Supplementary Material – A Simple Baseline for Video Restoration with Grouped Spatial-temporal Shift

Dasong Li1  Xiaoyu Shi1  Yi Zhang1  Ka Chun Cheung2  Simon See2
Xiaogang Wang1,4  Hongwei Qin3  Hongsheng Li1,4

1CUHK MMLab  2NVIDIA AI Technology Center  3SenseTime Research  4CPII under InnoHK

1. Optical Flow Analysis in Video Restoration

Optical flow, the core component to model the motion information, has been widely used in video super-resolution [2, 3], video deblurring [15, 18] and video denoising [21]. However, Zhu et al. [5] demonstrate that optical flow cannot estimate the alignment information well because of the significant influence of the motion blur. [6] also show that the optical flow is not accurate in noisy images.

We provide a quantitative analysis of optical flow in three video restoration tasks, including video super-resolution, video deblurring and video denoising. We select BasicVSR++ [3] (denoted as “BasicVSR++”) as the baseline model. To evaluate the importance of optical flow module, we remove the optical flow estimation from BasicVSR++ (denoted as “BasicVSR++ w/o flow”). We increase the number of residual blocks [9] and offsets computing layers in DCN to maintain the same running time as BasicVSR++. Then two models are trained for 200,000 iterations on video super-resolution (REDS4 dataset [13]), video deblurring (GoPro dataset [14]) and video denoising (Set8 dataset [20]), respectively. For a fair comparison of three tasks, we do not take the generalized version [4] of BasicVSR++ and the models for video deblurring and video denoising have the same parameters as BasicVSR++ [3] for video super-resolution. It is observed in Table 1 that the optical flow module makes different influences on different tasks. The optical flow can boost the performance of super-resolution by 0.56 dB. However, optical flow cannot improve video deblurring and denoising greatly because optical flow is not that accurate in blurry and noisy images as shown in [5,6,18,22].

We also provide a visualization of optical flow in Figure 1. Given degraded input frames $I_{t-1}, I_t$ and ground truth frames $GT_{t-1}, GT_t$, we utilize a pre-trained optical flow model [16] to estimate the optical flow of degraded pairs $I_t \rightarrow I_{t-1}$ and ground truth pairs $GT_t \rightarrow GT_{t-1}$.

<table>
<thead>
<tr>
<th>Method</th>
<th>SR REDS4</th>
<th>Deblurring GoPro $\sigma=10$</th>
<th>Denoising $\sigma=30$</th>
<th>Denoising $\sigma=50$</th>
<th>Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>BasicVSR++</td>
<td>32.01</td>
<td>33.22</td>
<td>36.19</td>
<td>31.75</td>
<td>7.3M</td>
</tr>
<tr>
<td>BasicVSR++ w/o flow</td>
<td>31.45</td>
<td>33.25</td>
<td>36.10</td>
<td>31.62</td>
<td>6.6M</td>
</tr>
</tbody>
</table>

Table 1. Analysis of optical flow on different tasks. Optical flow can improve video super-resolution greatly (+0.56 dB PSNR), but not in video deblurring and denoising.

We also utilize the optical flow network trained in the generalized version [4] of BasicVSR++ to visualize the task-oriented flow. It is shown in Figure 1 that the optical flow estimation is not accurate in noisy frames or blurry frames. Even with training on GoPro dataset [14] or DAVIS dataset [11], the task-oriented flow could not produce more accurate optical flow.

The different influences of optical flow estimation illustrate that the optical flow could help improve video super-resolution but make small contribution to video deblurring and denoising. Since it is difficult for optical flow to model motion information directly in video deblurring and video denoising, we design grouped spatial-temporal shift to achieve large receptive fields for implicit temporal correspondence modeling when optical flow is inaccurate.

The network is not designed for video super-resolution, which optical flow estimation could greatly help. Our network does not utilize optical flow and may not perform well on video super-resolution.

2. Qualitative Visualization

Video results and analysis. We provide four videos (Deblurring1.mp4, Deblurring2.mp4, Denoising1.mp4, Denoising2.mp4) in the project pages. Deblurring1.mp4 and Denoising1.mp4 provide the full-frame visualization of our restored video. It is shown that ours videos do not produce flicker cases and are temporal consist and stable. Deblurring2.mp4 and Denoising2.mp4 are provided to compare VRT [12] and our method clearly. It is shown that our
Figure 1. Optical flow visualization on GoPro testset [14] and Set8 testset [20]. The optical flow estimation of blurry images and noisy images is not accurate due to the negative influence of blur and noise. Task-oriented flow is usually smooth, but it is still inaccurate.

3. Further Analysis of Grouped Spatial Shift

Apart from the local attribute map (LAM) [8] visualization, we provide further analysis of grouped spatial shift. We perform LAM to obtain the contribution weights of four shifted feature groups of $o_{i+1}$ in helping restoring the local patch of $O_i$. According to the contribution weights, we sort $M$ shift vectors and divide them into different contribution classes. To find the connections between important shifted features and temporal motion information, we select optical flow to evaluate their shift vectors. $M$ shift vectors are sorted according to their contribution weights. We average top-4 important shift vectors obtain a pseudo optical flow $w_{i-1 \rightarrow i}$ for the local grids. We also calculate the pseudo optical flows of top $5 \sim 8$, $9 \sim 12$ and top $13 \sim 16$ important vectors. We utilize a pre-trained spynet [16] to estimate the optical flow $W_{i-1 \rightarrow i}$ from ground truth clean frames $H_{i-1}$ and $H_i$. The optical flow $W_{i \rightarrow i+1}$ is averaged in the every local grid. We calculate the correlations between optical flow $W_{i-1 \rightarrow i}$ and the pseudo optical flow $w_{i-1 \rightarrow i}$ along x-axis and y-axis, separately. It is shown in Table 2 that shifted feature groups usually make more contribution when the shift direction is similar to the optical flow $W_{i-1 \rightarrow i}$.

<table>
<thead>
<tr>
<th>Contribution</th>
<th>x-axis</th>
<th>y-axis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1 ~ 4</td>
<td>71.5 %</td>
<td>66.3 %</td>
</tr>
<tr>
<td>Top 5 ~ 8</td>
<td>32.5 %</td>
<td>34.7 %</td>
</tr>
<tr>
<td>Top 9 ~ 12</td>
<td>14.4 %</td>
<td>13.8 %</td>
</tr>
<tr>
<td>Top 13 ~ 16</td>
<td>3.7 %</td>
<td>3.5 %</td>
</tr>
</tbody>
</table>
Table 3. Deblurring performance of different motion magnitudes.

<table>
<thead>
<tr>
<th>Method</th>
<th>Largest 10 %</th>
<th>Smallest 10 %</th>
<th>Other 80 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>VRT</td>
<td>32.45</td>
<td>35.98</td>
<td>34.96</td>
</tr>
<tr>
<td>Ours</td>
<td>33.94 (+0.59)</td>
<td>36.25 (+0.25)</td>
<td>35.81 (+0.86)</td>
</tr>
</tbody>
</table>

4. Motion magnitudes

We categorize each frame of GoPro dataset according to motion magnitudes. For each blurry frame $I_i$ and its corresponding ground truth $O_i$, we utilize a pre-trained SPyNet to obtain optical flows between $O_i$ and two adjacent frames $O_{i-1}, O_{i+1}$. We obtain motion magnitudes by averaging the flows. The results in Table 3 show that our base model achieves 33.94dB in 10% largest magnitudes, which surpasses VRT (32.45dB) by +1.49dB. The gain of 10% smallest is 0.25dB.

5. Network Architecture

In our three-stage design, we take a three-scale U-Net [17] as our backbone. For each U-Net, we adopt the U-Net-like structure of MPRNet [23] to encode effective features. Average pooling and 2D bilinear upsampling is applied to obtain multi-scale features. Each feature in skip connections are processed by a Channel Attention Block (CAB) [24], which is the residual blocks equipped with a channel attention layer. The channel attention layer is first introduced in squeeze-and-excitation networks [10], and explored in low-level visions [23, 24]. In frame-wise feature extraction and final restoration, we take the Channel Attention Block (CAB) to extract frame-wise features. In multi-frame fusion, we utilize the proposed GSTS blocks to achieve multi-frame feature aggregation and communication. A GSTS block contains a grouped spatial-temporal shift operation and a lightweight fusion layer. The fusion layer, consisting of two lightweight convolution blocks (denoted as “FusionConv”), fuses the spatial-temporal shifted features effectively. Our FusionConv block takes the framework of Super Kernels (SKFlow) [19], which utilizes a small kernel convolution and a large kernel convolution as spatial filtering. The FusionConv block contains three point-wise convolution enable communication across channels and two depth-wise convolution for effective feature fusion. We utilize Layernorm [1] and channel attention [7] to improve the network capacity. Learning from NAFNet [7], we replace all GELU layers in SKFlow by gated layers to improve the performance further.

For our small model (“Ours-s”), we stack 3 slim U-Nets with 14 channels for frame-wise processing (Stage-1 and Stage-3) and the channel number of multi-frame fusion is set to be 64. For our base model “Ours” and enhanced version “Ours+”, we stack 5 slim U-Nets with 24 channels for frame-wise processing (Stage-1 and Stage-3) and the channel number of multi-frame fusion is set to be 80.

References


3


[22] Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, and Jingyu Yang. Supervised raw video denoising with a benchmark dataset on dynamic scenes. In IEEE Conference on Computer Vision and Pattern Recognition, 2020. 1
