Original Article

Deep learning-based super-resolution method for projection image compression in radiotherapy

Zhixing Chang[#], Jiawen Shang[#], Yuhan Fan[#], Peng Huang[#], Zhihui Hu[#], Ke Zhang, Jianrong Dai, Hui Yan

Department of Radiation Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Contributions: (I) Conception and design: K Zhang, J Dai, H Yan; (II) Administrative support: K Zhang, J Dai, H Yan; (III) Provision of study materials or patients: P Huang, Z Hu; (IV) Collection and assembly of data: Z Chang, J Shang, Y Fan; (V) Data analysis and interpretation: Z Chang, J Shang, Y Fan; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

Correspondence to: Ke Zhang, PhD; Jianrong Dai, PhD; Hui Yan, PhD. Department of Radiation Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, No. 17, Panjiayuan Nanli, Chaoyang District, Beijing 100021, China. Email: zk110105@163.com; dai_jianrong@cicams.ac.cn; hui.yan@cicams.ac.cn.

Background: Cone-beam computed tomography (CBCT) is a three-dimensional (3D) imaging method designed for routine target verification of cancer patients during radiotherapy. The images are reconstructed from a sequence of projection images obtained by the on-board imager attached to a radiotherapy machine. CBCT images are usually stored in a health information system, but the projection images are mostly abandoned due to their massive volume. To store them economically, in this study, a deep learning (DL)-based super-resolution (SR) method for compressing the projection images was investigated.

Methods: In image compression, low-resolution (LR) images were down-sampled by a factor from the high-resolution (HR) projection images and then encoded to the video file. In image restoration, LR images were decoded from the video file and then up-sampled to HR projection images via the DL network. Three SR DL networks, convolutional neural network (CNN), residual network (ResNet), and generative adversarial network (GAN), were tested along with three video coding-decoding (CODEC) algorithms: Advanced Video Coding (AVC), High Efficiency Video Coding (HEVC), and AOMedia Video 1 (AV1). Based on the two databases of the natural and projection images, the performance of the SR networks and video codecs was evaluated with the compression ratio (CR), peak signal-to-noise ratio (PSNR), video quality metric (VQM), and structural similarity index measure (SSIM).

Results: The codec AV1 achieved the highest CR among the three codecs. The CRs of AV1 were 13.91, 42.08, 144.32, and 289.80 for the down-sampling factor (DSF) 0 (non-SR) 2, 4, and 6, respectively. The SR network, ResNet, achieved the best restoration accuracy among the three SR networks. Its PSNRs were 69.08, 41.60, 37.08, and 32.44 dB for the four DSFs, respectively; its VQMs were 0.06%, 3.65%, 6.95%, and 13.03% for the four DSFs, respectively; and its SSIMs were 0.9984, 0.9878, 0.9798, and 0.9518 for the four DSFs, respectively. As the DSF increased, the CR increased proportionally with the modest degradation of the restored images.

Conclusions: The application of the SR model can further improve the CR based on the current result achieved by the video encoders. This compression method is not only effective for the two-dimensional (2D) projection images, but also applicable to the 3D images used in radiotherapy.

Keywords: Cone-beam computed tomography (CBCT); radiotherapy; projection image; compression; super-resolution (SR)

Submitted Dec 27, 2024. Accepted for publication Jun 06, 2025. This article was updated on Dec 29, 2025.
The original version is available at: https://dx.doi.org/10.21037/qims-2024-2962
doi: 10.21037/qims-2024-2962

Introduction

Cone-beam computed tomography (CBCT) is a routine imaging procedure used for positioning and target localization of cancer patients during radiotherapy (1). In this procedure, a sequence of two-dimensional (2D) kilo-voltage (kV) projection images are acquired by an on-board imager (OBI) attached to the gantry of a linear accelerator. Based on the projection images acquired at a series of continuous scan angles, a three-dimensional (3D) tomographic image can be reconstructed and compared with the planning computed tomography (CT) to verify the target volume and position. Due to the advantage of the fast imaging, CBCT is popularly used for daily pre-treatment, intra-treatment, and post-treatment target verification in radiotherapy (2-4). A busy radiotherapy clinic generates a massive volume of projection and CBCT images. Storing them on the local computer could expend hard disk space quickly. It could lower system performance and even cause crashing of the computer in the worst case. To maintain the normal clinical operation, the daily acquired projection and CBCT images should be backed up properly in a remote storage space.

CBCT images are usually stored in the oncology information system (OIS) as they are frequently used by physician in evaluating the accuracy of patient setup and target localization. The projection images are mostly discarded as they are less frequently used in clinic. However, these projection images they contain precious information about patient anatomy, warranting long-term storage. Considering the massive volume of projection images, backing up them is more time-consuming and complicated than it is for CBCT images. For example, one set of projection data contains 360–660 images (512×512 or 1,024×768 in 32-bit float), whereas the corresponding CBCT data only contains 50–100 images (256×256 or 512×512 in 16-bit integer). The size of projection data is at least 10 times larger than that of CBCT data (5-7). In clinical practice, projection images are mostly saved in local computers for few weeks and then deleted. This is a waste of precious clinical and scientific research resource as it contains daily changes of patient anatomy information and geometric parameters which would be highly valuable for the follow-up studies. For instance, projection images provide critical raw data for advancing reconstruction algorithm research and motion modeling in radiation oncology. As demonstrated by Zhang *et al.* (8), sparse-view reconstruction algorithm development directly relies on analyzing geometric and density features embedded in projection images. Their work achieved high-fidelity 3D reconstruction from merely 45–90 projections through compressed sensing theory and weighted Schatten p-norm minimization, attaining superior metrics compared to conventional FDK and SART methods. This validates how projection images mining enables novel algorithm designs to overcome the Nyquist sampling limit, significantly reducing radiation dose while maintaining diagnostic accuracy. Complementing this, Sakurai *et al.* (9) extracted diaphragm waveforms from projection images via Amsterdam Shroud analysis, developing a respiratory motion prediction model that improved 3D tumor tracking accuracy. These studies collectively demonstrate that projection images serve as a versatile platform for both image reconstruction innovation and physiological motion characterization, addressing two fundamental challenges in precision radiotherapy.

The clinically relevant images are mostly stored in medical information systems, such as Picture Archiving Communication System (PACS) and Hospital Information Systems (HIS), which enable online access. The compression algorithm, such as JPEG and JPEG-2000, can be selected and applied to the images to be stored. These compression algorithms are roughly categorized in lossless and lossy classes (10). The lossless algorithms compress images without any loss of image detail (11,12). Compared to the lossless algorithms, the lossy algorithms compress the image to an even smaller size but with a certain loss of image detail (13). The issue of applying lossy versus lossless algorithms in compressing medical images has long been a topic of discussion. Several professional organizations have issued guidelines and standards for the use of compression in medical imaging applications. They have suggested the appropriate use of the lossy compression algorithms for a specific image modality provided that there was no significant compromise to the clinical goals (14).

Nowadays, the medical images are mostly stored statically, namely, each image is saved as an independent file (15). As medical images are mostly acquired in

spatial and temporal domains, it could be feasible to save their differences instead of their contents for the better compression performance. Analogous to a movie or animation, medical images are highly correlated temporally or spatially. It would be more effective to compress medical images using the existing video coding-decoding (codec) algorithms. In the pioneering studies, our group applied the existing video codecs to compress the medical images [CT, CBCT, and magnetic resonance imaging (MRI)] which were popularly used in radiotherapy clinics, with promising results. These current video codecs utilize inter-frame prediction and intra-frame coding techniques to maximally reduce the size of videos for media players. In addition, many advanced techniques including motion compensation, variable block-size prediction, and sophisticated entropy coding, have been implemented to boost their performance for the lower bitrate and higher image quality (16).

Compressing medical images using video codecs is more effective compared with the conventional compression algorithms. However, as CBCT is more popularly used in clinical practice due to its dose-efficient imaging (17), seamless integration with linear accelerators for real-time 3D verification (18), and adaptive radiotherapy compatibility through synthetic CT conversion (19), this widespread adoption has driven a rapid accumulation of projection data across radiotherapy workflows. The current compression ratio (CR) provided by video codecs might be not sufficient. Therefore, a more effective solution for compressing projection images is needed. As a straightforward option, the image can be down-sampled to the coarse image. This could result in the reduction of image size but cause the loss of image resolution to certain degrees. To restore the fine image from the coarse image, the up-sampling method is used to make up the lost detail. Super-resolution (SR) is a technique that generates high-resolution (HR) images from low-resolution (LR) images (20). Traditional methods, such as bicubic interpolation, estimate new pixel values based on neighboring pixels, offering smoother images but often lacking fine detail. Advanced non-learning approaches, including reconstruction-based iterative methods (21) and sparse representation techniques, have addressed structural preservation but remain constrained by manual prior design. In contrast, deep learning (DL)-based SR models, such as convolutional neural networks (CNNs) (22), residual network (ResNet), and generative adversarial networks (GANs) (23), learn complex mappings between LR and HR images, achieving unprecedented detail recovery. Recent advancements in medical imaging underscore the synergistic integration of DL and domain-specific regularization (24,25), enabling robust SR for clinical applications. Recent work by Umirzakova *et al.* (26) further demonstrates the feasibility of medical image SR through deep neural networks, where innovations such as residual-within-residual architecture and channel attention mechanisms effectively preserve diagnostically critical details while maintaining computational efficiency, solidifying the advantage of data-driven approaches in clinical settings.

In this study, DL networks were investigated to improve the result achieved by the current video codecs. To compress the projection images, LR images were down-sampled from the HR images and encoded to a video file. To restore the projection images, the LR images were decoded from the video file and up-sampled to HR images by the DL network. The principles of video codecs and DL models are introduced in the methods section. The performance of the video codes and DL models are analyzed in the results section. Finally, the advantages and limitations of the proposed method are discussed.

Methods

Data

The proposed method underwent initial validation using the Ultra Video Group (UVG) dataset, a standardized benchmark for video processing. This dataset contains high-quality video sequences, typically in 4K resolution, and is designed to evaluate the efficiency and quality of video codecs. Each sequence in the UVG dataset is carefully selected to represent a variety of motion complexities, textures, and colors, making it a valuable resource for researchers and developers working on video coding and compression technologies (27). Although UVG contains natural video sequences rather than medical content, its precisely characterized temporal coherence and well-defined motion patterns provide an essentially controlled environment. This design specifically verifies our architecture's core capability in spatiotemporal feature learning and texture-preserving reconstruction—foundational competencies required for processing temporal medical projections. Our two-stage validation strategy first establishes baseline performance on natural videos before addressing medical imaging challenges, intentionally distinguishing architectural effectiveness from domain adaptation benefits. The UVG evaluation quantitatively

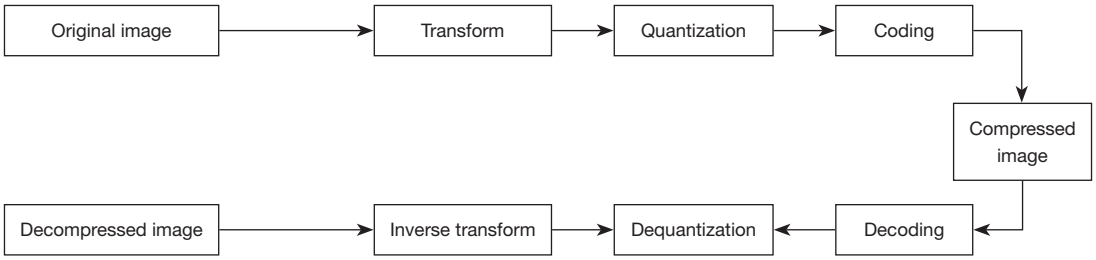

Figure 1 The workflow of intra-frame coding algorithm for image compression.

confirmed critical technical capabilities, including sub-frame motion estimation accuracy and multi-scale detail recovery, which proved directly transferable to preserving subtle anatomical textures in subsequent sparse-view CBCT projection processing. This methodological approach maintains continuity with established video restoration research while ensuring clinical improvements stem from core architectural merits rather than medical data overfitting. The data tested in this study is the Bosphorus dataset, which has a duration of 5 seconds, a resolution of 4,096×2,160 pixels, a frame rate of 120 fps (progressive), a subsampling format of 4:4:4, a bit depth of 16 bits, and a data format of RAW. To fit the SR model, the resolution was downloaded to 1,920×1,080, 8-bit, YUV, and RAW format. The dataset has been converted into 600 images. In consistency with the size of projection images, these images were cropped to the size of 512×512.

The proposed method was also evaluated on a clinical projection image dataset. These projection images were collected from patients under image-guided radiotherapy with treatment sites on the head and abdomen. For each patient, there are about 10–20 CBCT scans acquired over a period of 3–5 weeks. In each CBCT scan, 360 projection images are obtained during gantry rotation with 1° angle spacing. The CBCT scan was performed on the X-ray Volumetric Imaging (XVI) system which was attached to an Elekta VersaHD linear accelerator (Elekta, Stockholm, Sweden). The source to isocenter distance was 1,000 mm, whereas the isocenter to imager distance was 600 mm. The size of a projection image was 512×512 with a pixel size of 0.83 mm in isocenter plane.

This study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments. The Ethics Committee of National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College approved this study (No. NCC2018-016). The requirement for written informed consent was waived because of the retrospective design.

Video codecs

The intra-frame coding algorithms were mostly employed in compressing static images (the workflow is shown in *Figure 1*). This process begins by partitioning the input image into non-overlapping blocks. Each block then undergoes a frequency-domain transform—typically the discrete cosine transform (DCT) or, in some codecs, the discrete wavelet transform (DWT). This transform produces 2D blocks of coefficients representing the block's spatial frequency content. To achieve compression, these coefficients undergo quantization: small-magnitude coefficients are selectively discarded, and the precision of remaining coefficients is reduced, introducing controlled loss. The quantized coefficients are then commonly scanned, run-length encoded, and entropy-encoded to generate a compressed bitstream. Decompression reverses this sequence: the bitstream is entropy-decoded, inverse run-length processed, and de-quantized to reconstruct the coefficient blocks. Each block of the original image is finally restored by applying the corresponding inverse transform (inverse DCT or inverse DWT) to the coefficient blocks.

Motion images, such as video and animation, exhibit significant temporal redundancy due to feature similarity between consecutive frames. Algorithms exploiting this temporal redundancy, known as inter-frame coding, are detailed in the workflow shown in *Figure 2*. For each block within the current image, the encoder first performs motion estimation against one or more reference images to determine the optimal matching block. This search process yields motion vectors encoding the relative displacement between the current block and its best match. These motion vectors are then used to generate a predicted image through motion compensation. The encoder subsequently calculates the difference image by subtracting this predicted

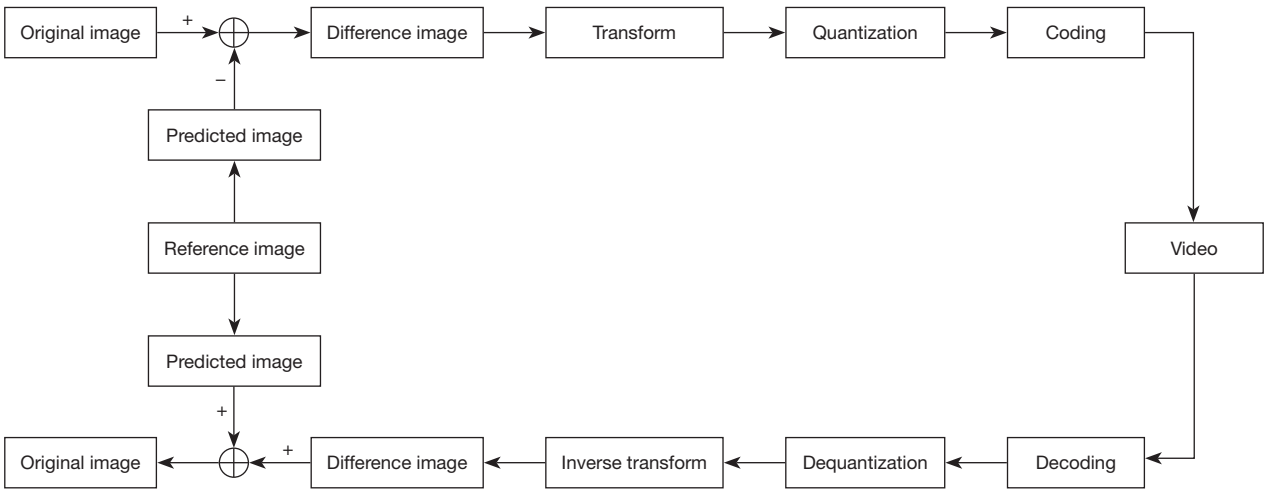

Figure 2 The workflow of inter-frame coding algorithm for image compression. The nodes annotated with '+' and '–' symbols perform addition and subtraction operations, respectively.

image from the original image. Typically exhibiting a sparse distribution, this difference image next undergoes the same core compression process used in intra-frame coding: frequency-domain transformation, quantization, scanning, run-length encoding, and entropy encoding. The resulting compressed bitstream contains the encoded difference data alongside the motion vectors. Decompression commences with entropy decoding of the bitstream to retrieve the motion vectors and the quantized difference coefficients. The difference coefficients are de-quantized and inverse transformed to reconstruct the difference image. Concurrently, motion vectors drive motion compensation applied to the reference image to regenerate the predicted image. The target image is ultimately recovered by adding the reconstructed difference image to this predicted image. Since storage requirements are limited to motion vectors and compressed difference data instead of complete images, inter-frame coding achieves substantially higher CRs compared to intra-frame coding methods.

There are many popular video codecs including Advanced Video Coding (AVC) (28), High Efficiency Video Coding (HEVC) (29), and AOMedia Video 1 (AV1) (30). AVC is also known as H.264 and the most popular video compression standard. HEVC is also known as H.265 and the successor of AVC/H.264. AV1 is an open video codec designed to provide high-quality video compression with greater efficiency than previous codecs. It enables high-quality video while reducing streaming and storage costs, offering significant reductions in file sizes without compromising quality. The detailed descriptions of these three video codecs are supplied in Appendix 1.

SR

The classic SR methods are bicubic and bilinear interpolation, which estimate new pixel values based on neighboring pixels (31). In contrast, DL-based SR methods can achieve better results by learning complex mappings from LR images to HR images through extensive training on the image databases (32). Three DL models, CNN, ResNet, and GAN, were investigated in this study.

SR-CNN adopts a lightweight three-layer convolutional structure for rapid medical projection image enhancement. The network sequentially processes inputs through a 9×9 convolutional feature extractor, followed by dual 5×5 nonlinear mapping layers with rectified linear unit (ReLU) activation, and concludes with a 5×5 reconstruction layer. Medical projection images from our custom dataset were divided into training, validation, and test set (8:1:1) through patient ID stratification to prevent data leakage. The model was trained for 100 epochs using an Adam optimizer with an initial learning rate of 1e–4, mean square error (MSE) loss, and a batch size of 8, augmented by random 90°, 180°, and 270° rotations to enhance generalization.

SR-ResNet leverages 16 residual blocks with integrated batch normalization and PixelShuffle upsampling to recover fine anatomical textures. Our proprietary medical dataset was partitioned into training, validation, and test sets using

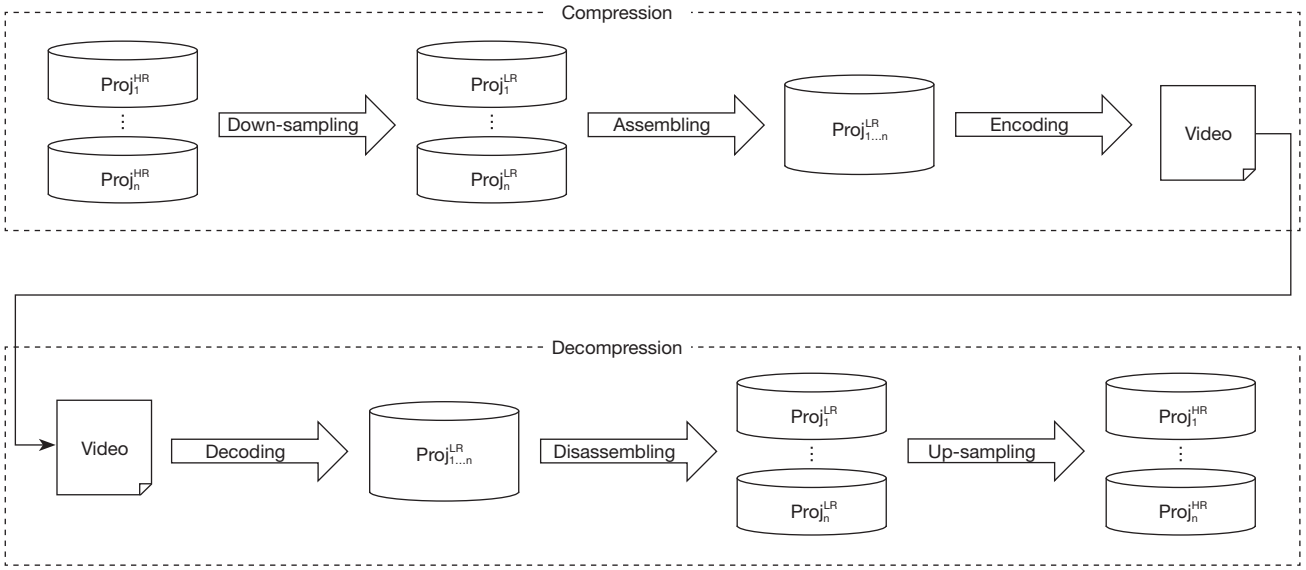

Figure 3 The workflow of image compression and decompression with the SR method. HR, high-resolution; LR, low-resolution; SR, super-resolution.

an 8:1:1 patient-wise split, ensuring all images from the same patient remained within a single subset. The training ran for 100 epochs per scale factor with a learning rate decay strategy where the initial rate of 1e–4 was halved every 50 epochs. The MSE loss function optimized pixel-level accuracy across multi-scale reconstructions.

SR-GAN combines residual blocks and adversarial training to generate clinically plausible enhancements. The generator employs 16 residual blocks with PixelShuffle upsampling, whereas the discriminator utilizes stride-2 convolutions layered with LayerNorm and LeakyReLU activation. Following the same 8:1:1 patient-wise data split strategy as SR-CNN and SR-ResNet, SR-GAN was trained on separated medical projections using a hybrid loss function comprising adversarial loss weighted by 1e–3 and perceptual loss based on VGG19 features weighted by 1e–4. To stabilize adversarial training, we implemented a label smoothing technique where real and fake labels were assigned values of 0.9999 and 0.0001, respectively. The model underwent 100 epochs of training with a reduced batch size of 8, optimized by Adam at a fixed learning rate of 1e–4 for both generator and discriminator.

The details of the network architectures and training schemes are supplied in Appendix 2. All networks were trained on an NVIDIA GeForce RTX 4080 SUPER (NVIDIA, Santa Clara, CA, USA) using a random sample of thousands of images from the dataset.

Evaluation

The workflow of image compression and decompression procedures is shown in *Figure 3*. The original HR projection images were first down-sampled to the LR images. Then, all LR projection images of one set were connected to those of the next set. Later, all LR projection images in all sets were assembled to an image sequence and encoded to a video file. The CR was calculated between the sizes of the image sequence and the video file. To restore images, the video file was first decoded to the LR projection image sequence via video decoder. Then, the LR projection image sequence was disassembled to their respective scan sets. Later, the HR projection images were up-sampled by the SR models. The restored HR projection images were compared to the original projection images to evaluate the restoration accuracy of the decoding algorithm. Three popular metrics, peak signal-to-noise ratio (PSNR) (33), video quality metric (VQM) (34), and structural similarity index measure (SSIM) (35), were employed in measuring the similarity between two images.

Our framework was systematically evaluated under four distinct compress modes: down-sampling factor (DSF) =0 (non-SR) and DSF =2, 4, and 6. Crucially, the DSF =0 condition represents a codec-only baseline where test images were directly processed through the video codecs without downsampling and upscaling. When DSF>1,

Table 1 The performance of the SR method on the natural images

Video codec	DSF	SR-CNN				SR-ResNet				SR-GAN			
		CR	PSNR (dB)	VQM (%)	SSIM	CR	PSNR (dB)	VQM (%)	SSIM	CR	PSNR (dB)	VQM (%)	SSIM
AVC	0	5.11	41.62	3.64	0.9138	5.11	41.62	3.64	0.9138	5.11	41.62	3.64	0.9138
	2	16.97	27.38	24.24	0.8071	16.97	29.21	19.56	0.8553	16.97	28.50	21.29	0.8347
	4	65.81	25.14	30.92	0.7233	65.81	26.97	25.38	0.7568	65.81	26.03	28.15	0.7412
	6	140.4	24.52	32.94	0.7014	140.4	25.17	30.83	0.7241	140.4	25.16	30.86	0.7198
HEVC	0	6.63	43.11	2.93	0.9226	6.63	43.11	2.93	0.9226	6.63	43.11	2.93	0.9226
	2	22.28	28.20	22.05	0.8413	22.28	30.79	16.09	0.8651	22.28	29.24	19.49	0.8547
	4	86.21	26.86	25.7	0.7362	86.21	28.43	21.46	0.7648	86.21	27.92	22.78	0.7488
	6	163.50	25.84	28.73	0.7196	163.50	26.57	26.54	0.7310	163.50	27.06	25.13	0.7305
AV1	0	7.34	44.56	2.37	0.9361	7.34†	44.56†	2.37†	0.9361†	7.34	44.56	2.37	0.9361
	2	25.17	28.94	20.2	0.8588	25.17†	31.42†	14.86†	0.8955†	25.17	29.99	17.78	0.8821
	4	97.94	27.14	24.9	0.7541	97.94†	29.47†	18.95†	0.7848†	97.94	28.73	20.71	0.7628
	6	173.88	26.05	28.09	0.7274	173.88	27.88	22.89	0.7483†	173.88†	28.22†	22†	0.7397

†, best-performing data. AV1, AOMedia Video 1; AVC, Advanced Video Coding; CR, compression ratio; DSF, down-sampling factor; HEVC, High Efficiency Video Coding; PSNR, peak signal-to-noise ratio; SR, super-resolution; SR-CNN, super-resolution convolutional neural network; SR-GAN, super-resolution generative adversarial network; SSIM, structural similarity index measure; VQM, video quality metric.

images underwent: (I) bicubic downsampling; (II) video codecs compression; (III) decoding; and (IV) neural network-based recovery that jointly addresses compression artifacts and reconstructs high-frequency details through our proposed SR method. The routines for data processing were developed using Python programming language (https://www.python.org/). An open-source audio and a video converter tool, ffmpeg (https://ffmpeg.org/), was used as API for image compression. The default configuration was selected for AV1 to maximize video compression. The constant rate factor (CRF) was set to 5 for AVC and HEVC and CRF was set to 1 for AV1. CRF is the quality control setting for the encoders. Lower values would result in better quality at the expense of higher file sizes.

Results

Natural image database

The performance of the proposed method on the UVG database with the different combination of video codecs, DSF, and SR models is summarized in *Table 1*. It shows that for the larger DSF, the CR was increased but the restored image quality was decreased. The combination of SR-ResNet and AV1 had the best performance in CR, PSNR, and SSIM among all 9 combinations. The visual effects of the restored and original images are also demonstrated in *Figure 4*. It shows that when DSF =2, the restored images were comparable to the original images. When DSF =4 and 6, the restored images were less clear than the original images.

Projection image database

The performance of the proposed method on the projection images at different scan sites (head and abdomen) are summarized in *Tables 2,3*. When DSF =2, the reconstruction retained the most of detailed image information. However, as DSF increased, the restored image quality gradually degraded. When DSF =6, the image was blurred severely. The combination of SR-ResNet and AV1 had the best performance in CR, PSNR, and SSIM among all 9 combinations.

The visual effects of the restored and original images are also demonstrated in *Figures 5,6*. It can be seen that SR-ResNet achieved the highest PSNR and SSIM in most tests, but the images produced by SR-GAN had better visual effect. This is because SR-ResNet focuses on pixel-wise fidelity, optimizing for higher PSNR and SSIM by

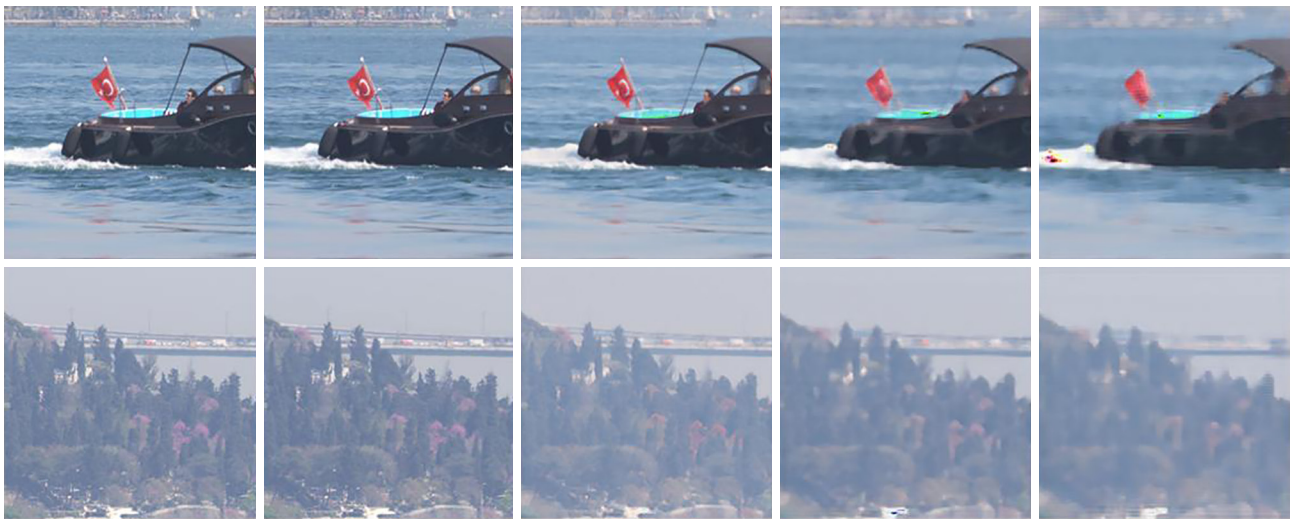

Figure 4 The visual comparison between the original and restored images with the different DSFs. The top line is the image with boat on the sea and the bottom line is the image with forest at the river. DSF, down-sampling factor; SR, super-resolution.

Table 2 The performance of the SR method on the clinical head projection images

Video codec	DSF	SR-CNN				SR-ResNet				SR-GAN			
		CR	PSNR (dB)	VQM (%)	SSIM	CR	PSNR (dB)	VQM (%)	SSIM	CR	PSNR (dB)	VQM (%)	SSIM
AVC	0	10.52	62.83	0.16	0.9940	10.52	62.83	0.16	0.9940	10.52	62.83	0.16	0.9940
	2	35.03	28.93	20.23	0.8468	35.03	32.56	12.82	0.8632	35.03	30.36	16.98	0.8501
	4	129.76	26.17	27.72	0.8255	129.76	29.67	18.49	0.8429	129.76	28.75	20.67	0.8347
	6	256.68	24.74	32.22	0.7958	256.68	26.40	27.04	0.8211	256.68	25.98	28.3	0.8036
HEVC	0	14.31	64.53	0.12	0.9941	14.31	64.53	0.12	0.9941	14.31	64.53	0.12	0.9941
	2	40.93	31.09	15.5	0.8520	40.93	34.17	10.36	0.8836	40.93	33.83	10.84	0.8693
	4	141.74	28.32	21.74	0.8317	141.74	31.48	14.75	0.8674	141.74	30.10	17.54	0.8552
	6	265.12	26.49	26.77	0.8104	265.12	28.03	22.49	0.8365	265.12	27.61	23.61	0.8249
AV1	0	17.07	65.82	0.1	0.9944	17.07†	65.82†	0.1†	0.9944†	17.07	65.82	0.1	0.9944
	2	47.88	32.70	12.59	0.8635	47.88†	36.78†	7.24†	0.8923†	47.88	34.95	9.32	0.8861
	4	152.48	29.63	18.58	0.8468	152.48†	32.80†	12.43†	0.8745†	152.48	30.62	16.44	0.8631
	6	279.36	27.17	24.82	0.8310	279.36	28.61	21.01	0.8513†	279.36†	29.59†	18.68†	0.8468

†, best-performing data. AV1, AOMedia Video 1; AVC, Advanced Video Coding; CR, compression ratio; DSF, down-sampling factor; HEVC, High Efficiency Video Coding; PSNR, peak signal-to-noise ratio; SR, super-resolution; SR-CNN, super-resolution convolutional neural network; SR-GAN, super-resolution generative adversarial network; SSIM, structural similarity index measure; VQM, video quality metric.

producing smoother images with minimal pixel-level discrepancies. In contrast, SR-GAN prioritizes perceptual quality, generating sharper textures and more realistic details, which may introduce subtle pixel variations. As a result, SR-GAN tends to have relatively smaller PSNR and SSIM but provides superior visual quality. Additionally, the results in *Tables 1-3* indicate that the quality of the restored images at the abdomen is higher than that of the

Table 3 The performance of the SR method on the clinical abdomen projection images

VIDEO CODEC	DSF	SR-CNN				SR-ResNet				SR-GAN			
		CR	PSNR (dB)	VQM (%)	SSIM	CR	PSNR (dB)	VQM (%)	SSIM	CR	PSNR (dB)	VQM (%)	SSIM
AVC	0	7.85	67.50	0.08	0.9923	7.85	67.50	0.08	0.9923	7.85	67.50	0.08	0.9923
	2	26.39	32.05	13.7	0.9313	26.39	31.56	14.59	0.9558	26.39	30.62	16.44	0.9432
	4	99.98	27.69	23.39	0.9147	99.98	30.04	17.67	0.9237	99.98	29.35	19.23	0.9219
	6	208.22	25.11	31.02	0.9085	208.22	28.91	20.27	0.9136	208.22	28.42	21.49	0.9104
HEVC	0	10.74	68.06	0.07	0.9981	10.74	68.06	0.07	0.9981	10.74	68.06	0.07	0.9981
	2	33.76	34.29	10.19	0.9415	33.76	40.54	4.25	0.9713	33.76	37.66	6.41	0.9602
	4	119.61	29.26	19.44	0.9269	119.61	38.63	5.59	0.9674	119.61	34.14	10.4	0.9380
	6	233.42	27.32	24.4	0.9152	233.42	31.69	14.35	0.9608	233.42	32.93	12.21	0.9295
AV1	0	13.91	69.08	0.06	0.9984	13.91†	69.08†	0.06†	0.9984†	13.91	69.08	0.06	0.9984
	2	42.08	35.36	8.81	0.9535	42.08†	41.60†	3.65†	0.9878†	42.08	38.92	5.36	0.9646
	4	144.32	32.43	13.04	0.9378	144.32†	37.08†	6.95†	0.9798†	144.32	35.62	8.5	0.9415
	6	289.80	29.78	18.25	0.9243	289.80	32.44	13.03	0.9518†	289.80†	33.28†	11.66†	0.9377

†, best-performing data. AV1, AOMedia Video 1; AVC, Advanced Video Coding; CR, compression ratio; DSF, down-sampling factor; HEVC, High Efficiency Video Coding; PSNR, peak signal-to-noise ratio; SR, super-resolution; SR-CNN, super-resolution convolutional neural network; SR-GAN, super-resolution generative adversarial network; SSIM, structural similarity index measure; VQM, video quality metric.

restored images at the head, and the compression rate is higher for the head projection images. The compression of the clinical projection images is higher than that of the natural images. *Tables 2,3* also show that when DSF <4, the VQM values obtained using the proposed method were mostly below 20%, which corresponds to an 'excellent' perceptual quality. This further validates the feasibility and superior performance of the method, demonstrating its effectiveness in maintaining high image quality under lower downsampling factors.

Figure 7 presents the performance comparison of different SR networks (SR-CNN, SR-ResNet, and SR-GAN) across varying DSF. As shown in the figure, image quality, as measured by PSNR, SSIM, and VQM, degraded with increasing DSF. This figure clearly demonstrates the contrasting behaviors of the different super-resolution networks in restoring image quality at various DSF levels. Notably, SR-ResNet and SR-GAN consistently outperformed SR-CNN, particularly at higher DSFs, indicating their superior capability in handling higher compression rates.

Video codecs

Figures 8,9 show the compression rates achieved by the three video encoders with respect to different DSF as a function of PSNR and SSIM. When all three video codecs had the same PSNR or SSIM, the AV1 displayed the best compression performance, followed by HEVC. The CR of AVC was the lowest. Ideally, with down-sampling operation, the CR should be increased by a factor equal to the square of DSF. However, as the image is down-sampled, the CR on the LR images could be compromised. Therefore, the ultimate CR under the down-sampling and video encoding operations would be less than the product of their respective contributions. For example, the CR with DSF =4 should be 4 times of the CR with DSF =2. Yet, as shown in *Figure 8*, the CR with DSF =4 is 100 whereas the CR with DSF =2 is 30. It only has about 3 folds instead of 4 folds.

Discussion

The common compression algorithms, such as JPEG, can only provide a modest reduction of image size. This is inadequate for the quickly increasing volume of on-board images generated in daily treatment of radiotherapy. To enable the storage of more images in a given hard disk space, a more effective compression algorithm should be developed. In this study, the original image was pre-processed by the down-sampling operation then exported

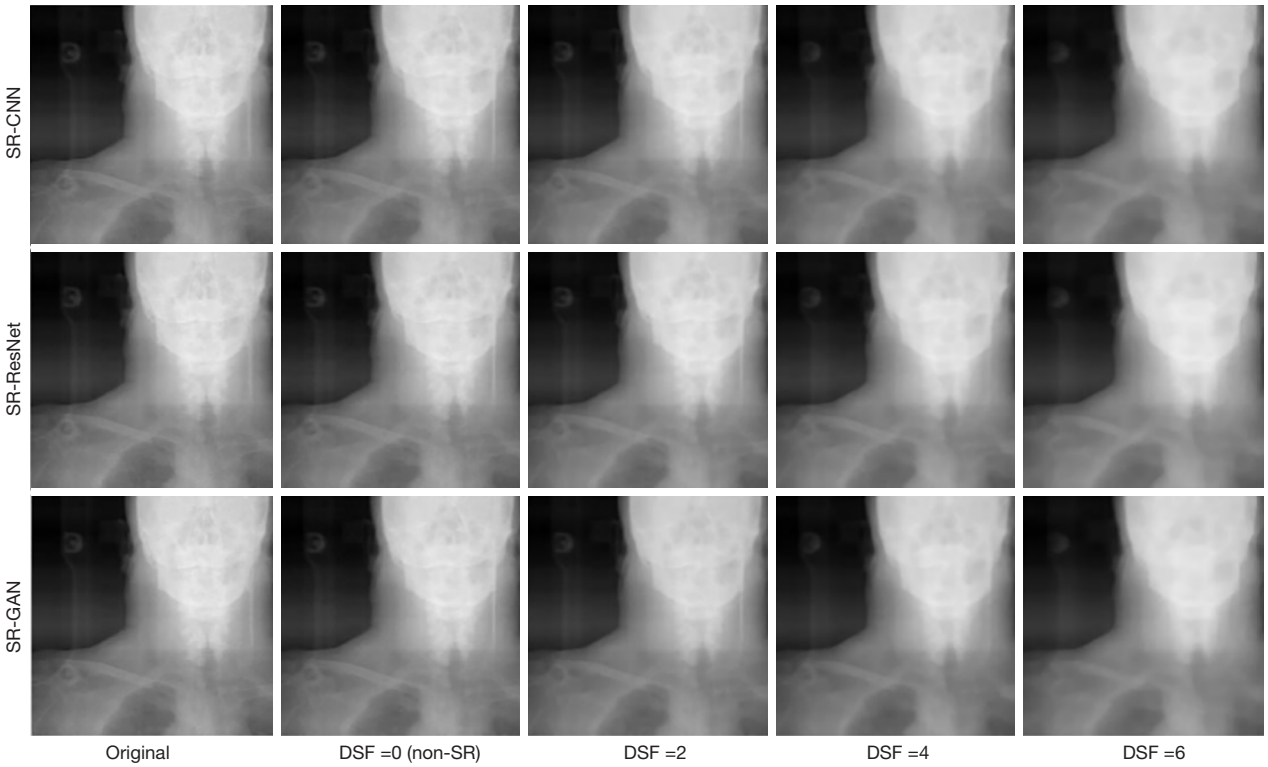

Figure 5 The visual comparison between the original and restored images for head projection image with the different DSFs. CNN, convolutional neural network; DSFs, down-sampling factors; GAN, generative adversarial network; SR, super-resolution.

to a video encoder for further compression. Based on the current result, the CR can be multiplied by a factor of 2^2–4^2 relative to that of the existing compression algorithm. This is encouraging as more images can be stored for long-term without visible loss of image quality.

The performance of SR models on the projection image database was better than that on the natural image database. The reason could be that the projection images are monochrome images that contain only luminance channel or intensity channel. The natural images are polychrome images that contain three color channels (red, green, and blue) in addition to luminance channel. The monochrome images have less information than the polychrome images. This reduced the data amount and complexity in training those DL networks. With the simplified data content, the network can focus more on the image details without the additional influence of color channels. In natural images, the network must manage both structural and color restoration, making it more difficult to achieve high accuracy across both factors. Furthermore, monochrome images are less prone to color-related noise and inconsistencies, resulting in more stable training and faster convergence.

The performance of the SR model was affected by the image content. As shown in *Figure 6*, the restored projection images had higher PSNR and SSIM. This is because the abdomen has fewer delicate structures and more smooth regions, which make the SR model easier to learn and predict. In contrast, projection images at the head have higher compression rates but relatively low restoration accuracy, as shown in *Figure 5*. This may be due to the head having more delicate structures in spatial and temporal domains, which allows encoders to compress more redundant content. Also, projection images showed higher compression rates compared to that of natural images. This is because projection images have fewer color channels than natural images.

The CR of AVC is less than that of HEVC, whereas the PSNR and SSIM of AVC are more than those of HEVC, as shown in *Figures 8*,*9*. This is because HEVC uses a flexible Coding Tree Unit (CTU) structure up to 64×64 pixels, compared to AVC's fixed 16×16 macroblocks, which allows more efficient handling of complex image regions and

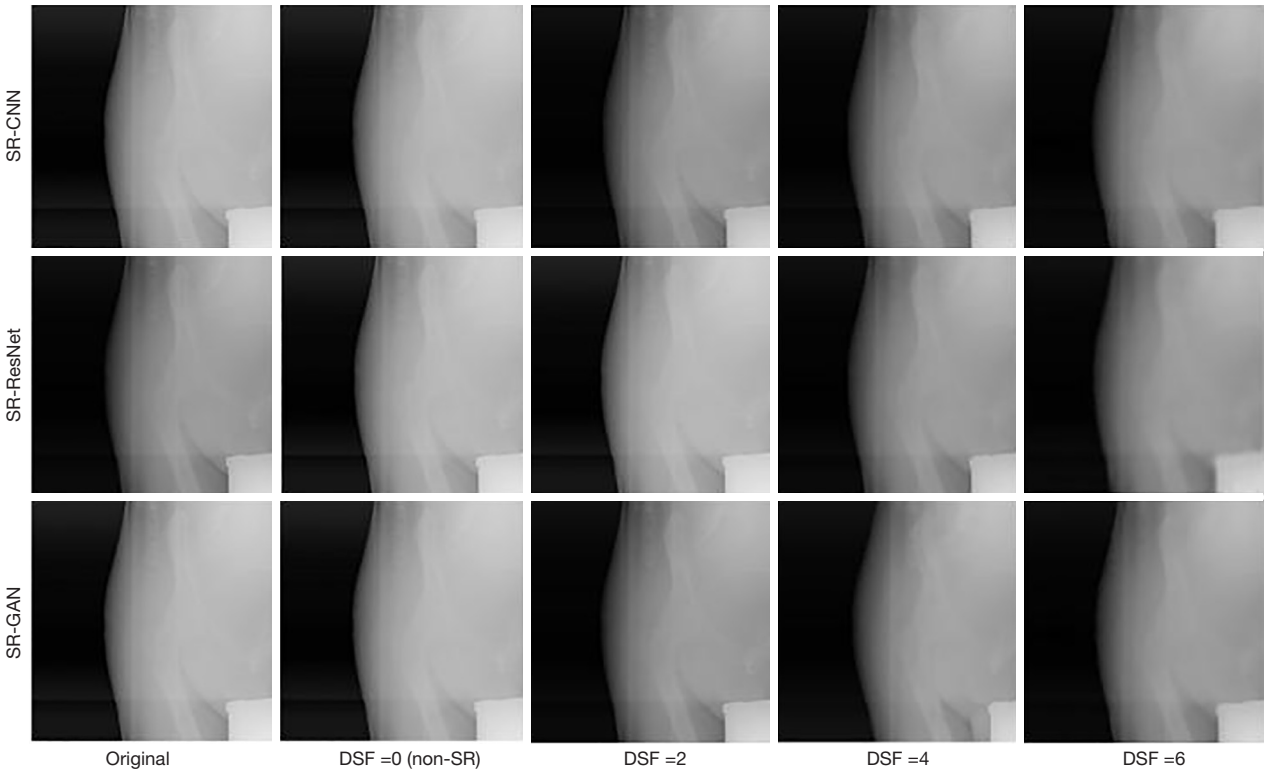

Figure 6 The visual comparison between the original and restored images for abdomen projection image with the different DSFs. CNN, convolutional neural network; DSFs, down-sampling factors; GAN, generative adversarial network; SR, super-resolution.

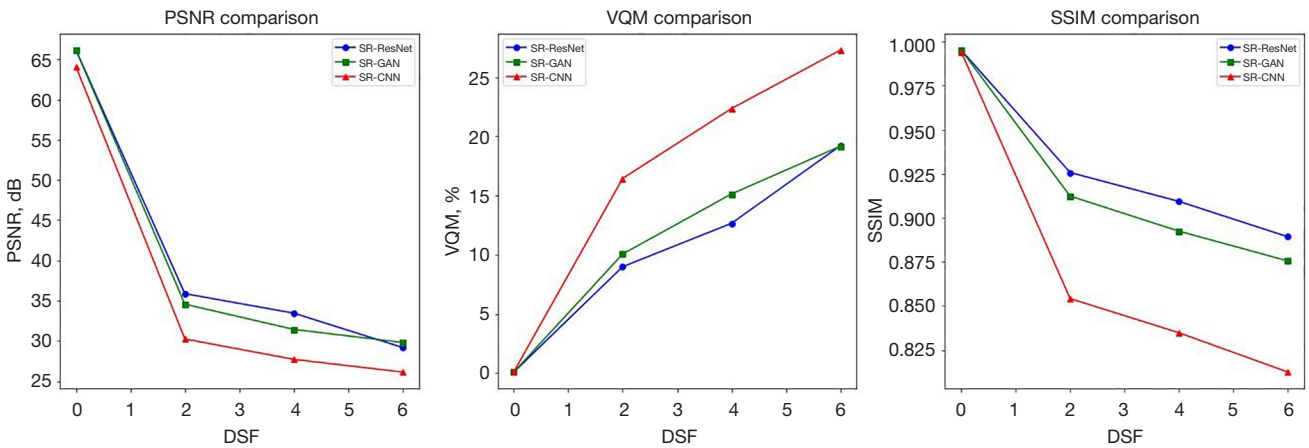

Figure 7 Comparison of SR networks in PSNR, VQM, and SSIM at different DSFs. CNN, convolutional neural network; DSFs, down-sampling factors; GAN, generative adversarial network; PSNR, peak signal-to-noise ratio; SR, super-resolution; SSIM, structural similarity index measure; VQM, Video Quality Metric.

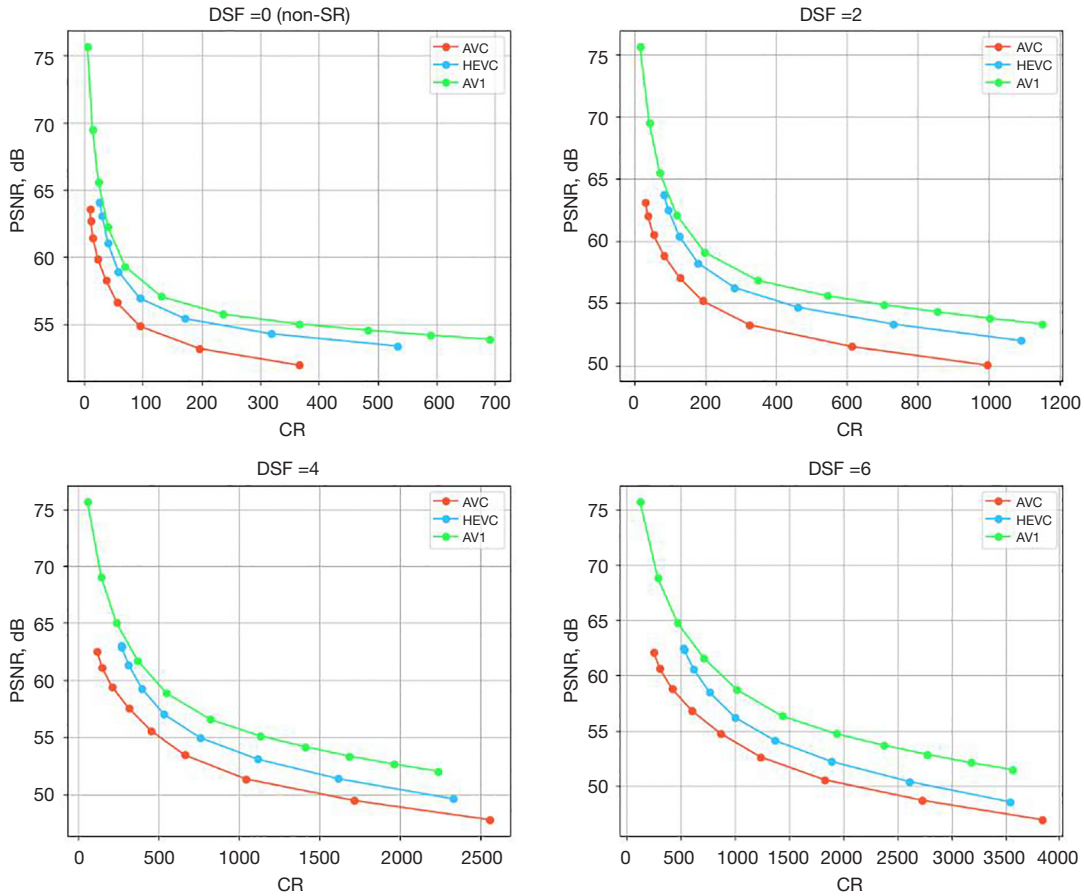

Figure 8 Relationship between CR and PSNR of three video encoders with respect to four DSFs. AV1, AOMedia Video 1; AVC, Advanced Video Coding; CR, compression ratio; DSFs, down-sampling factors; HEVC, High Efficiency Video Coding; PSNR, peak signal-to-noise ratio; SR, super-resolution.

reduces redundancy. In HEVC, the prediction was more accurate with the finer intra-frame motion compensation and the data compression was more effective with the enhanced CABAC entropy coding. In addition, the Sample Adaptive Offset (SAO) filter can further refine edge quality of image. These techniques help HEVC to maintain better restoration accuracy, which shows higher PSNR than that of AVC.

Among the three video codecs, AV1 achieved the best performance in both CR and restoration accuracy (PSNR and SSIM). This is attributed to its advanced prediction techniques, including improved intra-frame and inter-frame prediction, constrained directional enhancement filters (CDEF), and loop restoration. These technical advantages position AV1 as the most balanced codec among modern codecs, with high compression efficiency and restoration accuracy. AV1's architectural superiority over AVC/HEVC arises from its holistic re-engineering of compression paradigms. The codec fundamentally enhances prediction accuracy through directional granularity in intra-coding—its 56 angular modes with ±3° precision adaptively model texture gradients, while cross-component prediction dynamically transfers luminance features to chroma planes via nonlinear coefficient scaling. For inter-coding, affine motion compensation captures complex scene dynamics through 6-degrees-of-freedom transformations, synergizing with compound prediction that fuses multi-frame optical flow trajectories and spatial Overlapped Block Motion Compensation (OBMC) filtering. These innovations are reinforced by entropy coding optimized through machine-learned symbol distributions, achieving 30%+ bitrate reduction at equivalent visual quality compared to HEVC. Such systemic improvements in spatial–temporal redundancy elimination explain AV1's dominance in modern

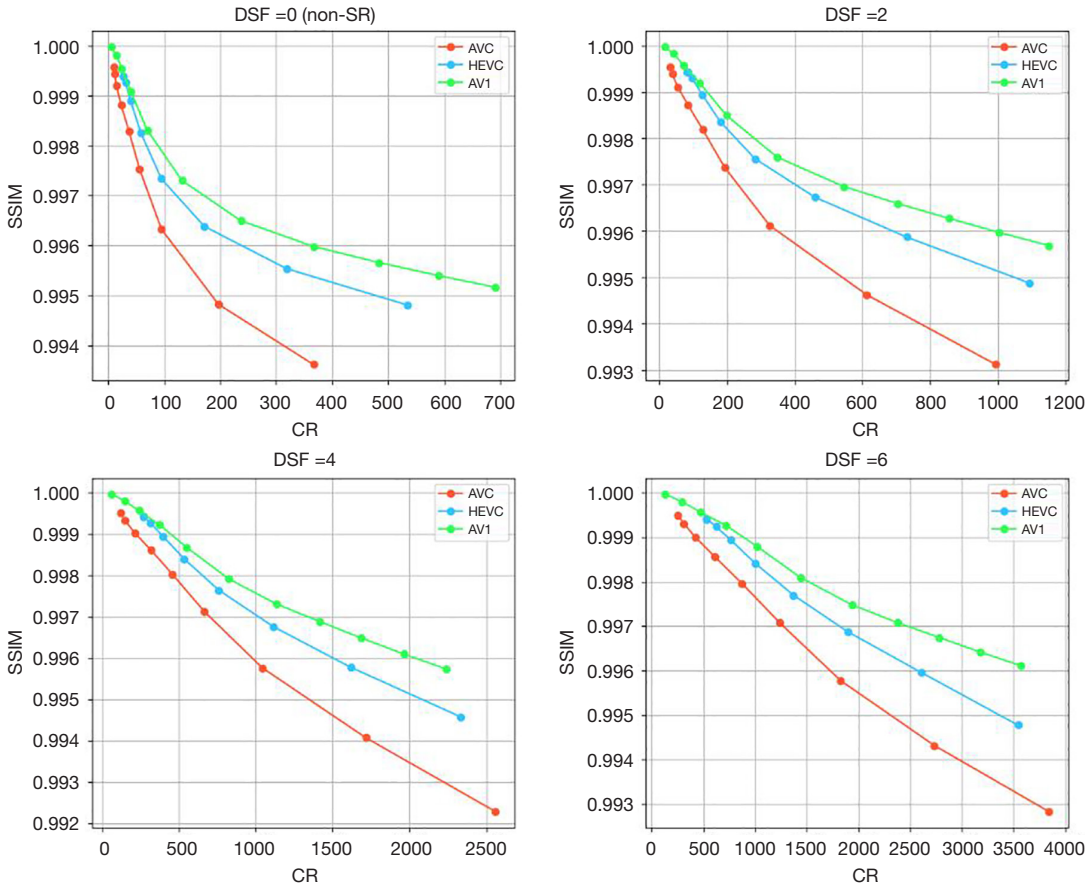

Figure 9 Relationship between CR and SSIM of three video encoders with respect to four DSFs. AV1, AOMedia Video 1; AVC, Advanced Video Coding; CR, compression ratio; DSFs, down-sampling factors; HEVC, High Efficiency Video Coding; SSIM, structural similarity index measure.

codec benchmarks.

The SR models showed different performance due to their distinct architectures. As the early model, SR-CNN is basically a convolutional network that falls short in capturing the detailed structures of projection images, resulting in only modest resolution improvements. SR-ResNet enhances this by leveraging deep residual learning, enabling more complex feature extraction and achieving higher PSNR with clearer reconstructions. However, SR-ResNet's outputs appear overly smooth and may lack the fine textures that are crucial for medical imaging analysis. SR-GAN, with its adversarial training framework, surpasses these limitations through a generator-discriminator structure that drives the model to produce HR images that close to their original images. The inclusion of perceptual loss based on VGG feature maps allows SR-GAN to prioritize visual fidelity, producing sharper textures and realistic details essential for accurate medical interpretation. Although SR-GAN may not always have the highest PSNR, its ability to create lifelike, detailed images makes it the best candidate for super-resolving projection images.

There are certain limitations in this research which can be improved in the future. First, we only selected several popular and representative video encoders and SR models for test. As of today, there newer codecs and DL models have emerged. They could be more powerful and suitable than current models. Second, when applying a video encoder on the projection images, different sorting methods can be used to assemble the projection images into a sequence. If the correlation quality of this sequence were to improve, the resulting CR by the video encoder could be increased. Finally, the current loss metrics may not be suitable to evaluate the loss of visual effect for restored images. For example, the PSNR and SSIM of SR-ResNet

are better than those of SR-GAN. However, in practice, the result achieved by SR-GAN has better visual perception. More advanced visual loss metrics, such as the mean opinion score (MOS), might be more appropriate to assess the effectiveness of the SR method.

Conclusions

The DL based SR method provides an effective way to restore HR images from LR images. Comparing to the classic linear or cubic interpolation methods which solely rely on the local image features, the DL network can learn global image features and extrapolate them to the new images. Our study shows that with a DSF of 2, the high-quality image can be restored perfectly. The current CR achieved by the existing video encoder can be escalated by a factor of 2^2–4^2. This is more important for those clinical applications in which a large volume of image data is to be stored for long term.

Acknowledgments

None.

Footnote

Data Sharing Statement: Available at https://qims.amegroups.com/article/view/10.21037/qims-2024-2962/dss

Funding: This work was supported by the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (No. 2024-RW320-05), CAMS Innovation Fund for Medical Sciences (CIFMS) (No. 2023-I2M-C&T-B-076), and the National High Level Hospital Clinical Research Funding (No. 2025-LYZX-R-B06) and Beijing Hope Run Special Fund of Cancer Foundation of China (No. LC2021B01).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-2024-2962/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments. The ethics committee of National Cancer Center/ Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College approved this study (No. NCC2018-016). The written informed consent was waived because of the retrospective design.

References

1. Létourneau D, Wong JW, Oldham M, Gulam M, Watt L, Jaffray DA, Siewerdsen JH, Martinez AA. Cone-beam-CT guided radiation therapy: technical implementation. Radiother Oncol 2005;75:279-86.
2. Li T, Xing L, Munro P, McGuinness C, Chao M, Yang Y, Loo B, Koong A. Four-dimensional cone-beam computed tomography using an on-board imager. Med Phys 2006;33:3825-33.
3. Park J, Park S, Kim J, Liu Z, Watkins W, Song W. TU-B-201B-04: Four-Dimensional Cone-Beam Computed Tomography and Digital Tomosynthesis Using Motion Signals Extracted from Fiducial Marker Inserted for Liver Cancer Radiation Therapy. Medical Physics 2010;37:3378.
4. Hugo GD, Weiss E, Sleeman WC, Balik S, Keall PJ, Lu J, Williamson JF. A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer. Med Phys 2017;44:762-71.
5. Waddington SP, McKenzie AL. Assessment of effective dose from concomitant exposures required in verification of the target volume in radiotherapy. Br J Radiol 2004;77:557-61.
6. Groh BA, Siewerdsen JH, Drake DG, Wong JW, Jaffray DA. A performance comparison of flat-panel imager-based MV and kV cone-beam CT. Med Phys 2002;29:967-75.
7. Howerton WB Jr, Mora MA. Advancements in digital imaging: what is new and on the horizon? J Am Dent Assoc 2008;139 Suppl:20S-4S.

8. Zhang J, He B, Yang Z, Kang W. A Novel Reconstruction of the Sparse-view CBCT Algorithm for Correcting Artifacts and Reducing Noise. Mathematics 2023;11:2127.
9. Sakurai Y, Ambo S, Nakamura M, Iramina H, Iizuka Y, Mitsuyoshi T, Matsuo Y, Mizowaki T. Development of a prediction model for target positioning by using diaphragm waveforms extracted from CBCT projection images. J Appl Clin Med Phys 2023;24:e14112.
10. Brennecke R, Bürgel U, Rippin G, Post F, Rupprecht HJ, Meyer J. Comparison of image compression viability for lossy and lossless JPEG and Wavelet data reduction in coronary angiography. Int J Cardiovasc Imaging 2001;17:1-12.
11. Liu X, An P, Chen Y, Huang X. An improved lossless image compression algorithm based on Huffman coding. Multimed Tools Appl 2022;81:4781-95.
12. Noreña T, Romero E. Compresión de imágenes médicas. Biomédica 2013;33:137-51.
13. Thayammal S, Selvathi D. A Review on Transform Based Image Compression Techniques. Int J Eng Res Technol 2013. doi: 10.17577/IJERTV2IS100437.
14. Liu F, Hernandez-Cabronero M, Sanchez V, Marcellin MW, Bilgin A. The Current Role of Image Compression Standards in Medical Imaging. Information (Basel) 2017;8:131.
15. Núñez-Gaona MA, Marcelín-Jiménez R, Gutiérrez-Martínez J, Aguirre-Meneses H, Gonzalez-Compean JL. A Dependable Massive Storage Service for Medical Imaging. J Digit Imaging 2018;31:628-39.
16. Bui V, Chang LC, Li D, Hsu LY, Chen MY. Comparison of lossless video and image compression codecs for medical computed tomography datasets. 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA; 2016:3960-2. doi: 10.1109/BigData.2016.7841075.
17. Huh SN, Zhang Y, Park JY, Indelicato DJ. Development of a Filtimator for Pediatric Image-Guided Radiation Therapy with Low Imaging Dose. International Journal of Radiation Oncology, Biology, Physics 2023;117:S178.
18. MacDonald RL, Fallone C, Chytyk-Praznik K, Robar J, Cherpak A. The feasibility of CT simulation-free adaptive radiation therapy. J Appl Clin Med Phys 2024;25:e14438.
19. Kong L, Li Z, Liu Y, Zhang J, Chen M, Zhou Q, Qi X, Deng X, Peng Y. A Generalized Deep Learning Method for Synthetic CT Generation. International Journal of Radiation Oncology, Biology, Physics 2023;117:e472.
20. Wang Z, Chen J, Hoi SCH. Deep Learning for Image Super-Resolution: A Survey. IEEE Trans Pattern Anal Mach Intell 2021;43:3365-87.
21. Zhang H, Huang J, Ma J, Bian Z, Feng Q, Lu H, Liang Z, Chen W. Iterative reconstruction for x-ray computed tomography using prior-image induced nonlocal regularization. IEEE Trans Biomed Eng 2014;61:2367-78.
22. Keys R. Cubic convolution interpolation for digital image processing. IEEE Transactions on Acoustics, Speech, and Signal Processing 1981;29:1153-60.
23. Ledig C, Theis L, Huszár F, Caballero J, Aitken AP, Tejani A, Totz J, Wang Z, Shi W. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016:105-14.
24. Shin M, Seo M, Lee K, Yoon K. Super-resolution techniques for biomedical applications and challenges. Biomed Eng Lett 2024;14:465-96.
25. He J, Ma H, Guo M, Wang J, Wang Z, Fan G. Research into super-resolution in medical imaging from 2000 to 2023: bibliometric analysis and visualization. Quant Imaging Med Surg 2024;14:5109-30.
26. Umirzakova S, Mardieva S, Muksimova S, Ahmad S, Whangbo T. Enhancing the Super-Resolution of Medical Images: Introducing the Deep Residual Feature Distillation Channel Attention Network for Optimized Performance and Efficiency. Bioengineering (Basel) 2023.
27. Mercat A, Viitanen M, Vanne J. UVG dataset: 50/120fps 4K sequences for video codec analysis and development. Proceedings of the 11th ACM Multimedia Systems Conference; 2020.
28. Wiegand T, Sullivan GJ, Bjøntegaard G, Luthra A. Overview of the H.264/AVC video coding standard. IEEE Transactions on Circuits and Systems for Video Technology 2003;13:560-76.
29. Sullivan GJ, Ohm J, Han W, Wiegand T. Overview of the High Efficiency Video Coding (HEVC) Standard. IEEE Transactions on Circuits and Systems for Video Technology 2012;22:1649-68.
30. Han J, Li B, Mukherjee D, Ching-Han C, Chen C, Su H, Parker S, Joshi U, Chen Y, Wang Y, Wilkins P, Xu Y, Bankoski J. A Technical Overview of AV1. Proceedings of the IEEE 2020;109:1435-62.
31. Park SC, Park MK, Kang MG. Super-resolution image reconstruction: a technical overview. IEEE Signal Processing Magazine 2003;20:21-36.
32. Dong C, Loy CC, He K, Tang X. Image Super-Resolution Using Deep Convolutional Networks. IEEE Trans Pattern Anal Mach Intell 2016;38:295-307.
33. Sara U, Akter M, Uddin M. Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative

Study. Journal of Computer and Communications 2019;7:8-18.
34. Frankewitsch T, Söhnlein S, Müller M, Prokosch HU. Computed Quality Assessment of MPEG4-compressed DICOM Video Data. Stud Health Technol Inform 2005;116:447-52.
35. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13:600-12.

Cite this article as: Chang Z, Shang J, Fan Y, Huang P, Hu Z, Zhang K, Dai J, Yan H. Deep learning-based super-resolution method for projection image compression in radiotherapy. Quant Imaging Med Surg 2025;15(9):8611-8626. doi: 10.21037/qims-2024-2962

Appendix 1 Introduction of video codecs

AVC, also known as H.264, is a video compression standard that incorporates advanced techniques to improve compression efficiency. It uses inter-frame prediction with motion compensation, allowing each macroblock to be predicted from multiple reference frames with sub-pixel motion estimation for higher accuracy. AVC also applies variable block-size motion estimation, encoding different parts of a frame with block sizes ranging from 16×16 to 4×4 pixels to optimize detail preservation. The standard supports multiple intra-prediction modes for more precise spatial prediction within frames. Additionally, it features Context-Adaptive Binary Arithmetic Coding (CABAC), an advanced entropy coding method that adapts to the statistical properties of the data, and a deblocking filter to reduce blocking artifacts and enhance visual quality.

HEVC, also known as H.265, is a video compression standard that builds on AVC with more advanced coding techniques. It introduces a flexible block partitioning structure called Coding Tree Units (CTUs), which can be as large as 64×64 pixels, enabling more efficient encoding of large, uniform regions. HEVC improves inter-frame prediction with hierarchical motion estimation and enhanced motion vector prediction, supporting more precise motion compensation. It also offers a more advanced intra-prediction scheme with up to 35 directional modes, compared to nine in AVC. Additionally, it uses advanced entropy coding methods like Context-Adaptive Binary Arithmetic Coding (CABAC) and incorporates loop filters, such as Sample Adaptive Offset (SAO) and deblocking filters, to minimize artifacts and enhance visual quality.

AV1 (AOMedia Video 1) is an open-source video compression standard that surpasses previous standards, such as AVC and HEVC, by integrating advanced coding tools to improve compression efficiency. It features flexible block partitioning with support for various sizes and shapes, allowing precise adaptation to the content's structure. AV1 enhances intra-prediction with 56 directional modes, offering greater accuracy in spatial prediction. It also employs sophisticated inter-frame prediction techniques, such as compound prediction, warp motion compensation, and temporal filtering, enabling more precise motion prediction and better use of temporal redundancy. The standard introduces advanced entropy coding tools like Constrained Directional Enhancement (CDEF) and Frame Super-Resolution to preserve detail while reducing bitrate. Additionally, AV1 integrates multiple loop filters to minimize artifacts and maintain high visual quality.

Appendix 2 Introduction of super-resolution deep-learning networks

The SR-CNN network structure used in this research is shown in *Figure S1*. SR-CNN employs a deep CNN with three essential layers: a patch extraction and representation layer, a nonlinear mapping layer, and a reconstruction layer. This structure enables the network to learn complex relationships between low- and high-resolution images directly from training data, effectively capturing details and textures. The simplicity and effectiveness of SR-CNN establish it as a foundational model in the field of image super-resolution. In this study, SR-CNN employs a lightweight three-layer convolutional structure, with 9×9 and 5×5 kernels sequentially extracting and restoring features, providing a simple yet effective solution for real-time super-resolution.

The SR-ResNet network structure used in this research is shown in *Figure S2*. SR-ResNet uses a series of residual blocks, allowing the model to effectively learn and retain fine image details by skipping unnecessary computations. The skip connections within these residual blocks help prevent the vanishing gradient problem, ensuring stable and efficient training. Additionally, SR-ResNet incorporates up-sampling layers at the end of the network, typically using pixel shuffling, to generate the final high-resolution output. This architecture balances performance and quality, delivering sharp and accurate images, making it well-suited for tasks that require high-fidelity visual enhancement. In this study, SR-ResNet builds on a residual network with 16 Batch Norm-enhanced residual blocks, using Pixel Shuffle up-sampling to recover details, making it suitable for tasks requiring high-detail restoration.

The SR-GAN network structure used in this research is shown in *Figure S3*. SR-GAN consists of two neural networks—a generator and a discriminator—trained through adversarial training. The generator, based on a deep CNN with residual blocks, creates high-resolution images from low-resolution inputs. The discriminator evaluates the outputs by distinguishing between generated images and real high-resolution images from the training dataset. A key feature of SR-GAN is its use of a perceptual loss function, which extends beyond traditional pixel-wise losses, such as mean squared error (MSE). This perceptual loss leverages high-level feature representations from a pre-trained deep network, typically the VGG network,

https://dx.doi.org/10.21037/qims-2024-2962

allowing the generator to produce images that are not only high in resolution but also visually appealing and closely aligned with human visual perception. In this study, the SR-GAN generator uses a deep CNN with residual blocks, combined with Pixel Shuffle up-sampling, to gradually restore image resolution. Through adversarial learning between the generator and discriminator, the network enhances the realism of the generated images. The discriminator applies convolutional layers, LayerNorm, and LeakyReLU activation to extract features and down-sample inputs, with a final fully connected layer and Sigmoid function providing the probability that the input is real or generated.

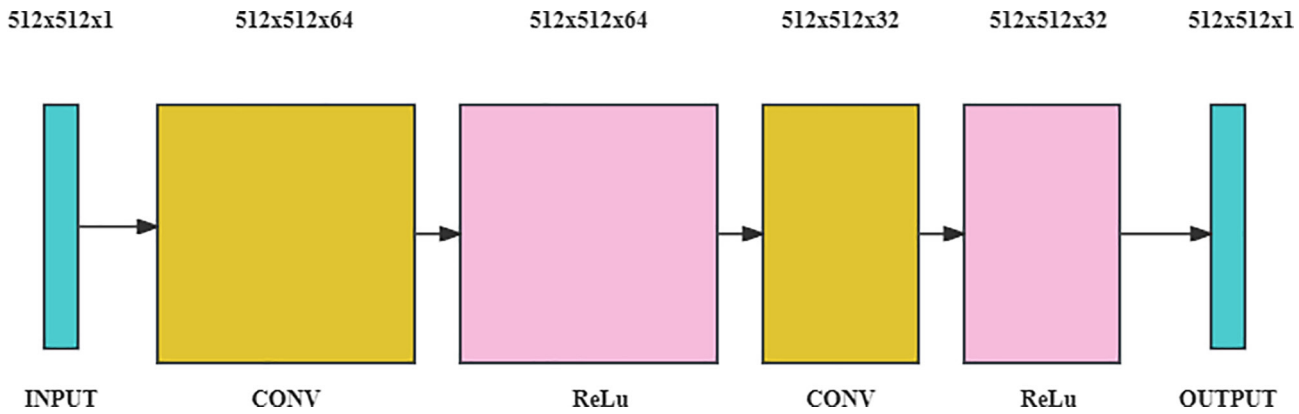

Figure S1 Architecture of SR-CNN.

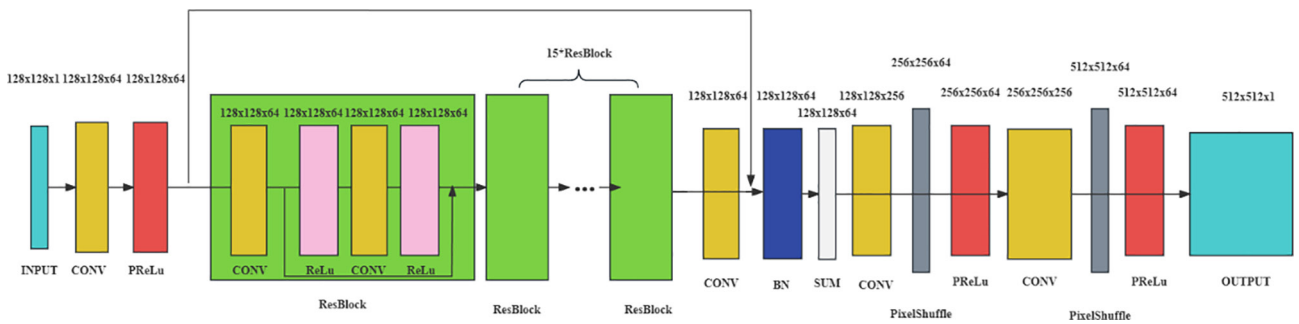

Figure S2 Architecture of SR-ResNet.

https://dx.doi.org/10.21037/qims-2024-2962

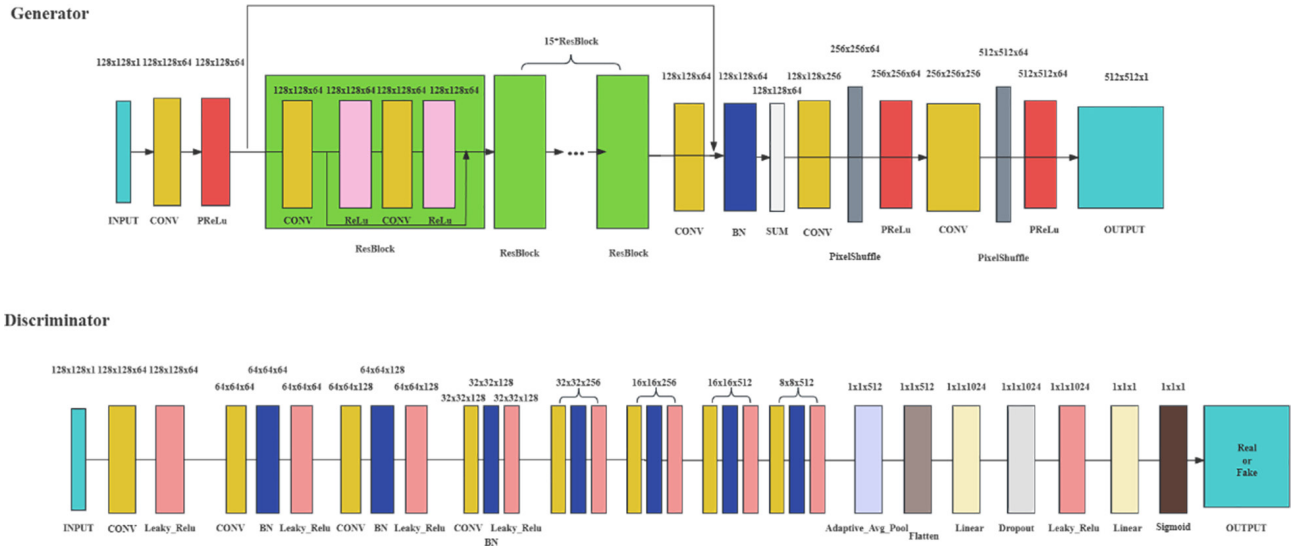

Figure S3 Architecture of SR-GAN.

https://dx.doi.org/10.21037/qims-2024-2962