\textsuperscript{1,2}, Zhaoyang Huang\textsuperscript{1,2,*}, Weikang Bian\textsuperscript{1}, Dasong Li\textsuperscript{1}, Manyuan Zhang\textsuperscript{1}, Ka Chun Cheung\textsuperscript{2}, Simon See\textsuperscript{2}, Hongwei Qin\textsuperscript{3}, Jifeng Dai\textsuperscript{4}, Hongsheng Li\textsuperscript{1,5,6,*}.

\textsuperscript{1}Multimedia Laboratory, The Chinese University of Hong Kong
\textsuperscript{2}NVIDIA AI Technology Center
\textsuperscript{3}SenseTime Research
\textsuperscript{4}Tsinghua University
\textsuperscript{5}Centre for Perceptual and Interactive Intelligence (CPII)
\textsuperscript{6}Shanghai AI Laboratory
\{xiaoyushi@link, drinkingcoder@link, hsli@ee\}.cuhk.edu.hk

1. More Implementation Details

Following SKFlow, we replace the vanilla convolutions with super-kernel blocks \cite{5, 3} for all encoders. As shown in Figure 1, SKBlock utilizes large-kernel depth-wise convolutions to enlarge the receptive fields while maintaining low computational cost.

As shown in Table 1, we follow SKFlow to implement these modules. We simply concatenate the correlation feature and bi-directional flows along the channel as inputs to CorrEncoder and FlowEncoder, respectively. The MotionEncoder of TROF is the same as that of SKFlow. For MOP, it additionally takes the 144-dim concatenation of motion features as input and additionally outputs a 48-dim updated motion feature.

2. More Details About Training Data

Note that the KITTI benchmark provides 20-frame sequences and each sequence has one GT flow map between frames 10-11. Therefore, we use the same number of GT for training as other two-frame methods. Since only the FlyingThings dataset provides bi-directional ground-truth optical flows, for other datasets, we randomly flip the input sequence and ground-truth optical flows with a probability of 0.3 to supervise the predicted backward flows. We follow FlowFormer for other settings of data augmentation.

3. Online Mode vs. Offline Mode

We conducted an experiment where only past frames are used (online setting). Our 5-frame VideoFlow achieves 4.08 Fl-all on the KITTI test set, outperforming 3-frame VideoFlow (4.44) and all previous two-frame methods. Specifically, we take KITTI images with indices 7, 8, 9, 10, 11 as inputs. The KITTI benchmark only evaluates the accuracy between images 10-11. The improvement reflects the superiority of VideoFlow in online setting.

4. Additional Qualitative Results

We provide additional visualisations in Figure 4, comparing our VideoFlow with previous best model FlowFormer++ \cite{4, 2, 1}.

5. Screenshots of Sintel Leaderboard

Figures 2 and 3 are the screenshots of evaluation results on the Sintel benchmarks. Our five-frame and three-frame models take the first and second places.

Table 1. Implementation details of encoders.

<table>
<thead>
<tr>
<th>Block</th>
<th>CorrEncoder</th>
<th>FlowEncoder</th>
<th>MotionEncoder</th>
<th>Updater</th>
<th>FlowHead</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>SKBlock</td>
<td>SKBlock</td>
<td>SKBlock</td>
<td>SKBlock</td>
<td>SKBlock</td>
</tr>
<tr>
<td>Input Dim.</td>
<td>648</td>
<td>4</td>
<td>256</td>
<td>512</td>
<td>128</td>
</tr>
<tr>
<td>Output Dim.</td>
<td>192</td>
<td>64</td>
<td>124</td>
<td>128</td>
<td>4</td>
</tr>
<tr>
<td>Hidden Dim.</td>
<td>256</td>
<td>96</td>
<td>256</td>
<td>512</td>
<td>128</td>
</tr>
</tbody>
</table>

References


Figure 4. VideoFlow can better distinguish foreground and background objects and is more robust to noise such as light reflection.