MDA-SR: Multi-level Domain Adaptation Super-Resolution for Wireless Capsule Endoscopy Images

Tianbao Liu[1,2], Zefeiyun Chen[1,2], Qingyuan Li[3], Yusi Wang[3,4], Ke Zhou[4], Weijie Xie[1,2], Yuxin Fang[3], Kaiyi Zheng[1,2], Zhanpeng Zhao[4], Side Liu[3,4(✉)], and Wei Yang[1,2(✉)]

[1] School of Biomedical Engineering, Southern Medical University, Guangzhou, China
[2] Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, China
weiyanggm@gmail.com
[3] Department of Gastroenterology, Nanfang Hospital, Southern Medical University, Guangzhou, China
liuside2011@163.com
[4] Guangzhou SiDe MedTech Company Ltd, Guangzhou, China

Abstract. Super-resolution (SR) of wireless capsule endoscopy (WCE) images is challenging because paired high-resolution (HR) images are not available. An intuitive solution is to simulate paired low-resolution (LR) WCE images from HR electronic endoscopy images for supervised learning. However, the SR model obtained by this method cannot be well adapted to real WCE images due to the large domain gap between electronic endoscopy images and WCE images. To address this issue, we propose a Multi-level Domain Adaptation SR model (MDA-SR) in an unsupervised manner using arbitrary set of WCE images and HR electronic endoscopy images. Our approach implements domain adaptation at the image level and latent level during the degradation and SR processes, respectively. To the best of our knowledge, this is the first work to explore an unsupervised SR approach for WCE images. Furthermore, we design an Endoscopy Image Quality Evaluator (EIQE) based on the reference-free image evaluation metric NIQE, which is more suitable for evaluating WCE image quality. Extensive experiments demonstrate that our MDA-SR method outperforms state-of-the-art SR methods both quantitatively and qualitatively.

Keywords: Wireless capsule endoscopy · Super-resolution · Domain adaptation

1 Introduction

Wireless capsule endoscopy (WCE) is an emerging examination technique that offers several advantages over traditional electronic endoscopy, including non-invasiveness, safety, and non-cross-infection. It enables the examination of the

T. Liu and Z. Chen—contributed equally to this work.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 518–527, 2023.
https://doi.org/10.1007/978-3-031-43907-0_50

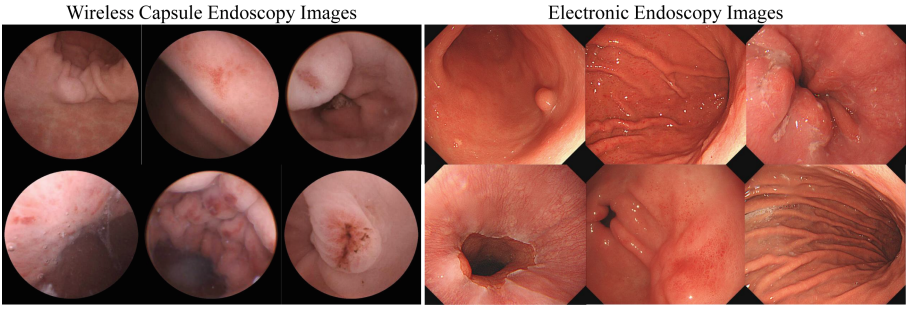


Fig. 1. The visual presentation of WCE images and electronic endoscopy images. The WCE images on the left are dim and blurred, while the electronic endoscopy images on the right are bright and clear.

entire human gastrointestinal tract and is widely used in clinical practice [15, 23]. Despite its successful clinical applications, the limited volume and data transmission bandwidth of WCE result in image quality drawbacks such as low resolution and poor quality [4]. Recent research has shown that high definition colonoscopy increases the identification of any polyps by 3.8% [19], and a 3-center prospective randomized trial has further proven the value of high resolution in invasive endoscopy [17]. Hence it is desirable to restore the image quality of WCE images via super-resolution techniques.

Most super-resolution methods based on deep learning are supervised by paired low-resolution (LR) and high-resolution (HR) images [8,21,26]. However, the corresponding HR WCE image is currently unavailable due to the hardware limitations in capsule size and transmission bandwidth. Alternatively, researchers have adopted predefined simple linear degradation assumptions (e.g., bicubic downsampling, gaussian downsampling) to feasibly synthesize corresponding LR samples from ground-truth HR images. Similarly, Almalioglu et al. [1] adopted this assumption to synthesis paired LR electronic endoscopy images from the HR ones for supervised super-resolution learning. However, this method is difficult to generalize to real WCE images due to the domain gap between WCE images and electronic endoscopy images.

What causes this domain gap? It might seem reasonable to adopt the simple linear degradation assumption by simply analogizing the domain gap between a mobile camera and a professional camera to a WCE and an electronic endoscope. However, what cannot be ignored is the different examination environment, where the WCE requires filling the stomach with water, while the electronic endoscopy inflates the stomach, which directly leads to the difference between the two image domains in terms of villi pose and speckle reflection, as shown in Fig. 1. This domain gap cannot be described by a simple linear degradation matrix.

Recently, many studies have utilized the CycleGAN [28] to combine explicit domain adaptation into SR. The basic idea is to generate LR versions of HR

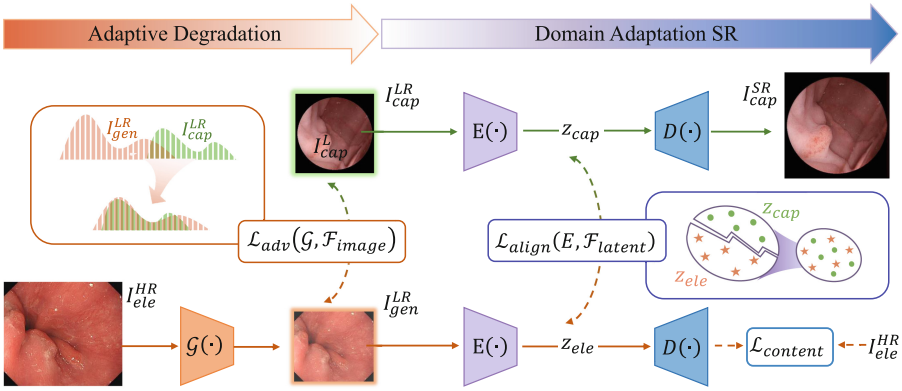


Fig. 2. Overview of the proposed MDA-SR, which consists of two parts: adaptive degradation and domian adaptation SR.

images with degenerate distributions similar to the real LR images, and then train the SR model on the generated LR-HR paired dataset [20,25,27]. The challenge is that the large domain gap between WCE and electronic endoscopic images makes the generator sensitive to learning shallow differences in content or style and unable to effectively bridge the domain gap.

In this work, we propose a Multi-level Domain Adaptation Super-Resolution (MDA-SR) for WCE images to bridge the domain gap between electronic endoscopy images and WCE images. MDA-SR leverages prior knowledge of HR electronic endoscopy images to guide the SR process of WCE images, as illustrated in Fig. 2. First, we train the adaptive degradation at the image level, employing a generative adversarial network to learn the complex and variable degradation distribution in WCE images, while incorporating an adaptive data loss [18] as the fidelity term of the image content. In contrast to previous methods that assume generated LR images are free from domain shift [3,9,25], we propose to minimize the domain gap during the SR process by aligning the latent feature distribution of electronic endoscopy images and WCE images. We further proposed EIQE for improving the reference-free image evaluation metric NIQE to be more suitable for endoscopy images. Through extensive experiments on real WCE images, we demonstrate the superiority of our method over other state-of-the-art SR methods, and its efficacy in reality.

2 Methods

2.1 Overview of the Proposed Method

Given a set of WCE images and HR electronic endoscopy images $\{I^{LR}_{cap}, I^{HR}_{ele}\}$, we aim to learn a SR function $\mathcal{R}(\cdot)$ that maps an observed I^{LR}_{cap} to its HR version according to the distribution defined by I^{HR}_{ele} in testing. As shown in Fig. 2,

the proposed scheme consists of two major parts: an adaptive degradation that generates the LR version of I_{ele}^{HR}, denoted as I_{gen}^{LR}, and a domain adaptation SR that aligns the latent features of the WCE and electronic endoscopy datasets during the SR process.

2.2 Adaptive Degradation

The purpose of the adaptive degradation is to obtain I_{gen}^{LR} with a degradation distribution similar to I_{cap}^{LR}. To achieve this, we employ the architecture of GAN [5], where the degradation generator $\mathcal{G}$ maps the generated I_{gen}^{LR} to I_{cap}^{LR} and the image-level discriminator $\mathcal{F}_{image}$ learns to distinguish between generated samples I_{gen}^{LR} and realistic samples I_{cap}^{LR} with adversarial loss:

$$\mathcal{L}_{adv}\left(\mathcal{G}, \mathcal{F}_{image}\right) = \mathcal{F}_{image}\left(\mathcal{G}\left(I_{ele}^{HR}\right)\right)^2 + \left[1 - \mathcal{F}_{image}\left(I_{cap}^{LR}\right)\right]^2 \tag{1}$$

$\mathcal{G}$ is optimized by maximizing the loss in Eq. (1) against an adversarial $\mathcal{F}_{image}$ that tries to minimize the loss.

A critical requirement is that the LR image generated by $\mathcal{G}$ should be consistent with the low-frequency information of the HR image. To enforce this constraint, we incorporate data loss as an additional supervision information.

The process of degradation from HR images to LR images is unknown. We adopt an adaptive downsampling kernel $\bar{k}$ [18] to approximate the unknown degradation process, since a widely-used approach uses predefined downsampling assumptions, such as bicubic downsampling or $s \times s$ average pooling [3]. It has been observed in previous works [2,6,12] that an appropriate downsampling function consists of low-filtering and decimation, we linearize the degradation generator $\mathcal{G}$ to a corresponding 2D kernel $\bar{k}$:

$$\bar{k} = \arg\min_{k} \sum_{i=1}^{N} \left\| \left(I_{ele,i}^{HR} * k\right)_{\downarrow s} - \mathcal{G}\left(I_{ele,i}^{HR}\right) \right\|_2^2 \tag{2}$$

where $I_{ele,i}^{HR}$ denotes an i-th example to estimate the kernel, N is the total number of samples that have been used and $\downarrow s$ represents downsampling operation with scale factor s. Finally, the data fidelity term $\mathcal{L}_{data}$ is defined as follows:

$$\mathcal{L}_{data}\left(\mathcal{G}\right) = \left\| \left(I_{ele}^{HR} * \bar{k}\right)_{\downarrow s} - \mathcal{G}\left(I_{ele}^{HR}\right) \right\|_1 \tag{3}$$

Given the definitions of adversarial and data losses above, the training loss of our adaptive degradation is defined as:

$$\mathcal{L}_{LR}\left(\mathcal{G}, \mathcal{F}_{image}\right) = \mathcal{L}_{adv}\left(\mathcal{G}, \mathcal{F}_{image}\right) + \lambda \mathcal{L}_{data}\left(\mathcal{G}\right) \tag{4}$$

where λ is a hyperparameter.

2.3 Domain Adaptation SR

The SR function $\mathcal{R}(\cdot)$ can then be supervised by the aligned image pair set $\{I_{gen}^{LR}, I_{ele}^{HR}\}$ obtained from adaptive degradation. As shown in Fig. 2, we use the pixel-wise content loss on the SR results $\mathcal{R}\left(I_{gen}^{LR}\right)$, which ensures the accuracy of HR image composition:

$$\mathcal{L}_{content} = \left\|\mathcal{R}\left(I_{gen}^{LR}\right) - I_{ele}^{HR}\right\|_1 \tag{5}$$

To further improve the performance of the WCE SR, it is crucial to bridge the domain gap between WCE images and electronic endoscopy images, even though the adaptive degradation generator $\mathcal{G}$ already learns the degradation distribution of WCE images through domain adaptation at the image level. To achieve this, we improve the domain adaptation at the latent level during the SR process.

A straightforward way is to adopt a GAN [5] structure to reduce the distribution shift. As shown in Fig. 2, the SR function $\mathcal{R}(\cdot)$ consists of an encoder E and a decoder D. We use two encoders E with shared weights to generate the electronic endoscopy latent feature z_{ele} as well as the WCE latent feature z_{cap}. We introduce latent-level discriminator $\mathcal{F}_{latent}$ to distinguish the domain for each latent feature, while the encoder E is trained to deceive $\mathcal{F}_{latent}$. The optimization of E and $\mathcal{F}_{latent}$ is achieved via the adversarial way, we use LSGAN [11] here:

$$\mathcal{L}_{align}(E) = \left(\mathcal{F}_{latent}\left(E\left(I_{cap}^{LR}\right) - 0.5\right)\right)^2 + \left(\mathcal{F}_{latent}\left(E\left(I_{gen}^{LR}\right)\right) - 0.5\right)^2 \tag{6}$$

$$\mathcal{L}_{align}(\mathcal{F}_{latent}) = \left(\mathcal{F}_{latent}\left(E\left(I_{cap}^{LR}\right) - 1\right)\right)^2 + \left(\mathcal{F}_{latent}\left(E\left(I_{gen}^{LR}\right)\right) - 0\right)^2 \tag{7}$$

As a result, the discriminator $\mathcal{F}_{latent}$ is trained with its corresponding loss in Eq. (7). The SR function $\mathcal{R}(\cdot)$ is trained with the following loss function:

$$\mathcal{L}_{SR} = \mathcal{L}_{content} + \mu\mathcal{L}_{align}(E) \tag{8}$$

where μ is a hyperparameter.

3 Experiments

3.1 Experiment Settings

Datasets. Our proposed model is trained on WCE image dataset and electronic endoscopy image dataset, and tested on WCE images. The WCE image dataset contains 14090 images, and the electronic endoscopy image dataset contains 2033 images, including 1302 images from the public Kvasir dataset [16] and 731 images from the local hospital. We perform a strict quality selection and remove the problematic images such as blurry, low-resolution and poor quality images.

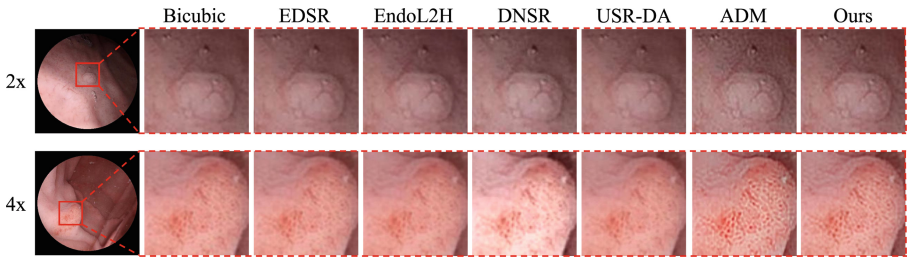


Fig. 3. The visual comparisons for 2x and 4x SR on the WCE images. Obviously, our result contains more natural details and textures suffering from less blur and artifacts.

Evaluation Metrics. In WCE SR problem, there is no corresponding ground-truth image used to calculate the evaluation metric. To address this issue, we design a no-reference Endoscopy Image Quality Evaluator (EIQE), which derives the quality-aware features from the Endoscopy Scene Statistic (ESS) model, inspired by the reference-free Natural Image Quality Evaluator (NIQE) [14]. The quality of a given test image is then expressed as the distance between a multivariate gaussian (MVG) fit of the ESS features extracted from the test image and a MVG model of the quality-aware features extracted from the corpus of HR electronic endoscopy images. Additionally, we use the no-reference metric BRISQUE [13] for evaluation purposes.

To better illustrate the subjective quality, we conduct a mean opinion score (MOS) test for comparison with other methods. We randomly select 100 different WCE images from the test set to subjectively evaluate the quality of the 2x and 4x WCE SR images. Four gastroenterology clinicians rate the visual perceptual qualities by assigning scores. Scores from 0 to 5 are used to indicate the qualities from low to high.

3.2 Training Details

Throughout the framework, the discriminators $\mathcal{F}_{image}$ and $\mathcal{F}_{latent}$ use the patch-based discriminator [7] with the instance normalization [22]. The generator $\mathcal{G}$, encoder E and decoder D in the model follow the residual block structure from the EDSR [8]. The parameters in Eq. (4) and Eq. (8) are set to be $\lambda = 1$ and $\mu = 0.001$, respectively. During training, we use the Adam optimizer and set the batch size and learning rate as 10 and 1×10^{-4}, respectively. To streamline the model training and reduce its complexity, we have divided the training process into two stages. In the first stage, we focus on training the adaptive degradation, which stabilizes the quality of generated LR images after 50 epochs. Following this, we incorporate the domain adaptation SR into the training process by training another 50 epochs. The experiments are implemented with Pytorch platform on NVIDIA GeForce RTX 2080 Ti.

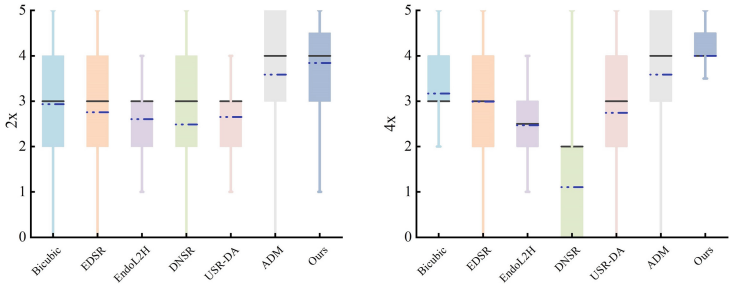


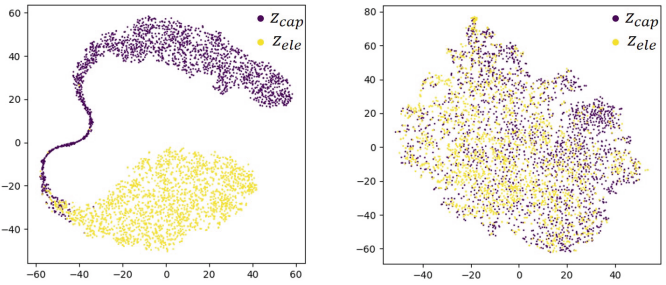
Fig. 4. Mean opinion of the subjective evaluation for different SR methods.

Table 1. Comparison results of the proposed MDA-SR model and other state-of-the-art methods on WCE image dataset, where DNSR, USR-DA and ADM incorporate domain adaptation but EDSR and EndoL2H do not. The best values are highlighted in bold font.

Method	2x		4x	
	EIQE↓	BRISQUE↓	EIQE↓	BRISQUE↓
Bicubic	5.65	58.03	7.36	76.75
EDSR [8]	6.06	59.06	6.81	75.89
EndoL2H [1]	6.63	57.42	6.66	69.92
DNSR [27]	5.50	58.16	6.71	74.72
USR-DA [24]	5.23	51.82	6.42	75.30
ADM [18]	5.19	52.62	6.21	**59.49**
MDA-SR(ours)	**5.14**	**50.70**	**6.16**	64.79

3.3 Results and Discussions

Comparison with Previous Methods. To validate the effectiveness of our proposed method, we compare it with existing state-of-the-art conventional SR methods without domain adaptation [1,8] and SR methods with domain adaptation [18,24,27]. Quantitative results are shown in Table 1, while visualization results are provided in Fig. 3. As EIQE is the most important metric in Table 1 and BRISQUE [13] is provided for reference since BRISQUE is based on natural scene statistic. As can be seen in Table 1, approaches that incorporated domain adaptation generally outperformed those that did not, thus highlighting the value of domain adaptation for the WCE SR problem. For the WCE image dataset, our MDA-SR method achieves the top EIQE performance.Note that in Fig. 3, the proposed method produces results containing clean and natural textures, while the result of ADM [18] are overly sharpened, producing unreal artifacts. The MOS results are shown in Fig. 4, indicating that our MDA-SR model produced the highest scores on average and with relatively smaller variance.



(a) w/o domain adaptation (b) w/ multi-level domain adaptation

Fig. 5. Visual representation of electronic endoscopy latent feature z_{ele} and WCE latent feature z_{cap} via t-SNE. z_{cap} in both (a) and (b) is obtained by encoding the I_{cap}^{LR}. z_{ele} in (a) is obtained by encoding the bicubic downsampled electronic endoscopy image I_{ele}^{LR}, while the z_{ele} in (b) is obtained by encoding the adaptively degraded electronic endoscopy image I_{gen}^{LR}

Table 2. Ablation study on our proposed method. We report four different levels of domain adaptation for the MDA-SR affect the SR results on WCE images.

Image Level	Latent Level	EIQE↓	
		2x	4x
		5.96	6.85
✓		5.76	6.35
	✓	5.48	6.26
✓	✓	**5.14**	**6.16**

Ablation Study. We use the t-SNE [10] for the visual representation of the image latent features distribution during SR process. It can be observed from Fig. 5(a) that there is a significant domain gap between WCE images and electronic endoscopy images, and from Fig. 5(b) show that our MDA-SR effectively bridges the domain gap. To better verify the effectiveness of our proposed model, we perform ablation experiments on WCE dataset. We obtain four different frameworks by removing different levels of domain adaptation structures. As shown in Table 2, domain adaptation at both the image and latent levels is effective. The worst framework is to train SR model on electronic endoscopy dataset and test it on WCE dataset. Our proposed MDA-SR model has the best SR effect on WCE images.

4 Conclusion

In this paper, we propose a multi-level domain adaptation SR for real WCE images. Our method first utilizes adaptive degradation to simulate the degradation distribution of WCE and generate LR electronic endoscopy images. We then

employ implicit domain adaptation at the latent level during the SR process to further bridge the domain gap between WCE images and electronic endoscopy images. Through extensive experiments on real WCE images, we demonstrate the superiority of our method over other state-of-the-art SR methods, and its efficacy in reality. Further evaluation for downstream tasks such as disease classification, region segmentation, or depth and pose estimation from the generated SR WCE images is warranted.

Acknowledgements. This research is supported by the Guangdong Province Key Field Research and Development Plan Project (2022B0303020003).

References

1. Almalioglu, Y., et al.: EndoL2H: deep super-resolution for capsule endoscopy. IEEE Trans. Med. Imaging **39**(12), 4297–4309 (2020)
2. Bell-Kligler, S., Shocher, A., Irani, M.: Blind super-resolution kernel estimation using an internal-GAN. In: Advances in Neural Information Processing Systems 32 (2019)
3. Bulat, A., Yang, J., Tzimiropoulos, G.: To learn image super-resolution, use a GAN to learn how to do image degradation first. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 187–202. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_12
4. Fante, K.A., Abdurahman, F., Gemeda, M.T.: An ingenious application-specific quality assessment methods for compressed wireless capsule endoscopy images. Trans. Environ. Electr. Eng. **4**(1), 18–24 (2020)
5. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
6. Gu, J., Lu, H., Zuo, W., Dong, C.: Blind super-resolution with iterative kernel correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1604–1613 (2019)
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
8. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144 (2017)
9. Lugmayr, A., Danelljan, M., Timofte, R.: Unsupervised learning for real-world super-resolution. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 3408–3416. IEEE (2019)
10. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11), 2579–2605 (2008)
11. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802 (2017)
12. Michaeli, T., Irani, M.: Nonparametric blind super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 945–952 (2013)
13. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: 2011 Conference Record of the Forty Fifth Asilomar Conference on SignAls, Systems And Computers (ASILOMAR), pp. 723–727. IEEE (2011)

14. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a completely blind image quality analyzer. IEEE Signal Process. Lett. **20**(3), 209–212 (2012)
15. Muhammad, K., Khan, S., Kumar, N., Del Ser, J., Mirjalili, S.: Vision-based personalized wireless capsule endoscopy for smart healthcare: taxonomy, literature review, opportunities and challenges. Futur. Gener. Comput. Syst. **113**, 266–280 (2020)
16. Pogorelov, K., et al.: KVASIR: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference, pp. 164–169 (2017)
17. Rex, D.K., et al.: High-definition colonoscopy versus Endocuff versus Endorings versus Full-spectrum Endoscopy for adenoma detection at colonoscopy: a multicenter randomized trial. Gastrointest. Endosc. **88**(2), 335–344 (2018)
18. Son, S., Kim, J., Lai, W.S., Yang, M.H., Lee, K.M.: Toward real-world super-resolution via adaptive downsampling models. IEEE Trans. Pattern Anal. Mach. Intell. **44**(11), 8657–8670 (2021)
19. Subramanian, V., Mannath, J., Hawkey, C., Ragunath, K.: High definition colonoscopy vs. standard video endoscopy for the detection of colonic polyps: a meta-analysis. Endoscopy **43**(06), 499–505 (2011)
20. Sun, W., Gong, D., Shi, Q., van den Hengel, A., Zhang, Y.: Learning to zoom-in via learning to zoom-out: real-world super-resolution by generating and adapting degradation. IEEE Trans. Image Process. **30**, 2947–2962 (2021)
21. Tong, T., Li, G., Liu, X., Gao, Q.: Image super-resolution using dense skip connections. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4799–4807 (2017)
22. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
23. Wang, A., et al.: Wireless capsule endoscopy. Gastrointest. Endosc. **78**(6), 805–815 (2013)
24. Wang, W., Zhang, H., Yuan, Z., Wang, C.: Unsupervised real-world super-resolution: a domain adaptation perspective. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4318–4327 (2021)
25. Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., Lin, L.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 701–710 (2018)
26. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301 (2018)
27. Zhao, T., Ren, W., Zhang, C., Ren, D., Hu, Q.: Unsupervised degradation learning for single image super-resolution. arXiv preprint arXiv:1812.04240 (2018)
28. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)