RETHINKING IMAGE AND VIDEO RESTORATION: AN INDUSTRIAL PERSPECTIVE

Huan Yang
Microsoft Research Asia
Beijing, China
huayan@microsoft.com

Abstract

Image and video restoration as a fundamental low-level vision task can significantly improve the visual quality and benefit a lot of downstream computer vision tasks (e.g., video surveillance and satellite imagery). However, early works mainly focus on some ideal settings that strongly limit their applications. Recent years have witnessed increasing interest in designing restoration approaches under real-world scenarios. In this article, we rethink the challenges of restoration deployment from an industrial perspective and share our experiences from three aspects: network design, model training, and deployment environments. According to those thinking and our solutions, we conclude the current progress of restoration tasks and point out some future opportunities that we will focus on.

I. Industrial Applications of Restoration

Image and video restoration aims to recover high-quality content from its degraded counterpart. It consists of many low-level vision tasks, e.g., super-resolution, inpainting, light enhancement, etc. The degradation usually varies between those tasks. In super-resolution, it could be a down-sampling process that reduces the content resolution. Specific to video super-resolution, reducing the frame rate of videos in temporal dimension could also be an option for degradation. Moreover, in light enhancement, the degradation will be exposure adjustments. Although the degradation is varied, those tasks share the same optimization target which is to recover the high-quality content and improve its visual quality. Such a goal encourages the industry to deploy those methods in real scenarios to advance user experiences (as shown in Fig. 1).

With the development of high-definition display devices (e.g., 8K televisions) in recent years, there is an increasing need for high-quality content to release the power of those devices and bring new visual enjoyment. However, such high-quality content is hard to access due to the limitation of network bandwidth and media sources. To mitigate this problem, in real deployments, restoration techniques play an important role to bridge the gap between content sources and display devices. Specific to high-definition television, super-resolution and frame interpolation techniques are usually adopted to align the spatial and temporal resolution, respectively [8], [16], [17]. From the user aspect, restoration techniques could also benefit and level up their content quality during the image and video capturing and retouching [28], [29]. User capturing is usually subject to light conditions and camera sensors and results in low-quality content. With the help of restoration techniques, e.g., color enhancement and relighting could significantly improve the visual quality and make a more vivid illustration [22].

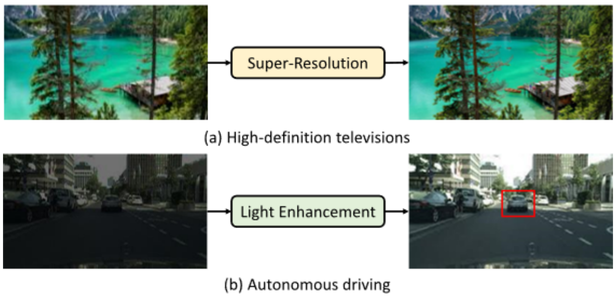

Fig. 1 Industrial applications of restoration. (a) Enhancing content quality by super-resolution techniques for high-definition televisions. (b) Improving detection accuracy by light enhancement techniques for autonomous driving.

Except for the remarkable progress of the restoration in low-level vision applications that aim to improve visual quality, it also benefits a broad range of high-level vision tasks, e.g., recognition and detection [10], [19]. Specific to the industrial scenarios, in video surveillance, restoration especially for super-resolution, deblur, and denoising has been widely used [1], [33]. With the help of these techniques, the content quality could be recovered with more details and further boost the accuracy of anomaly detection and face verification, etc. In other words in autonomous driving, restoration techniques could also encourage the applications of traffic lane detection as well as traffic light classification under extreme weather conditions (e.g., the foggy weather) [4]. By equipping the techniques of deraining and dehazing, a clearer scene could be captured, and detailed traffic information could be analyzed which leads to a smarter decision of the control system.

Overall, in recent years, many industrial applications have witnessed the power of restoration and tried to deploy it in their scenarios to further boost performance or reduce cost. In the following sections, we will guide the readers to see the challenges of restoration deployment in Sec. II and introduce our industrial solutions and thinking in Sec. III. In the last, we will summarize the current progress and point out some potential opportunities that people could work on in the future in Sec. IV.

II. Industrial Challenges of Restoration

Even though some restoration methods have already been successfully deployed in real industrial scenarios, there still have a lot of challenges that limit these methods to release their full capacities [24], [30].

The first challenge is the trade-off between computational costs and performance improvements. In real industrial scenarios, slight computational cost increases will affect other components a lot (e.g., power consumption, and memory cost). How to design specific networks that fit the industrial requirements and achieve a good balance between the costs and gains will be the most important problem [12].

The second challenge is about the settings and scenarios. There is a big gap between research settings and industrial scenarios. For example, in image super-resolution, research settings usually take bicubic downsampling as its only degradation method [2]. While in real products, such scenarios could be very complex including noise, blur as well as JPEG compression artifacts. Directly applying existing research methods to products may result in visual unpleasant images [9], [34].

The last challenge is the hardware deployment problems. Restoration, as a dense prediction task, usually requires more resources than traditional high-level vision tasks and its computational cost highly depends on its input resolution. This requires the model to be specially designed for some specific hardware like INT8 inference for NPU. In recent years, more and more industrial companies find that designing hardware-friendly models would also be a challenging but of great potential direction in restoration [3], [20].

III. Industrial Solutions to Restoration

In this section, we will introduce some of our existing solutions to the above three challenges: network design, model training settings, and hardware deployment environments.

A. Network Design

Convolutional neural network (CNN) has been widely used in computer vision tasks and achieved great success [5-7], [11], [23], [31]. However, recent works on Transformer [21] further improve the performance by a large margin [16], [25], [30], [32]. This is achieved by leveraging the long-range dependency between different regions. Specific to restoration, such a design could make full use of the self-exemplar prior to input images and produce visually more pleasant results than CNN based methods under a fixed computational cost. Motivated by this, we propose a novel Texture Transformer network for image Super-Resolution (TTSR) [26], in which the LR and Ref images are formulated as queries and keys in a Transformer, respectively. As shown in Fig. 2, TTSR consists of four closely-related modules optimized for image restoration tasks, including a learnable texture extractor by DNN, a relevance embedding module, a hard-attention module for texture transfer, and a soft-attention module for texture synthesis. Such a design encourages joint feature learning across LR and Ref images, in which deep feature correspondences can be discovered by attention, and thus accurate texture features can be transferred. Extensive experiments show that TTSR achieves significant improvements over state-of-the-art approaches on both quantitative and qualitative evaluations.

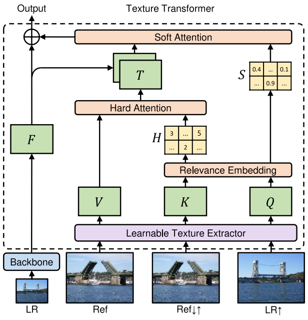

Fig. 2. The first Transformer based image super-resolution network [26].

To extend the capability of TTSR, we further consider the temporal information in video super-resolution and proposed a novel Trajectory-aware Transformer for Video Super-Resolution (TTVSR) [15]. As shown in Fig. 3, we formulate video frames into several pre-aligned trajectories which consist of continuous visual tokens. For a query token, self-attention is only learned on relevant visual tokens along spatial-temporal trajectories. Compared with vanilla vision Transformers, such a design significantly reduces the computational cost and enables Transformers to model long-range features. Experimental results demonstrate the superiority of the proposed TTVSR over state-of-the-art models, by extensive quantitative and qualitative evaluations in four widely used video super-resolution benchmarks.

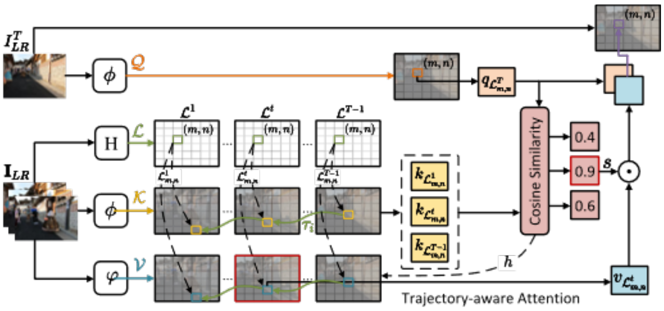

Fig. 3. An overview of our proposed trajectory-aware transformer for video super-resolution [15].

B. Model Training Settings

In the real deployment of restoration methods, there is a large gap between research settings and real-world scenarios. Directly applying those methods may result in visually unpleasant results. To mitigate this problem, we propose a Degradation-guided Meta-restoration network for blind Super-Resolution (DMSR) that facilitates image restoration for real cases [27]. As shown in Fig. 4, DMSR consists of a degradation extractor that estimates the degradations in LR inputs and guides the restoration networks to predict restoration parameters for different degradations on-the-fly. Through such an optimization, DMSR outperforms SOTA by a large margin on three widely used benchmarks.

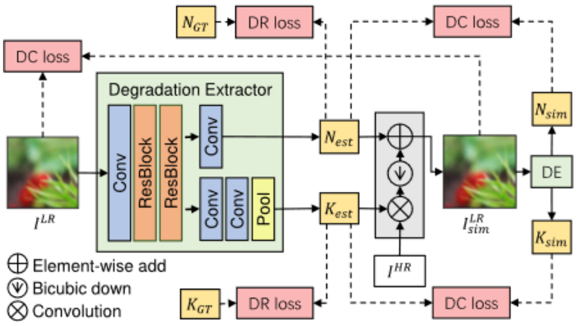

Fig.4. The core component of our proposed degradation-guided meta-restoration network for blind super-resolution [27].

Compared with image super-resolution, real-world video super-resolution further introduces compression artifacts [13]. To attack this challenge, we propose a novel Frequency-Transformer for compressed Video Super-Resolution (FTVSR) that conducts self-attention over a joint space-time-frequency domain [18]. As shown in Fig. 5, we first divide a video frame into DCT patches. Then we study different self-attention schemes and discover that a ``divided attention'' which conducts a joint space-frequency attention before applying temporal attention on each frequency band, leads to the best video enhancement quality. Experimental results on two widely used video super-resolution benchmarks show that FTVSR outperforms state-of-the-art approaches on both uncompressed and compressed videos with clear visual margins.

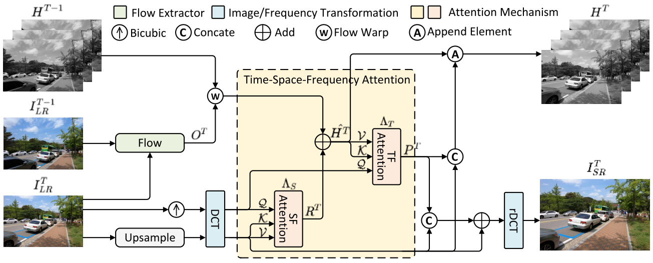

Fig. 5. An overview of our proposed frequency-transformer for compressed video super-resolution [18].

C. Hardware Deployment Environments

Recent works have less explored the advantage of different operations on real hardware. Compared with matrix multiplication operations that take most of the inference time, addressing operations take only a small portion of the whole computational costs. Motivated by this, we propose a novel learnable context-aware 4-Dimensional LookUp Table (4D LUT) for image enhancement [14]. As shown in Fig. 6, we first introduce a lightweight context encoder and a parameter encoder to learn a context map and a group of coefficients for LUTs, respectively. Then, the context-aware 4D LUT is generated by integrating multiple basis 4D LUTs via the coefficients. Finally, the input image is enhanced by feeding into the fused context-aware 4D LUT with the context map via quadrilinear interpolation. With such a design, most computational costs are spent on the addressing operation which is super-fast on real hardware. Experimental results demonstrate that our proposed 4D LUT outperforms other state-of-the-art methods in widely used benchmarks while keeping a real-time speed on most low-end devices.

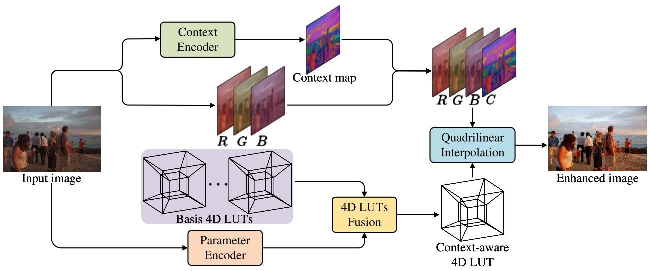

Fig. 6. A system overview of our proposed 4-dimensional lookup table for image enhancement [14].

IV. Industrial Opportunities for Restoration

Despite the remarkable progress in restoration deployments in industrial scenarios, there is still a long way to go. In the future, two potential opportunities have been witnessed to break the gap and take a step further in this area. The first is the restoration of extremely low-quality content under real scenarios with complex degradations. Recovering highly damaged content could not only improve the visual quality but also bring a new high-level understanding of the content and benefit many downstream applications. The second opportunity is to design models that highly depend on the hardware. Such a strategy could enable hardware-dependent optimizations and make full use of the hardware to achieve higher quality improvements. In the future, we will focus on these proposed opportunities and design practical solutions to restoration in more industrial deployment scenarios.

References

[1] Marco Cristani, Dong Seon Cheng, Vittorio Murino, and Donato Pannullo. Distilling information with super-resolution for video surveillance. In Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks, pages 2–11, 2004.
[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. TPAMI, 38(2):295– 307, 2015.
[3] Zongcai Du, Jie Liu, Jie Tang, and Gangshan Wu. Anchor-based plain net for mobile image super-resolution. In CVPR, pages 2494–2502, 2021.
[4] Syeda Nyma Ferdous, Moktari Mostofa, and Nasser M Nasrabadi. Super resolution-assisted deep aerial vehicle detection. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, volume 11006, pages 432–443. SPIE, 2019.
[5] Jianlong Fu, Tao Mei, Kuiyuan Yang, Hanqing Lu, and Yong Rui. Tagging personal photos with transfer deep learning. In WWW, pages 344–354, 2015.
[6] Jianlong Fu and Yong Rui. Advances in deep learning approaches for image tagging. APSIPA Transactions on Signal and Information Processing, 6, 2017.
[7] Jianlong Fu, Jinqiao Wang, Yong Rui, Xin-Jing Wang, Tao Mei, and Hanqing Lu. Image tag refinement with view-dependent concept representations. TCSVT, 25(8):1409–1422, 2014.
[8] Kyohei Goto, Fumiya Nagashima, Tomio Goto, Satoshi Hirano, and Masaru Sakurai. Super-resolution for high-resolution displays. In GCCE, pages 309–310. IEEE, 2014.
[9] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind superresolution with iterative kernel correction. In CVPR, pages 1604–1613, 2019.
[10] Bahadir K Gunturk, Aziz Umit Batur, Yucel Altunbasak, Monson H Hayes, and Russell M Mersereau. Eigenface-domain super-resolution for face recognition. TIP, 12(5):597–606, 2003.
[11] Kibeom Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, and Hyeran Byun. Domain-aware universal style transfer. In ICCV, pages 14609– 14617, 2021.
[12] Xiangtao Kong, Hengyuan Zhao, Yu Qiao, and Chao Dong. ClassSR: A general framework to accelerate super-resolution networks by data characteristic. In CVPR, pages 12016–12025, 2021.
[13] Yinxiao Li, Pengchong Jin, Feng Yang, Ce Liu, Ming-Hsuan Yang, and Peyman Milanfar. COMISR: Compression-informed video superresolution. In ICCV, pages 2543–2552, 2021.
[14] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. 4D LUT: Learnable context-aware 4d lookup table for image enhancement. arXiv preprint arXiv:2209.01749, 2022.
[15] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In CVPR, pages 5687–5696, 2022.
[16] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. TTVFI: Learning trajectory-aware transformer for video frame interpolation. arXiv preprint arXiv:2207.09048, 2022.
[17] Yasutaka Matsuo and Shinichi Sakaida. Super-resolution for 2k/8k television using wavelet-based image registration. In GlobalSIP, pages 378–382. IEEE, 2017.
[18] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for compressed video superresolution. arXiv preprint arXiv:2208.03012, 2022.
[19] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In CVPR, 2019.
[20] Dehua Song, Yunhe Wang, Hanting Chen, Chang Xu, Chunjing Xu, and DaCheng Tao. AdderSR: Towards energy efficient image superresolution. In CVPR, pages 15648–15657, 2021.
[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. NeurIPS, 30, 2017.
[22] Baoyuan Wang, Yizhou Yu, Tien-Tsin Wong, Chun Chen, and Ying-Qing Xu. Data-driven image color theme enhancement. TOG, 29(6):1–10, 2010.
[23] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. AI Coach: Deep human pose estimation and analysis for personalized athletic training assistance. In ACM MM, pages 374–382, 2019.
[24] Jun Xiao, Xinyang Jiang, Ningxin Zheng, Huan Yang, Yifan Yang, Yuqing Yang, Dongsheng Li, and Kin-Man Lam. Online video superresolution with convolutional kernel bypass graft. arXiv preprint arXiv:2208.02470, 2022.
[25] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In CVPR, pages 5036–5045, 2022.
[26] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In CVPR, pages 5791–5800, 2020.
[27] Fuzhi Yang, Huan Yang, Yanhong Zeng, Jianlong Fu, and Hongtao Lu. Degradation-guided meta-restoration network for blind super-resolution. arXiv preprint arXiv:2207.00943, 2022.
[28] Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In ICCV, pages 4633–4641, 2015.
[29] Huan Yang, Baoyuan Wang, Noranart Vesdapunt, Minyi Guo, and Sing Bing Kang. Personalized exposure control using adaptive metering and reinforcement learning. TVCG, 25(10):2953–2968, 2018.
[30] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatialtemporal transformations for video inpainting. In ECCV, pages 528–543. Springer, 2020.
[31] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. TVCG, 2022.
[32] Yanhong Zeng, Huan Yang, Hongyang Chao, Jianbo Wang, and Jianlong Fu. Improving visual quality of image synthesis by a token-based generator with transformers. NeurIPS, 34:21125–21137, 2021.
[33] Liangpei Zhang, Hongyan Zhang, Huanfeng Shen, and Pingxiang Li. A super-resolution reconstruction algorithm for surveillance images. Signal Processing, 90(3):848–859, 2010.
[34] Heliang Zheng, Huan Yang, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning conditional knowledge distillation for degraded-reference image quality assessment. In ICCV, pages 10242–10251, 2021.