Multi-Frequency Representation Enhancement with Privilege Information for Video Super-Resolution

Fei Li *
China Agricultural University
leefly072@126.com

Linfeng Zhang *
Tsinghua University
zhanglinfeng1997@outlook.com

Zikun Liu
Samsung Research China
zikun.liu@samsung.com

Juan Lei
Samsung Research China
juan.lei@samsung.com

Zhenbo Li
China Agricultural University
National Innovation Center for Digital Fishery
lizb@cau.edu.cn

The supplementary material is organized as follows: Section 1 presents additional experiments conducted with our MFPI on other tasks, while Section 2, a theoretical analysis of spatial frequency representation enhancement (SFE) and energy frequency representation enhancement (EFE) is provided. Section 3 contains additional visualization results.

1. Extend MFPI to other tasks

Capturing long-range dependencies is a major challenge in super-resolution (SR). However, convolutional neural networks (CNNs) usually cannot capture long-range dependencies due to their limited receptive fields. This problem is more crucial in video super-resolution (VSR) than in single image super-resolution (SISR) because VSR models need to capture not only spatial but also temporal dependencies. We intuitively speculate that our proposed modules are generic and will be useful for other enhancement or restoration tasks such as SISR, denoising, and so on.

Therefore, we applied our method directly to image denoising tasks, such as boosting InvDN [6] on the SIDD [1] dataset by 0.06 dB, and SISR tasks, such as improving TTSR [15] performance by 0.11 dB on the CUFED5 [17] dataset. These results indicate that: (1) our proposed MFPI is also effective on single image (SI) tasks; and (2) the performance improvements on SI tasks are lower than on VSR, indicating that our method is more suitable for VSR.

2. Theoretical analysis

2.1. The effectiveness of fast Fourier transform in SFE

We first elaborate on the properties of the fast Fourier transform (FFT) in the SFE and analysis the computational complexity. The discrete Fourier transform (DFT) has been widely adopted in digital image processing [11, 18], and the FFT can facilitate and improve the speed of the DFT [4]. For clarification, here we only discuss 1-D case with the following formulation:

$$X_r = \sum_{k=0}^{c-1} x_k \exp(-2\pi jrk/C) := \sum_{k=0}^{C-1} x_k W^{rk} \quad (1)$$

where X_r is the r-th coefficient of the FFT, $r = 0, \cdots, C-1$ denotes frequency of the FFT, and x_k denotes the k-th channel, $j = \sqrt{-1}$ and $W = \exp(-2\pi j/C)$ for simplicity.

Proposition 1. *The computation of FFT can be reduced by a factor of* $(\log_2 C)/C$*, where* C *is the number of channels.*

Proof. Assume that x_k is split into y_k, z_k, each of which has half the channels. y_k is composed of the even-numbered channels, and z_k is composed of the odd-numbered channels, which can be formulated as follows:

$$\begin{cases} y_k = x_{2k} \\ z_k = x_{2k+1} \end{cases} \quad k = 0, 1, \cdots, \frac{C}{2} - 1 \quad (2)$$

The corresponding FFT can be formulated as follows:

$$\begin{cases} Y_r = \sum_{k=0}^{C/2-1} y_k \exp(-4\pi jrk/C) \\ Z_r = \sum_{k=0}^{C/2-1} z_k \exp(-4\pi jrk/C) \end{cases} \quad r = 0, 1, \cdots, \frac{C}{2} - 1 \quad (3)$$

* The first two authors contribute equally. This work is done during the internship of F. Li in Samsung. Z. Li is the corresponding author.

Notice the equations 1 and 3 and the properties of FFT [11, 3], we have:

$$X_r = \sum_{k=0}^{C/2-1} \left\{ y_k \exp(\frac{-4\pi jrk}{C}) + z_k \exp\left(-\frac{2\pi jr}{C}(2k+1)\right)\right\}$$
$$= Y_r + \exp\left(-2\pi jr/C\right) \cdot Z_r \quad (4)$$

where $r = 0, 1, \cdots, N-1$. When r values greater than $C/2$, the transformations Y_r and Z_r periodically repeat the case when $r < C/2$. Hence, by replacing r in equation 4 to $r + C/2$, we have:

$$X_{r+C/2} = Y_r + \exp\left(-2\pi j\left[r + \frac{C}{2}\right]/C\right) \cdot Z_r$$
$$= Y_r - \exp(-2\pi jr/C) \cdot Z_r, \quad 0 \le r < C/2 \quad (5)$$

Considering the above cases and combining the equations 1, 4, 5, we can derive:

$$X_r = Y_r + W^r Z_r$$
$$X_{r+\frac{c}{2}} = Y_r - W^r Z_r \quad (6)$$

The equation 6 shows that it is possible to simplify the calculation of FFT with C channels. By computing the FFTs with two sequences of $C/2$ channels each, followed by the computation of Y_r (or Z_r) with sequences of $C/4$ channels, the process can be continued, provided that each function has a number of channels divisible by 2. Thus, the conclusion can be proven.

□

2.2. The effectiveness of discrete cosine transform in EFE

To enhance its representational capabilities, the EFE employs the discrete cosine transform (DCT) to convert features into the frequency domain [2, 10]. Before delving into its application, we offer a concise introduction to the DCT, a commonly employed technique in signal processing and data compression [13, 9].

We carry over the notations from the previous for simplicity, where $f \in \mathbb{R}^{H\times W}$ is the input feature with two dimensions, H, W is the height, width of x. Then the two-dimensional DCT can be formulated as follows:

$$\mathcal{F}_c(i,j) = \frac{2}{\sqrt{HW}}\alpha(i)\alpha(j)\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} f_{i,j} \quad (7)$$

$$\mathcal{K}_{h,w}^{i,j} = \cos\left(\frac{(2h+1)i\pi}{2H}\right)\cos\left(\frac{(2w+1)j\pi}{2W}\right) \quad (8)$$

$$f_c^{h,w} = \mathcal{K}_{h,w}^{i,j} \times \mathcal{F}_c\left(f_{i,j}\right) \quad (9)$$

where $\mathcal{F}_c$ denotes DCT operation, $\mathcal{K}$ denotes the basis function of DCT, $\alpha(x) = 1/\sqrt{2}$ for $x = 0$ and $\alpha(x) = 1$ otherwise [13].

The mean-square reconstruction error (MSRE) between transformed images f_c and reference images $\hat{f}$, we could be defined as:

$$\bar{E}_{mse} = \frac{1}{N}\sum_{n=0}^{N}\left\{\frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\left[f_{c,n}^{i,j} - \hat{f}_{c,n}^{i,j}\right]^2\right\}$$
$$= \frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} E\left\{\left[f_c^{i,j}\right]^2\right\} \cdot [1 - \phi(i,j)] \quad (10)$$

where $E\left\{\left[f_c^{i,j}\right]^2\right\}$ is the expectation of DCT components at location (i, j), n denotes the number of input frames and H, W denote the height, width of the input feature, $\phi(i, j)$ is the masking function [13]. And for the variance of DCT components, we denote $\bar{f}$ as the mean of input feature set $\{f_1, f_2, \cdots, f_n\}$, then we replace f_n with $f_n - \bar{f}$ and $\hat{f}_n$ with $\hat{f}_n - \bar{f}$ in the equation 10, the MSRE between the feature set $\left\{f_1 - \bar{f}, f_2 - \bar{f}, \cdots, f_n - \bar{f}\right\}$ and their approximations $\left\{\hat{f}_1 - \bar{f}, \hat{f}_2 - \bar{f}, \cdots, \hat{f}_n - \bar{f}\right\}$:

$$E_{mse} = \frac{1}{N}\sum_{n=0}^{n}\left\{\frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\left[f_n^{i,j} - \bar{f}^{i,j} - \left(\hat{f}_n^{i,j} - \bar{f}^{i,j}\right)\right]^2\right\}$$
$$= \frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\sigma^2_{f_c^{i,j}} \cdot [1 - \phi(i,j)] \quad (11)$$

where $\sigma^2_{f_c^{i,j}}$ is the variance of DCT components. The equation 11 denotes the total MSRE which is equal to the average of the variances of the transform components when $\phi(u, w) = 0$. Therefore, the equation 11 holds when we regard the pixels of the feature $f_n - \bar{f}$ generated by a random process with zero mean and *known* variance. And the DCT features are selected to minimize the MSRE in equation 11 to obtain optimal solution [12].

In the EFE, the energy function refines the input feature set as $\{\mathbf{e}_1, ..., \mathbf{e}_N\}$, then we rearrange the mean and variance as follows:

$$\bar{E}_{en\_mse} = \frac{1}{N}\sum_{n=0}^{N}\left\{\frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\left[e_{c,n}^{i,j} - \hat{f}_{c,n}^{i,j}\right]^2\right\}$$
$$= \frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} E\left\{\left[e_c^{i,j}\right]^2\right\} \cdot [1 - \phi(i,j)] \quad (12)$$

$$E_{en\_mse} = \frac{1}{N}\sum_{n=0}^{n}\left\{\frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\left[e_n^{i,j} - \bar{f}^{i,j} - \left(\hat{f}_n^{i,j} - \bar{f}^{i,j}\right)\right]^2\right\}$$
$$= \frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\sigma^2_{e_c^{i,j}} \cdot [1 - \phi(i,j)] \quad (13)$$

Since the traditional DCT truncate the coefficients from the representation, and thereby introduces errors [8]. In the EFE, the energy DCT can be estimated from energy value e, and the $e_{i,j}$ act as the scale factors for the unknown feature. Moreover, the equation 13 holds when $e_{i,j}$ can be obtained from the input feature set and applied to the learnable DCT filter. Therefore, our EFE, compared to standard DCT, can be adaptive to process unknown inputs. In the ablation experiments, EFE achived 0.6 dB higher than fixed coefficients with DCT initialization.

3. Visual comparisons on the test dataset.

Fig. 1 shows VSR results of MFPI and those of the state-of-the-art methods on several challenging images from different datasets [7, 14, 5, 16].

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1692–1700, 2018. 1

[2] Vladimir Britanak, Patrick C Yip, and Kamisetty Ramamohan Rao. *Discrete cosine and sine transforms: general properties, fast algorithms and integer approximations*. Elsevier, 2010. 2

[3] William T Cochran, James W Cooley, David L Favin, Howard D Helms, Reginald A Kaenel, William W Lang, George C Maling, David E Nelson, Charles M Rader, and Peter D Welch. What is the fast fourier transform? *Proceedings of the IEEE*, 55(10):1664–1674, 1967. 2

[4] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965. 1

[5] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):346–360, 2013. 3, 4

[6] Yang Liu, Zhenyue Qin, Saeed Anwar, Pan Ji, Dongwoo Kim, Sabrina Caldwell, and Tom Gedeon. Invertible denoising network: A light solution for real noise removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13365–13374, 2021. 1

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 3, 4

[8] Zhengjun Pan, Alistair G Rust, and Hamid Bolouri. Image redundancy reduction for neural network classification using discrete cosine transforms. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN)*, volume 3, pages 149–154. IEEE, 2000. 3

[9] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 783–792, 2021. 2

[10] K Ramamohan Rao and Ping Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014. 2

[11] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:980–993, 2021. 1, 2

[12] Zhengjun Pan Rod, Rod Adams, and Hamid Bolouri. Dimensionality reduction of face images using discrete cosine transforms for recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000. 2

[13] Xing Shen, Jirui Yang, Chunbo Wei, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Xiaoliang Cheng, and Kewei Liang. Dct-mask: Discrete cosine transform mask representation for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8720–8729, 2021. 2

[14] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 3, 4

[15] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5791–5800, 2020. 1

[16] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3106–3115, 2019. 3, 4

[17] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7982–7991, 2019. 1

[18] Man Zhou, Hu Yu, Jie Huang, Feng Zhao, Jinwei Gu, Chen Change Loy, Deyu Meng, and Chongyi Li. Deep fourier up-sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

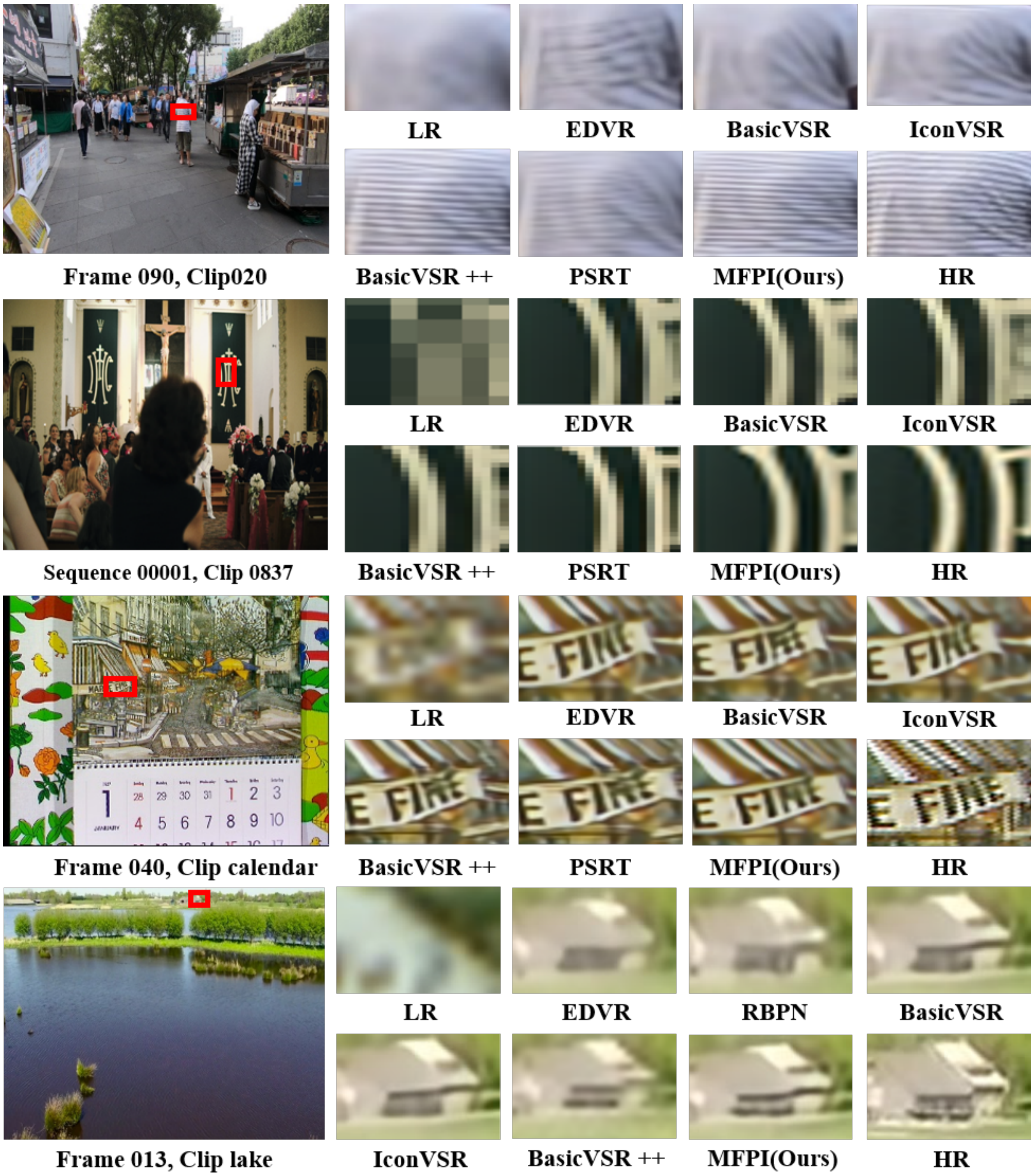

Figure 1: Visual results on REDS4 [7], Vimeo-90K-T [14], Vid4 [5], and UDM10 [16]. Zoom in to see better visualization.