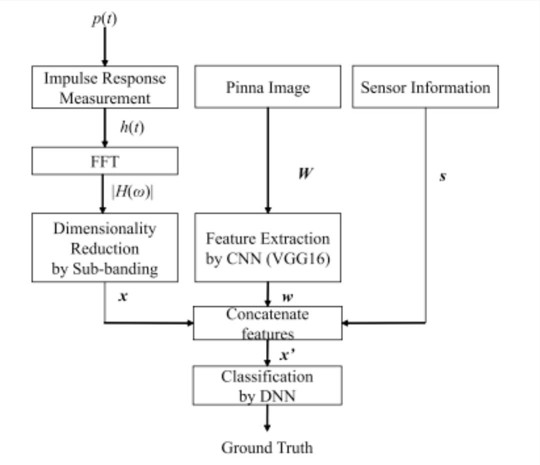
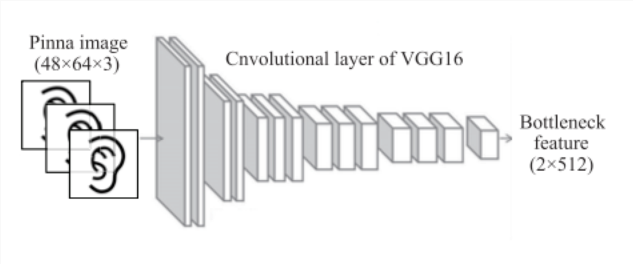
IEEE CONSUMER TECHNOLOGY SOCIETY (CTSOC)

ISSUE: SEPTEMBER 2022

CTSOC-NCT NEWS ON CONSUMER TECHNOLOGY





AUTHENTICATION SYSTEMS BASED ON THE ACOUSTIC TRANSFER FUNCTION OF THE PINNA (PRTF: PINNA RELATED TRANSFER FUNCTION) HAVE BEEN INVESTIGATED. OVERVIEW OF PERSONAL AUTHENTICATION USING PINNA. THE FEATURE EXTRACTION OF PINNA IMAGES WAS MADE USING VGG16. PINNA IMAGES ARE INPUT INTO THE MODEL FROM WHICH THE FULLY CONNECTED LAYER OF VGG16 IS REMOVED AND THE VALUE OF THE LAST CONVOLUTION LAYER IS EXTRACTED AS A BOTTLENECK FEATURE.

TABLE OF CONTENTS

EDITOR'S NOTE

On behalf of the Editorial Board of IEEE CTSoc News on Consumer Technology editor-in-chief Wen-Huang Cheng and co-editors, Yafei Hou, Luca Romeo, and Yuen Peng Loh, I am delighted to introduce the 2022 September issue of the News on Consumer Technology.

This issue starts with a cover story about biometric authentication technology for smartphones published in IEEE Transactions on Consumer Electronics. Besides face and fingers, the ears as a biometric feature enables human identification, and has been the subject of research on personal authentication when hands are wet, or a person is wearing a mask. The article proposes a multimodal personal authentication on smartphones. Half total error rate (HTER) has been improved with large margins for multimodal authentication using images and sensor data.

Next, we provide an interview with Prof. Minoru Kuribayashi, from Okayama University, Japan. Minoru Kuribayashi received B.E., M.E., and D.E degrees from Kobe University, Japan, in 1999, 2001, and 2004. His research interests include multimedia security, digital watermarking, cryptography, and coding theory. The interview shows the protection of multimedia content from illegal copying and ownership infringement, and the specific technologies that Prof. Minoru has devoted to in image, video and audio domains.

Finally, I would like to draw your attention the featured article from Dr. Huan Yang, a Senior Researcher at Microsoft Research, which introduces a rethinking-type article on image and video restoration from an industrial perspective. As a fundamental low-level vision task, visual restoration can significantly improve the visual quality and benefits a lot of downstream computer vision tasks, like video surveillance and satellite imagery. Recent years have witnessed the increasing interest for real-world enhancement scenarios. To solve the real challenges, the article has discussed network design, model training settings, and hardware deployment environments.

Happy reading!

Dr. Jianlong Fu

Editor of NCT

Editor: Luca Romeo

ARTICLE TITLE

Multimodal Personal Ear Authentication Using
Acoustic Ear Feature for Smartphone Security

AUTHOR(S)

Shunji Itani , Shunsuke Kita, and Yoshinobu Kajikawa, Senior Member, IEEE

JOURNAL TITLE

IEEE Transactions on Consumer Electronics

JOURNAL VOLUME AND ISSUE

Volume: 68, Issue: 1

DATE OF THE ARTICLE

February 2022

PAGE NUMBERS FOR THE ARTICLE

77 - 84

In recent years, biometric authentication technology for smartphones has become widespread, with the mainstream methods being fingerprint authentication and face recognition. However, fingerprint authentication cannot be used when hands are wet, and face recognition cannot be used when a person is wearing a mask. Like the face and fingers, the ear as a biometric contains features that enable human identification and has been the subject of research on personal authentication. Authentication systems based on the acoustic transfer function of the pinna (PRTF: Pinna Related Transfer Function) have been investigated. However, the authentication accuracy decreases due to the positional fluctuation across each measurement. In this paper, they propose multimodal personal authentication on smartphones using PRTF. The pinna image and positional sensor information are used with the PRTF, and the effectiveness of the authentication method is examined. Half total error rate (HTER) of 9.3% for single-modal authentication using only PRTF was improved to 1.6% for multimodal authentication using images and sensor data. They demonstrate that the proposed authentication system can compensate for the positional changes in each measurement and improve the robustness.

INTERVIEW WITH PROF. MINORU KURIBAYASHI, OKAYAMA UNIVERSITY, JAPAN

Minoru Kuribayashi received B.E., M.E., and D.E degrees from Kobe University, Japan, in 1999, 2001, and 2004. He was a Research Associate and an Assistant Professor at Kobe University from 2002 to 2007 and from 2007 to 2015, respectively. Since 2015, he has been an Associate Professor in the Graduate School of Natural Science and Technology, Okayama University. His research interests include multimedia security, digital watermarking, cryptography, and coding theory. He serves as an associate editor of JISA and IEICE. He is a vice chair of APSIPA TC of Multimedia Security and Forensics, and a TC member of IEEE SPS Information Forensics and Security. He received the Young Professionals Award from IEEE Kansai Section in 2014, and the Best Paper Award from IWDW 2015 and 2019. He is a senior member of IEEE and IEICE.

Could you briefly introduce your research?

My primacy research interest is the protection of multimedia content from illegal copying and ownership infringement. One promising solution is the use of data hiding techniques that insert tiny signals into multimedia content without degrading its perceptual quality. This is a kind of active approach to check the ownership and soundness of the content by slightly distorting it. On the other hand, analyzing distortions caused by editing or modifying content is the passive approach, which is called multimedia forensics. In both cases, the handling of tiny signals involved in multimedia content is important, and a combination of signal processing and machine learning techniques is inevitable in this research

What is your main work?

My main contribution in this research area is to achieve traceability of multimedia content, which is called digital fingerprinting. Uniquely assigned information, called fingerprint, are embedded into multimedia content with the help of data hiding techniques. Fingerprinting requires consideration of two difficult requirements: asymmetric property in buyer-seller protocols and collusion resistance.

Asymmetry refers to the information gap between buyers and sellers. Typically, it is assumed that the buyer will violate the ownership rights of the content by redistributing the illegal copies. However, seller can frame innocent buyers by distributing the fingerprinted content and claiming that the illegal copies were leaked from the buyer. To address this issue, cryptographic protocols between the buyer and seller have been investigated to assure that only the buyer can obtain the content containing his/her fingerprint information.

In a fingerprinting setup, different versions of the same content are distributed to multiple users, and hence, a coalition of illegal users can compare uniquely fingerprinted content and modify or delete the fingerprint information, which is called a collusion attack. Therefore, resistance against collusion attack has been investigated both in terms of encoding fingerprint (approach of coding theory) and modulating signals (approach of communication theory).

What is a challenging topic in multimedia security?

Due to the advance of deep learning (DL) technology, creation and manipulation of multimedia content have progressed to the point where they can now ensure a high degree of realism. In movies, realistic characters and exciting scenes can be created according to the interest of movie director without having them perform dangerous actions. Artificially created newscasters can continue to work on news multicasts in smooth tones. On the other hand, the DL-based signal processing operations of image, video, and audio may be abused to generate fake news like misinformation that mimics famous people. By using DL technology with multiple videos of people as supervised data, it is possible to create fake content that realistically reproduces false statements. One famous fake content is DeepFake, which is hard to distinguish from real or fake. DeepFake is basically reproduced media obtained by injecting or replacing some information into the target content. For the classification of DeepFake, unnatural signals involved in multimedia content are analyzed by using various signal processing operations and the DL techniques, which is called multimedia forensics.

What are the difficult problems in the multimedia forensics?

With the progress of defense techniques, attackers will develop content generator models according to the weaknesses exploited by the defense techniques. The use of generative adversarial network enables the generator to update and improve the performance without any theoretical and mathematical formulation if a reasonable amount of computing resources is available.

DL technology helps us to analyze the traces (tiny signals) in multimedia content for classifying whether it is fake or not. However, the reliability of DL-based system will be dropped due to the vulnerabilities against adversarial attacks such that intentional perturbations are crafted to fool the system by misleading the results. When considering defense techniques, it is necessary to assume defense-specific adversarial attacks. From the attacker’s perspective, such defense techniques can be considered to craft adversarial perturbations.

Could you explain about your current research projects?

- **EIG CONCERT-Japan**

The European Interest Group (EIG) CONCERT-Japan is an international joint initiative to support and enhance science, technology and innovation (STI) cooperation between European countries and Japan. **Detection of fake newS on SocIal MedIa pLAtfoRms (DISSIMILAR)** is a project within **CONCERT-Japan** (Connecting and Coordinating European Research and Techology Development with Japan) programme and 7th Joint Call on **"ICT for Resilient, Safe and Secure Society"** which is realized by the consortium consisting of researchers from: **Okayama University** (Japan), **Fundació per a la Universitat Oberta de Catalunya** (Spain), and **Warsaw University of Technology** (Poland). The project will be realized between **June 2021 and May 2024**. The funding is provided through grants number **PCI2020-120689-2** (Spanish Government), **JPMJSC20C3** (Japanese Government), and **EIG CONCERT-JAPAN/05/2021** (National Centre for Research and Development, Poland).

DISSIMILAR

http://dissimilar.ii.pw.edu.pl/

DISSIMILAR combines research on watermarking and machine learning with a user experience study to develop novel technological tools to help users to distinguish between original and altered media content. With the proposed tools we expect online social media users to be able to identify the authorship of a content to distinguish legitimate from fake multimedia content in an autonomous manner, without the need for the platform manager to control or validate any content. Furthermore, the watermarking tools will also provide content creators with a way to protect their creations against manipulation. Hence, the aim of this project is to provide user centric tools that combat disinformation and that contribute minimizing the redistribution of fake news in online social media.

What do you expect about the changes in the consumer technology?

In the early days of the Internet, anyone was free to enter the network market and offer services such as e-mail, chat, social network service, video streaming. With the proliferation of networks, it becomes necessary to protect against malicious activities such as eavesdropping, malicious software (malware), denial-of-service attack, phishing attack, and spread of fake news.

As for the e-mail, when setting up a new mail server, formal registration with authentication servers on a public network is inevitable to legitimize the services offered by the server.

In the near future, the distribution of multimedia content will be controlled over the network to prevent the spread fake content. Content Authenticity Initiative, which a community of media and tech companies, NGOs, academics, and others working to promote adoption of an open industry standard for content authenticity and provenance, is one of the new trends of services related to multimedia. One of potential framework is the Content Credentials functionality which allows the history of content capture, editing, and publication to be verified, including the provenance and attribution. By making the history of multimedia content transparent, it can ensure the reliability of information distribution on cyberspace.

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].

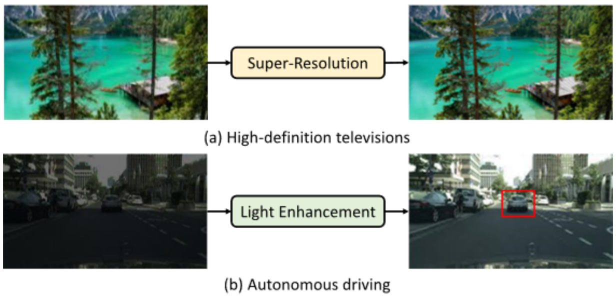


Fig. 1 Industrial applications of restoration. (a) Enhancing content quality by super-resolution techniques for high-definition televisions. (b) Improving detection accuracy by light enhancement techniques for autonomous driving.

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.

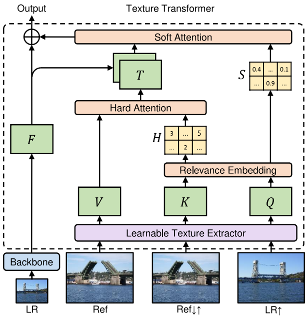


Fig. 2. The first Transformer based image super-resolution network [26].

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.

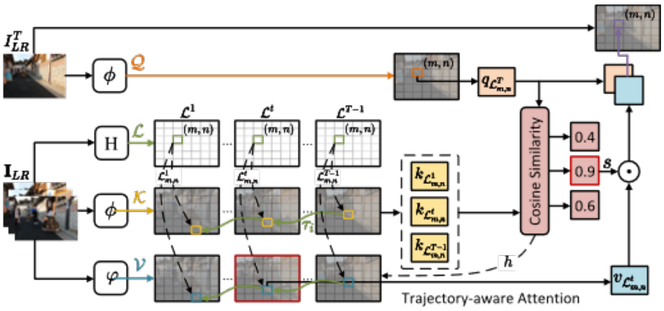


Fig. 3. An overview of our proposed trajectory-aware transformer for video super-resolution [15].

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.

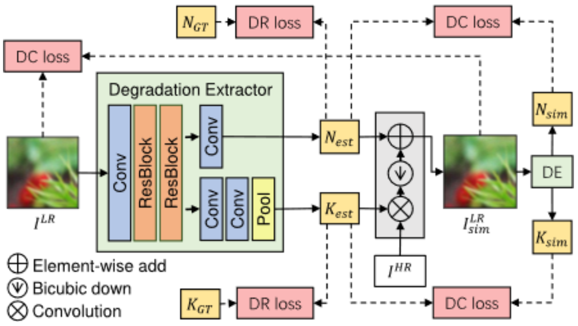


Fig.4. The core component of our proposed degradation-guided meta-restoration network for blind super-resolution [27].

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.

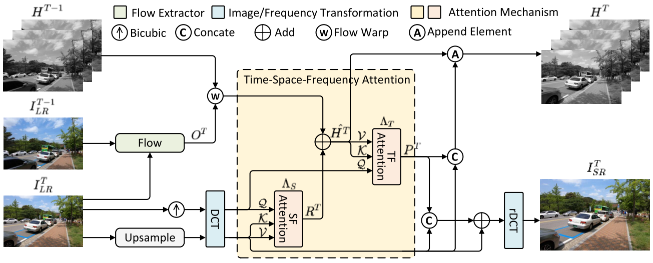


Fig. 5. An overview of our proposed frequency-transformer for compressed video super-resolution [18].

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.

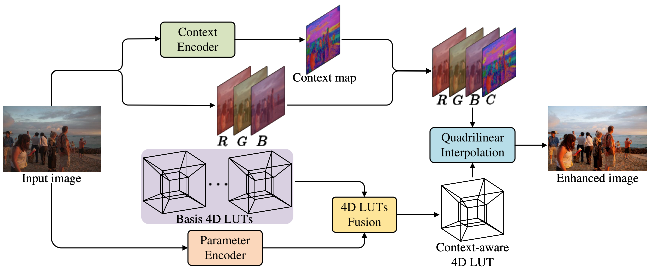


Fig. 6. A system overview of our proposed 4-dimensional lookup table for image enhancement [14].

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.