VideoGigaGAN: Towards Detail-rich Video Super-Resolution
Supplementary Material

This supplementary document includes additional quantitative results with SSIM scores, our network architecture, and training and evaluation details.

A. Additional quantitative comparison

We include SSIM scores in Table 1. Similar to our conclusion in the main paper, our model achieves the lowest LPIPS error, showing a detail-rich result.

	BI degradation			BD degradation		
	REDS4 [12]	Vimeo-90K-T [16]	Vid4 [9]	UMD10 [17]	Vimeo-90K-T	Vid4
TOFlow [16]	-/27.98/0.7990	-/33.08/0.9054	-/25.89/0.7651	-/36.26/0.9438	-/34.62/0.9212	-
RBPN [4]	-/30.09/0.8590	-/37.07/0.9435	-/27.12/0.8180	-/38.66/0.9596	-/37.20/0.9458	-
PFNL [17]	-/29.63/0.8502	-/36.14/0.9363	-/26.73/0.8029	-/38.74/0.9627	-	-/27.16/0.8355
EDVR [15]	0.2097/31.05/0.8793	-/37.61/0.9489	-/27.35/0.8264	-/39.89/0.9686	-/37.81/0.9523	-/27.85/0.8503
MuCAN [7]	0.2162/30.88/0.8750	0.1523/37.32/0.9465	-	-	-	-
BasicVSR [1]	0.2023/31.42/0.8909	0.1616/37.18/0.9450	0.2812/27.24/0.8251	0.1148/39.96/0.9694	0.1551/37.53/0.9498	0.2555/27.96/0.8553
IconVSR [1]	0.1939/31.67/0.8948	0.1587/37.47/0.9476	0.2739/27.39/0.8279	0.1152/40.03/0.9694	0.1531/37.84/0.9524	0.2462/28.04/0.8570
TTVSR [10]	0.1836/32.12/0.9021	-	-	0.1112/40.41/0.9712	0.1507/37.92/0.9526	0.2381/28.40/0.8643
BasicVSR++ [2]	0.1786/32.39/0.9069	0.1506/37.79/0.9500	0.2627/27.79/0.8400	0.1131/40.72/0.9722	0.1440/38.21/ 0.9550	0.2390/ 29.04/0.8753
RVRT [8]	0.1727/**32.74**/**0.9113**	0.1502/**38.15**/ **0.9527**	0.2500/**27.99**/**0.8464**	0.1100/**40.90**/**0.9729**	0.1465/**38.59**/ **0.9576**	0.2219/**29.54**/**0.8811**
Ours	**0.1582**/30.46/0.8718	**0.1120**/35.97/0.9238	**0.1925**/26.78/0.8029	**0.1060**/36.57/0.9521	**0.1129**/35.30/0.9317	**0.1832**/27.04/0.8365

Table 1. **Quantitative comparisons of VideoGigaGAN and previous VSR approaches.** We report LPIPS↓/PSNR↑/SSIM↑. Similar to our conclusion in the main paper, our model achieves the lowest LPIPS error, showing a detail-rich result.

B. Network architecture

B.1. GigaGAN upsampler

We show the configurationss of our GigaGAN upsampler in Table 2. For the low-pass filters, we use a kernel of $\frac{1}{16}[1, 4, 6, 4, 1]$ before the downsampling.

B.2. Flow-guided feature propagation module

We follow the architecture in BasicVSR++ [2]. We use SPyNet [13] as our flow estimator to reduce memory cost. For the feature extraction, we use 5 residual blocks. The number of residual blocks for propagation is set to 7. The kernel size of the deformable convolutional netowrk (DCN) is 3. We encourage readers to refer to BasicVSR++ [2] for more details.

C. Training and evaluation details

Datasets. Following previous works [1, 3], we use REDS [12] and Vimeo-90K [16] for training purpose. For REDS, we use clips 000, 011, 015, 020 of the training set for testing, and clips 000, 001, 006, 017 are used for validation, the rest of the clips are used for training. The ground truth has a resolution of 1280×720. For Vimeo-90K, in addition to its official test set Vimeo-90-K, we use UDM10 [17] and Vid4 [9] for testing purpose. The ground truth has a resolution of 448×256.

Degradation. We use MMagic's [11] script for degradations - Bicubic (BI) and Blur Downsampling (BD). For BD, the ground truth is blurred by a Gaussian filter with $\sigma = 1.6$, followed by a $4\times$ subsampling.

Table 2. GigaGAN model configurations

$\mathbf{z}$ dimension	512
$\mathbf{w}$ dimension	512
Mapping network layers	4
Activation	LeakyReLU
$\mathcal{G}$ channel base	32768
$\mathcal{G}$ channel max	512
$\mathcal{G}$ # of filters N for adaptive kernel selection	[1, 1, 1, 1, 1, 2, 4, 8, 16, 16, 16, 16]
$\mathcal{G}$ spatial self-attention resolutions	[8, 16]
$\mathcal{G}$ temporal attention resolutions	[8, 16, 32, 64]
$\mathcal{G}$ attention depth	[2, 2, 2, 1]
$\mathcal{G}$ temporal attention window size	1
$\mathcal{G}$ temporal convolution kernel size	3
$\mathcal{G}$ # synthesis block per resolution	[4, 4, 4, 4, 4, 4, 3]
$\mathcal{G}$ # downsampling blocks	3
$\mathcal{D}$ channel base	32768
$\mathcal{D}$ channel max	512
$\mathcal{D}$ attention depth	[2, 2, 1]
$\mathcal{D}$ attention resolutions	[8, 16]
$\mathcal{G}$ model size	369M
$\mathcal{D}$ model size	179M

Training settings. We use Adam optimizer [5] for training with a fixed learning rate of 5×10^{-5}. During training, we randomly crop a 64×64 patch from each LR input frames at the same location. We use 10 frames of each video and a batch size of 32 for training. The batch is distributed into 32 NVIDIA A100 GPUs. The total number of training iterations for each model is $100,000$.

Test settings. During the testing, we use the full-frame of the videos. Particularly, for Vimeo-90K-T, we follow its tradition and only evaluate PSNR, SSIM and LPIPS [18] on the center frame.

Metrics. We consider two aspects in our evaluation: per-frame quality and temporal consistency.

For **per-frame quality**, we use **PSNR, SSIM, and LPIPS** [18]. Except for REDS4, we evaluate PSNR and SSIM on y-channel following previous works [1, 2, 10].

For **temporal consistency**, we use warping error E_{warp} [6] and proposed referenced warping error $E_{\text{warp}}^{\text{ref}}$. Please refer to our main paper for the definition of E_{warp}^{ref}. We use RAFT [14] as our flow estimator when computing temporal consistency.

D. More visual results

We encourage readers to refer to our project website more visual results.

References

[1] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2021. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. In *CVPR*. 1, 2

[2] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. 2022. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*. 1, 2

[3] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. 2022. Investigating Tradeoffs in Real-World Video Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5962–5971. 1

[4] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. 2019. Recurrent Back-Projection Network for Video Super-Resolution. In *CVPR*. 1

[5] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 2

[6] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. 2018. Learning blind video temporal consistency. In *ECCV*. 170–185. 2

[7] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. 2020. Mucan: Multi-correspondence aggregation network for video super-resolution. In *ECCV*. Springer, 335–351. 1

[8] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. 2022. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems* 35 (2022), 378–393. 1

[9] Ce Liu and Deqing Sun. 2013. On Bayesian adaptive video super resolution. *TPAMI* 36, 2 (2013), 346–360. 1

[10] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. 2022. Learning Trajectory-Aware Transformer for Video Super-Resolution. In *CVPR*. 1, 2

[11] MMagic Contributors. 2023. MMagic: OpenMMLab Multimodal Advanced, Generative, and Intelligent Creation Toolbox. `https://github.com/open-mmlab/mmagic`. 1

[12] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. 2019. NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study. In *CVPRW*. 1

[13] Anurag Ranjan and Michael J Black. 2017. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4161–4170. 1

[14] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 402–419. 2

[15] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*. 0–0. 1

[16] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. 2019. Video Enhancement with Task-Oriented Flow. *IJCV* 127, 8 (2019), 1106–1125. 1

[17] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. 2019. Progressive Fusion Video Super-Resolution Network via Exploiting Non-Local Spatio-Temporal Correlations. In *IEEE International Conference on Computer Vision (ICCV)*. 3106–3115. 1

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595. 2