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

Dasong Li[1] Xiaoyu Shi[1] Yi Zhang[1] Ka Chun Cheung[2] Simon See[2]
Xiaogang Wang[1,4] Hongwei Qin[3] Hongsheng Li[1,4]

[1]CUHK MMLab [2]NVIDIA AI Technology Center [3]SenseTime Research [4]CPII 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 verison [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_{i-1}, I_i and ground truth frames GT_{i-1}, GT_i, we utilize a pre-trained optical flow model [16] to estimate the optical flow of degraded pairs $I_i \rightarrow I_{i-1}$ and ground truth pairs $GT_i \rightarrow GT_{i-1}$.

Method	SR	Deblurring	Denoising			Params
	REDS4	GoPro	σ=10	σ=30	σ=50	
BasicVSR++	**32.01**	33.22	36.19	31.75	29.56	7.3M
BasicVSR++ w/o flow	31.45	33.25	36.10	31.62	29.49	6.6M

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

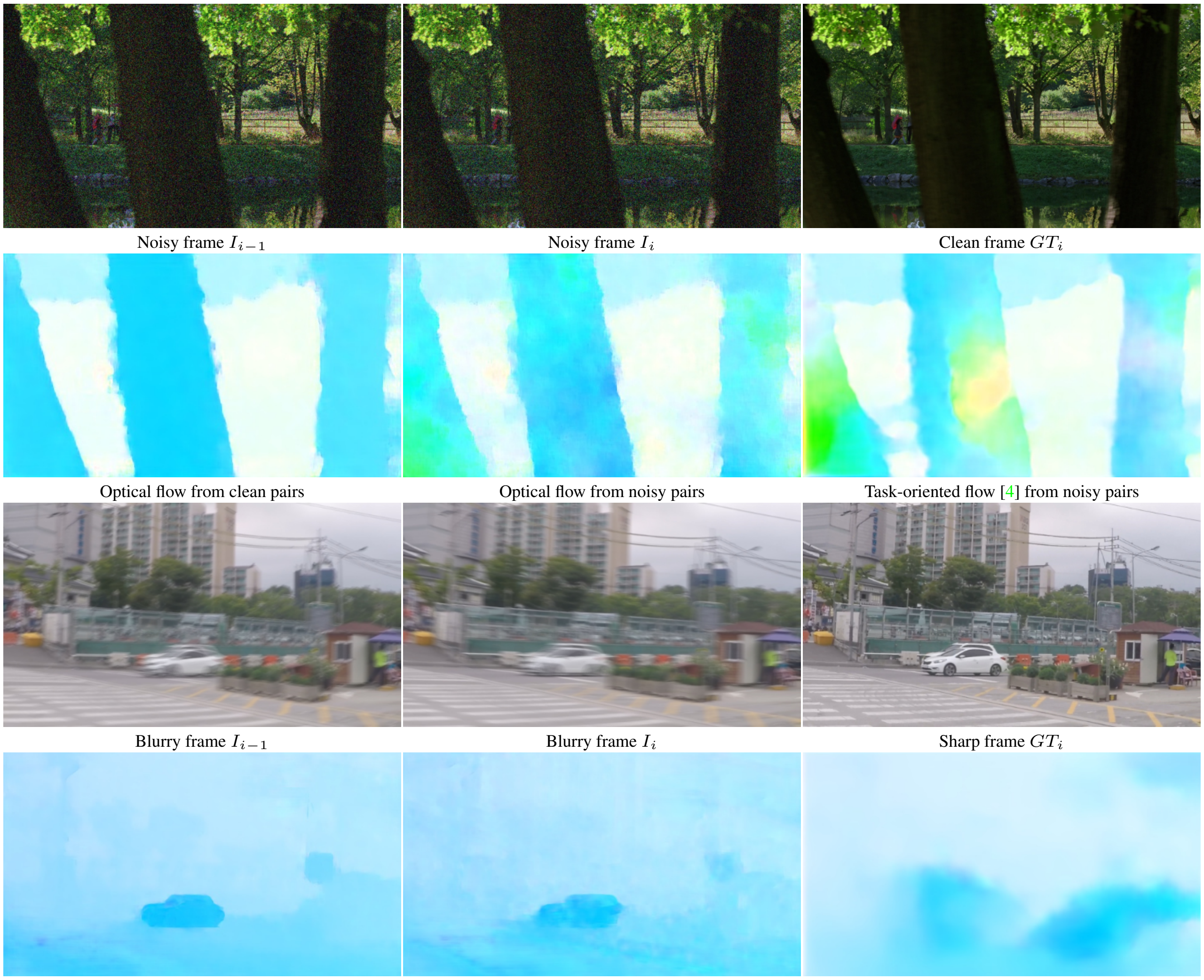

Figure 1. Optical flow visualization on GoPro testset [14] and Set8 testset [20]. The optial flow estimation of blurry images and noisy images is not accurate due to the negative influence of blur and noise. Task-oriented flow is usually smooth, but it is still inaccurate.

Contribution	x-axis	y-axis
Top 1 $\sim$ 4	71.5 %	66.3 %
Top 5 $\sim$ 8	32.5 %	34.7 %
Top 9 $\sim$ 12	14.4 %	13.8 %
Top 13 $\sim$ 16	3.7 %	3.5 %

Table 2. The correlations between the optical flow $W_{i-1\to i}$ of ground truth frames and the pseudo optical flow $w_{i-1\to i}$ produced from shifted features of different contribution weights.

method can restore more textures and details than VRT in both video deblurring and video denoising.

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 $w_{i-1\to i}$ 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 $W_{i-1\to i}$ from ground truth clean frames H_{i-1} and H_i. The optical flow $W_{i\to i+1}$ is averaged in the every local grid. We calculate the correlations between optical flow $W_{i-1\to i}$ and the pseudo optical flow $w_{i-1\to i}$ 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 $W_{i-1\to i}$.

Method	Largest 10 %	Smallest 10 %	Other 80 %
VRT	32.45	35.98	34.96
Ours	33.94 (**+1.49**)	36.23 (+0.25)	35.61 (+0.65)

Table 3. Deblurring performance of different motion magnitudes.

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

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 3

[2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 1

[3] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[4] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. On the generalization of BasicVSR++ to video deblurring and denoising. *arXiv preprint arXiv:2204.05308*, 2022. 1, 2

[5] Zhu Chao, Dong Hang, Pan Jinshan, Liang Boyang, Huang Yuhao, Fu Lean, and Wang Fei. Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In *AAAI*, 2022. 1

[6] Chen Chen, Qifeng Chen, Minh N. Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022. 3

[8] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 3

[11] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*, pages 123–141. Springer, 2018. 1

[12] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 1

[13] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1996–2005, 2019. 1

[14] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

[15] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[16] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2720–2729, 2017. 1, 2

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015. 3

[18] Hyeongseok Son, Junyong Lee, Jonghyeop Lee, Sunghyun Cho, and Seungyong Lee. Recurrent video deblurring with blur-invariant motion estimation and pixel volumes. *ACM Transactions on Graphics (TOG)*, 40(5), 2021. 1

[19] Shangkun Sun, Yuanqi Chen, Yu Zhu, Guodong Guo, and Ge Li. SKFlow: Learning optical flow with super kernels. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 3

[20] Matias Tassano, Julie Delon, and Thomas Veit. dvdnet: a fast network for deep video denoising. In *2019 IEEE International Conference on Image Processing*, Taipei, Taiwan, Sept. 2019. 1, 2

[21] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8):1106–1125, 2019. 1

[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

[23] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 3

[24] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 3