Is There Tradeoff between Spatial and Temporal in Video Super-Resolution?

Haochen Zhang, Dong Liu, Zhiwei Xiong
CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System,
University of Science and Technology of China, Hefei 230027, China
zhc12345@mail.ustc.edu.cn, {dongeliu, zwxiong}@ustc.edu.cn

1. Motivation

Recent advances of deep learning lead to great success of image and video super-resolution (SR) methods that are based on convolutional neural networks (CNN). For video SR, advanced algorithms have been proposed to exploit the temporal correlation between low-resolution (LR) video frames, and/or to super-resolve a frame with multiple LR frames [3, 8, 9, 11]. These methods pursue higher quality of super-resolved frames, where the quality is usually measured frame by frame in *e.g*. PSNR. However, as mentioned in [6], frame-wise quality may not reveal the consistency between frames. If an algorithm is applied to each frame independently (which is the case of most previous methods), the algorithm may cause temporal inconsistency, which can be observed as flickering. It is a natural requirement to improve both frame-wise fidelity and between-frame consistency, which are termed spatial quality and temporal quality, respectively. Then we may ask, is a method optimized for spatial quality also optimized for temporal quality? Can we optimize the two quality metrics jointly? In short, we want to understand the relationship between spatial quality and temporal quality, in the context of video SR.

2. Experiments and Analyses

Dataset. We use the HMDB51 dataset, which is a collection of real-world videos and was widely used for action recognition research [5]. The dataset includes 6,766 video clips that belong to 51 action categories. The dataset provides three training/testing splits. We use split1 as a representative. We down-sample the video clips by a factor of 4 using bicubic interpolation, and then super-resolve the video clips with different methods.

Compared methods. As a naive baseline we use bicubic interpolation to up-sample. We test four image SR methods: VDSR [4], RCAN [15], SRGAN [7], and ESRGAN [13], where we super-resolve each frame independently. We test four video SR methods: SPMC [11], DUF [3], SoSR, and ToSR. SoSR and ToSR are proposed by ourselves as new video SR methods for facilitating action recognition rather than for PSNR [14]. All the compared methods, excluding bicubic, are based on CNN. For each method we use the pretrained model provided by the corresponding authors. It is worth noting that the compared methods use different training data, but none of the training data has overlap with HMDB51.

Spatial quality metrics. For spatial quality, we consider MSE and SSIM. They are calculated by comparing the super-resolved videos against the original videos frame by frame. For each frame, MSE and SSIM are calculated on RGB and luma component, respectively. Then, they are averaged over each video and then over the entire test set. Moreover, we consider the quality not only at the signal level, but also at the semantic level. We use a pretrained action recognition network, TSN [12], to evaluate the action recognition accuracy based on the super-resolved videos. Our used TSN model is trained with HMDB51 split1 training set. TSN is a two-stream network, so for spatial quality we use its spatial stream, *i.e*. action recognition based on the frames.

Temporal quality metrics. For temporal quality, we use the warping error proposed in [6], which is calculated for a super-resolved video V:

$$E_{warp}(V) = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{1}{\sum_{p=1}^{N} M_t(p)} E_{warp}(V_t, V_{t+1})$$
$$E_{warp}(V_t, V_{t+1}) = \sum_{p=1}^{N} M_t(p)[V_t(p) - V_{t+1}^{w}(p)]^2 \quad (1)$$

where T is the number of frames. V_t and V_{t+1} are two consecutive frames in the video. N is the number of pixels per frame, and p denotes each pixel. M_t is a mask standing for whether each pixel is occluded or not. V_{t+1}^{w} is a *warped* frame, *i.e*. $V_{t+1}^{w}(p) = V_{t+1}(p + F_{t\to t+1}(p))$, where $F_{t\to t+1}$ is the optical flow from V_t to V_{t+1}. Optical flow is calculated by Flownet2.0 [2]. Occlusion mask is estimated by [10]. In addition, we use the temporal stream of TSN [12] to evaluate the action recognition accuracy based on the optical flows extracted from the super-resolved videos.

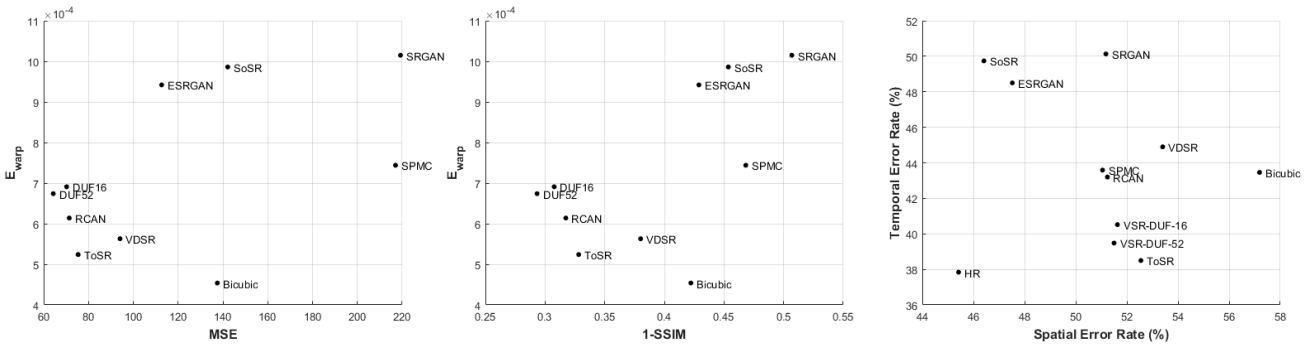

Figure 1. Evaluation results with different spatial and temporal quality metrics, including MSE, 1−SSIM, warping error, recognition error rate in spatial/temporal stream. In each plot, the bottomleft corner is the best. In the third plot, error rates of HR videos are shown for reference.

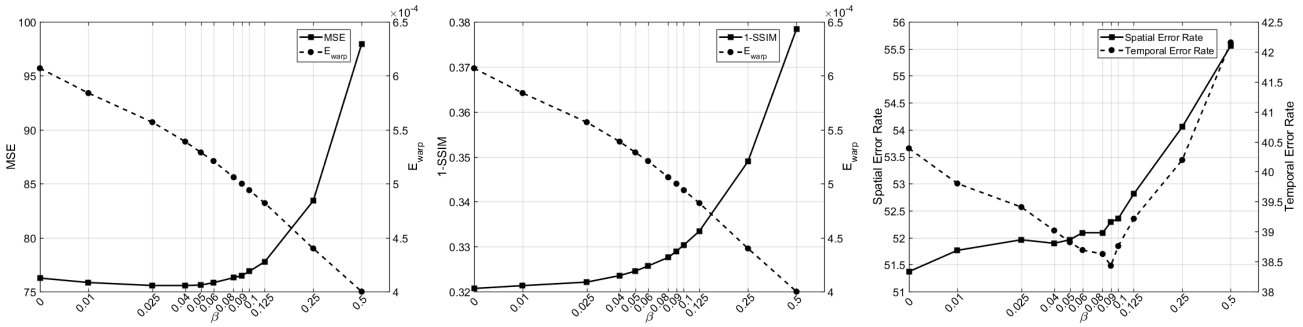

Figure 2. Performance of our SR method trained with 2 and different β values. The horizontal axis is shown in logarithmic scale.

Results. Figure 1 displays the evaluation results of the spatial and temporal quality metrics for each method. For spatial quality, the relative trends of MSE and SSIM are almost consistent, so we analyze the MSE. Of these methods, DUF, RCAN, and VDSR are optimized for MSE. We can find that in terms of MSE, DUF is the best and VDSR is the worst among the three, and all the three are far better than bicubic. However, in terms of warping error, DUF is the worst and VDSR is the best among the three. Indeed, bicubic performs the best among all the compared methods in terms of warping error! This can be understood, since bicubic tends to generate oversmooth frames, which have less temporal inconsistency. To improve spatial quality, advanced SR methods try to add details into the frames, but take the risk of producing flickering artifacts. This seems to indicate a tradeoff between spatial and temporal quality. Moreover, we look at the spatial/temporal quality metrics evaluated by recognition accuracy. The best methods in the spatial stream, SoSR and ESRGAN, all use adversarial loss in their training. However, these methods perform the worst in the temporal stream. These methods also lead to very high warping error. This seems another evidence of the spatial-temporal tradeoff. It is worth noting that the tradeoff is only to some extent. DUF with 52 layers is consistently better than DUF with 16 layers, ESRGAN is consistently better than SRGAN, in every considered metric. The consistently better performance is achieved at the cost of much increased network complexity.

Joint Optimization of Spatial and Temporal. We extend the ToSR method, where we use a siamese network to super-resolve two consecutive frames simultaneously. We use the following loss function:

$$\mathcal{L} = \alpha\|\hat{I}_t - I_t\|_F^2 + \alpha\|\hat{I}_{t+1} - I_{t+1}\|_F^2 + \beta E_{warp}(\hat{I}_t, \hat{I}_{t+1}) \quad (2)$$

where I and $\hat{I}$ denote HR and SR frames. $E_{warp}(\hat{I}_t, \hat{I}_{t+1})$ is similar to that defined in (1), except that the mask is defined as $M_t(p) = \exp(-50[I_t(p) - I_{t+1}^w(p)]^2)$ [6]. We conduct experiments with fixed $\alpha = 0.5$ and variable β. The results are shown in Figure 2. As β increases, warping error decreases monotonously, but MSE or 1−SSIM increases; error rate in the spatial stream increases, but error rate in the temporal stream decreases to some extent. In summary, it seems a difficulty to optimize the spatial and temporal quality metrics simultaneously.

3. Conclusion

In [1], it was proved that minimizing distortion and optimizing perceptual naturalness can be contradictory, which was named perception-distortion tradeoff. Similarly, our empirical results imply a tradeoff between spatial and temporal in video SR. We are seeking a theoretical proof.

References

[1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, pages 6228–6237, 2018.

[2] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, volume 2, pages 2462–2470, 2017.

[3] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, pages 3224–3232, 2018.

[4] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016.

[5] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011.

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

[7] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, and Zehan Wang. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017.

[8] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *ICCV*, pages 2507–2515, 2017.

[9] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, Xinchao Wang, and Thomas S Huang. Learning temporal dynamics for video super-resolution: A deep learning approach. *IEEE Transactions on Image Processing*, 27(7):3432–3445, 2018.

[10] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *German Conference on Pattern Recognition*, pages 26–36, 2016.

[11] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, pages 22–29, 2017.

[12] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016.

[13] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCVW*, pages 63–79, 2018.

[14] Haochen Zhang, Dong Liu, and Zhiwei Xiong. Two-stream oriented video super-resolution for action recognition. *arXiv preprint arXiv:1903.05577*, 2019.

[15] 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*, pages 1–16, 2018.