DDASR: Domain-Distance Adapted Super-Resolution Reconstruction of MR Brain Images

Shan Cong[1,2,†], Kailong Cui[1,2,†], Yuzun Yang[2,†], Yang Zhou[3], Xinxin Wang[3], Haoran Luo[1,2], Yichi Zhang[4,*], Xiaohui Yao[1,2,*]

***Abstract*— High detail and fast magnetic resonance imaging (MRI) sequences are highly demanded in clinical settings, as inadequate imaging information can lead to diagnostic difficulties. MR image super-resolution (SR) is a promising way to address this issue, but its performance is limited due to the practical difficulty of acquiring paired low- and high-resolution (LR and HR) images. Most existing methods generate these pairs by down-sampling HR images, a process that often fails to capture complex degradations and domain-specific variations. In this study, we propose a domain-distance adapted SR framework (DDASR), which includes two stages: the domain-distance adapted down-sampling network (DSN) and the GAN-based super-resolution network (SRN). The DSN incorporates characteristics from unpaired LR images during down-sampling process, enabling the generation of domain-adapted LR images. Additionally, we present a novel GAN with enhanced attention U-Net and multi-layer perceptual loss. The proposed approach yields visually convincing textures and successfully restores outdated MRI data from the ADNI1 dataset, outperforming state-of-the-art SR approaches in both perceptual and quantitative evaluations. Code is available at https://github.com/Yaolab-fantastic/DDASR.**

I. INTRODUCTION

Magnetic resonance imaging (MRI) is widely used as a crucial neuroimaging method because it delivers detailed insights into brain tissue structure without the need for invasive procedures. However, obtaining high-quality MRI images often involves challenges such as the need for extended scanning times, high magnetic field strengths, and patient discomfort, which can lead to motion-sensitive images and increased costs [1]. Additionally, the complex nature of MR signals makes them prone to various distortions and artifacts, including motion blur, noise, and signal dropout. These further complicate the imaging process and potentially hinder accurate diagnoses. Single image super-resolution (SISR) techniques have been developed to enhance a low-resolution (LR) image by improving signal-to-noise ratio [2], accentuating texture details, and eliminating visual artifacts [3]. This technology is particularly beneficial in brain imaging as it offers a solution to the labor-intensive and demanding process of analyzing substandard MR images that lack sufficient information. While recent advances in deep learning-based MR image SR offer a promising solution to enhance image quality and maintain diagnostic quality [4], there are several challenges that must be addressed before clinical implementation.

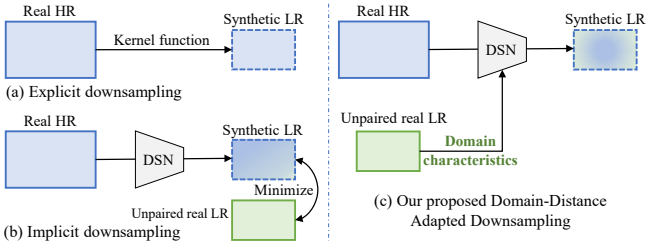

Fig. 1. Synthesizing downsampled LR images from the real HR and unpaired real LR datasets. (a) shows the explicit down-sampling. (b) shows the conventional implicit down-sampling, which proposes minimizing the distance between synthetic and real LR images. (c) shows our proposed domain-distance adapted down-sampling method, in which real LR domain characteristics were embedded into the down-sampling process to facilitate the domain translation.

One major challenge is modeling the unknown and complex degradation of a real LR image in the absence of paired LR/HR MR images. Since paired LR/HR MR images are usually unavailable, existing methods are dedicated to generating an LR/HR pair by synthesizing an LR image from the HR image to model the transformation [5], [6]. Recent works primarily focused on explicit modeling (as shown in Fig. 1(a)), such as bicubic down-sampling [7], k-space filtering [8], and combination of image degradation operations [5], [9]. These works performed well when the true degradation is known prior. However, the performance is significantly compromised when the real degradation differs from their estimation. Alternatively, a few other works employ implicit modeling (as shown in Fig. 1(b)), such as domain-distance adaptation using convolutional neural network (CNN) [6], data distribution learning with GAN [10], and representation learning [11]. *Although they offer greater flexibility in handling various degradation scenarios, as shown in Fig. 1(b), existing implicit methods primarily focus on minimizing the distance between real and synthetic images and overlook the transfer of domain representations.*

† These authors contributed equally to this work.
* Corresponding authors: Yichi Zhang and Xiaohui Yao.
[1,2]Shan Cong, Kailong Cui, Haoran Luo, and Xiaohui Yao are with the College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, Heilongjiang, China, and with the Qingdao Innovation and Development Center, Harbin Engineering University, Qingdao 266000, Shandong, China. shan.cong; cuikailong; luohaoran; xiaohui.yao@hrbeu.edu.cn
[2]Yuzun Yang is with the Qingdao Innovation and Development Center, Harbin Engineering University, Qingdao 266000, Shandong, China. yangyuzun@hrbeu.edu.cn
[3]Yang Zhou and Xinxin Wang are with the Harbin Medical University Cancer Hospital, Harbin 150000, China. zhouyang094@126.com; wangxinxin@hrbmu.edu.cn
[4]Yichi Zhang is with the School of Data Science, Fudan University, Shanghai, China. zhangyichi23@m.fudan.edu.cn

Another challenge is recovering texture details while preserving the topological structures. An enhanced image may pass quantitative assessment but fail visual inspection as existing algorithms usually fail in balancing sharpness and structural details [12]. In some cases, tissue structure is even altered, which can be a fatal error as the structural texture is critical for visual diagnostics (i.e., lesion localization) and neuroscience studies (i.e., brain parcellation). There are two common directions to prioritize the preservation of topological structures and texture details: 1) design proper deep learning architectures for SR, where recent efforts have been made to improve autoregressive models [13], GAN [14], [15], variational autoencoder (VAE) [16], [17], encoding-decoding networks [18], [11], and diffusion probabilistic models [19]; and 2) design loss functions that consider visual effects to produce convincing results like perceptual loss [14], [7] and texture matching loss [20], [21] functions. Overall, the investigation in both directions offers exciting opportunities for advancing MR image SR, and we believe that their combination has the potential to yield even more impressive results.

To tackle the aforementioned challenges, we develop a realistic down-sampling strategy, as illustrated in Fig. 1(c). This strategy aims to incorporate LR domain characteristics while learning the mapping relationship between the target domain (i.e., real LR) and the source domain (i.e., real HR). In addition, it includes an effective texture restoration component that utilizes multi-scale image features for enhancing MR brain image super-resolution. By combining these innovative designs, our proposed method significantly improves the preservation of both structural integrity and texture details, leading to better visual and quantitative quality. Our contributions can be summarized as follows:

- **Realistic down-sampling:** We introduce a transformer-based, domain-distance adapted down-sampling network (DSN) to efficiently capture domain representations and incorporate these features into the mapping relationship between HR images and LR images.
- **Structure and texture preservation:** We propose a super-resolution network (SRN) that employs a generator enhanced with multi-layer perceptual loss and a discriminator utilizing an attention U-Net to produce biologically reasonable textures and precise contours.
- **Denoising and super-resolution:** We propose a novel domain-distance adapted SR framework to enhance both the image quality and resolution of MR brain images, achieving substantial improvements over state-of-the-art models in both qualitative visual assessments and quantitative evaluations.

II. Methodology

Given two MR image domains characterized by two sets of unpaired HR images $\mathbf{X}_{\text{HR}}^{(\text{r})}$ and LR images $\mathbf{Y}_{\text{LR}}^{(\text{r})}$, our goal is to train an SR model that can increase the resolution of an LR domain image while enabling the synthesized HR image align with the real HR domain. To achieve this, we propose a two-stage framework that comprises a down-sampling network and a super-resolution network.

The down-sampling stage (Fig. 2(a)) employs a transformer architecture to embed the real LR domain representation into the downsampled images and introduces the content-aware positional encoding (CAPE, Fig. 2(c)) to capture invariant brain structural representation from source HR images and preserve fine-grained details. The SR stage (Fig. 2(b)) uses a GAN architecture, where the generator (Fig. 2(d)) employs an encoder-bottleneck-decoder scheme and the discriminator (Fig. 2(e)) adopts an attention U-Net. Meanwhile, to ensure visually convincing textures, we incorporate multi-layer perceptual loss into the SR stage. We present the details of the domain-distance adapted down-sampling strategy in Sect. II-A, the GAN-based SR framework in Sect. II-B for reproducibility. In principle, one can apply any image SR task in the proposed architecture.

A. Domain-Distance Adapted Down-Sampling

Existing down-sampling methods, such as linear, k-space, blur kernels, or a combination of down-sampling operations have low generalization ability and cannot adapt to complex degradation processes during MRI scans. To address this, we propose a domain distance-adaptive down-sampling network (DSN) based on the transformer architecture. As shown in Fig. 2(c), we employ transformer blocks with attention mechanisms to extract image domain characteristics, textures, and structural features, thereby facilitating the construction of source and target feature embeddings in the latent space [22]. This mechanism enables the transformer to capture local and global contextual information and generate multi-scale feature representation. In addition, we apply CAPE [23] on the source HR images ($\mathbf{X}_{\text{HR}}^{(\text{r})}$) during multi-scale down-sampling, which aims to ensure that the brain structure and texture information remains invariant during the transformation process by encoding the positional information.

Image positional coding. Given a source image $X_{\text{HR}}^{(\text{r})} \in \mathbb{R}^{H_s \times W_s}$ and a target image $Y_{\text{LR}}^{(\text{r})} \in \mathbb{R}^{H_t \times W_t}$, patches are projected into a sequential feature embedding E of size $N \times C$ using a linear projection layer, where $N = \frac{H}{m} \times \frac{W}{m}$ and C are the length and dimension of E respectively, and $m \times m$ is the patch size. The CAPE P_c was proposed to integrate the semantics of image content into the positional encoding, thereby reducing the impact of the image scale on the relative distance. Specifically, $P_c(i,j)$ between patches (i,j) is calculated as follows:

$$\begin{aligned} P_L &= F_{\text{pos}}\left(\text{AvgPool}_{n\times n}(E)\right), \\ P_c(i,j) &= \sum_{k=0}^{s}\sum_{l=0}^{s}\left(a_{kl} \times P_L\left(i_k, j_l\right)\right), \end{aligned} \tag{1}$$

where F_{pos} is a 1×1 convolution operation, P_L is the learnable relative positional relationship, n is the block size, a_{kl} is the interpolation weight, and s is the number of adjacent patches.

Transformer encoder. Two transformer encoders are employed to encode the structure information from the source

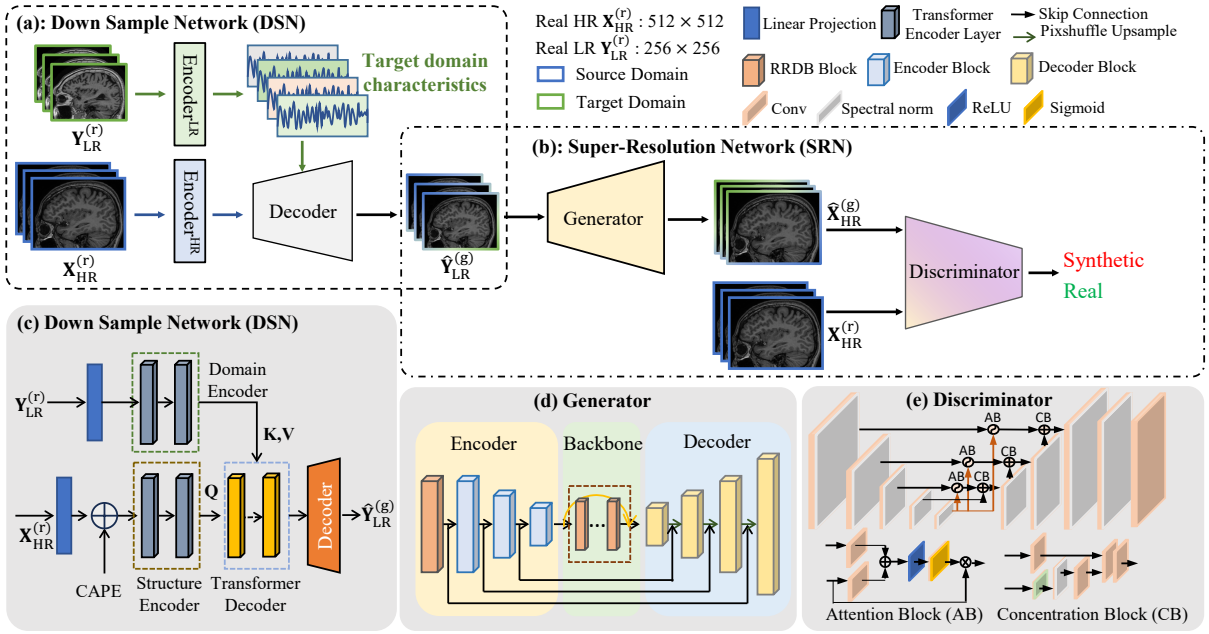

Fig. 2. Overview of the proposed two-stage SR framework. (a) Stage 1: taking the real HR and real unpaired LR data as input, a transformer-based down-sampling network (DSN) is designed to generate paired (to the HR images) LR images, aligning to the real LR domain. (b) Stage 2: the synthesized (LR, HR) pairs are used to train the super-resolution network (SRN), where the GAN architecture with multi-layer perceptual loss (for the generator) and attention U-Net mechanisms (for the discriminator) is employed. (c-e) illustrate the details of DSN, GAN generator, and discriminator, respectively.

domain image $X_{\text{HR}}^{(\text{r})}$ and encode the domain characteristic information from the target domain image $Y_{\text{LR}}^{(\text{r})}$. Specifically, the embeddings of the source domain image sequence $Z_s = \{E_{s1} + P_{c1}, E_{s2} + P_{c2}, \cdots, E_{sL} + P_{cL},\}$ and the target domain image sequence $Z_t = \{E_{t1}, E_{t2}, \cdots, E_{tL},\}$ are fed into each transformer encoder to produce the encoded structure sequence $\xi_{\text{HR}}^{(\text{r})}$ and domain characteristic sequence $\xi_{\text{LR}}^{(\text{r})}$. Each encoder layer is composed of a multi-head attention (MHA) module and a feed-forward network (FFN) [24], which together encode the input sequence as follows:

$$\begin{aligned} Q = ZW^Q, \quad K = ZW^K, \quad V = ZW^V, \\ \text{MHA}(Q,K,V) = \Big\Vert_{i=1}^{h} \text{Attention}_i\,(Q,K,V)\cdot W^O, \end{aligned} \tag{2}$$

$$\begin{aligned} \xi' &= \text{MHA}(Q,K,V) + Q, \\ \xi &= \text{E}(Z) = \text{FFN}(\xi') + \xi'. \end{aligned} \tag{3}$$

For each input sequence, the queries (Q), keys (K), and values (V) are generated via three learnable weight matrices W^Q, W^K, and W^V, respectively. The $\Vert$ denotes the concatenation of h independent heads, and W^O is the projection matrix that aggregates the output from all attention heads.

Following Eq. 2 and Eq. 3, the source HR images and the target LR images are encoded as $\xi_{\text{HR}}^{(\text{r})} = \text{E}(Z_s)$ and $\xi_{\text{LR}}^{(\text{r})} = \text{E}(Z_t)$.

Transformer decoder. As shown in Fig. 2(c), the transformer decoder is used to generate decoded image $Y_{\text{LR}}^{(\text{g})'}$ from the source image sequence $\xi_{\text{HR}}^{(\text{r})}$ while absorbing the domain characteristic information from the target image sequence $\xi_{\text{LR}}^{(\text{r})}$. In particular, the transformer decoder consists of two cross-attention layers and one FFN, from which the decoded image $Y_{\text{LR}}^{(\text{g})'}$ is calculated as:

$$\begin{aligned} &Q = \hat{\xi}_{\text{HR}}^{(\text{r})}W^Q, \quad K = \xi_{\text{LR}}^{(\text{r})}W^K, \quad V = \xi_{\text{LR}}^{(\text{r})}W^V, \\ &Y_{\text{LR}}^{(\text{g})'''} = \text{MHA}(Q,K,V) + Q, \\ &Y_{\text{LR}}^{(\text{g})''} = \text{MHA}(Y_{\text{LR}}^{(\text{g})'''} + P_c, K, V) + Y_{\text{LR}}^{(\text{g})'''}, \\ &Y_{\text{LR}}^{(\text{g})'} = \text{FNN}(Y_{\text{LR}}^{(\text{g})''}) + Y_{\text{LR}}^{(\text{g})''}. \end{aligned} \tag{4}$$

Here, $\hat{\xi}_{\text{HR}}^{(\text{r})} = \{\xi_{\text{HR},1}^{(\text{r})} + P_{c1}, \xi_{\text{HR},2}^{(\text{r})} + P_{c2}, \cdots, \xi_{\text{HR,L}}^{(\text{r})} + P_{cL}\}$ is the CAPE encoded source image sequence.

Resizing. For the training purposes, the decoded output $Y_{\text{LR}}^{(g)'}$ undergoes additional processing through a three-layers up-sampling decoder [25] and a linear degradation operation to obtain the downsampled result $\hat{Y}_{\text{LR}}^{(g)}$. There is a convolution layer of 3×3 convolution kernel, ReLU activation function, and $2\times$ upsample applied in each three-layer convolution up-sampling module, respectively. By controlling the number of layers of the CNN up-sampling module, we can output images with any integer multiple downsampled multiple.

DSN loss. There are two objectives for down-sampling: preserving the original structure and texture of the source HR MR image and attaining the domain characteristics of the target LR images. To achieve these, two loss functions are introduced: the structural loss $\mathcal{L}_s$ that reduces the structure and texture discrepancies between the generated LR ($\hat{Y}_{\text{LR}}^{(g)}$) and real HR images ($X_{\text{HR}}^{(\text{r})}$), and the domain loss $\mathcal{L}_d$ to address the gap between synthesized LR and target LR domains (i.e., real LR domain). To improve the model stability, we employ the pre-trained VGG19 to extract feature maps and

train the model:

$$\mathcal{L}_s = \sum_{l=0}^{N_l} \left\| \phi_l \left(\hat{Y}_{\mathrm{LR}}^{(\mathrm{g})} \right) - \phi_l \left(X_{\mathrm{HR}}^{(\mathrm{r})} \right) \right\|_2, \tag{5}$$

$$\begin{aligned} \mathcal{L}_d = & \sum_{l=0}^{N_l} \left\| \mu \left(\phi_l \left(\hat{Y}_{\mathrm{LR}}^{(\mathrm{g})} \right) \right) - \mu \left(\phi_l \left(Y_{\mathrm{LR}}^{(\mathrm{r})} \right) \right) \right\|_2 \\ & + \left\| \sigma \left(\phi_l \left(\hat{Y}_{\mathrm{LR}}^{(\mathrm{g})} \right) \right) - \sigma \left(\phi_l \left(Y_{\mathrm{LR}}^{(\mathrm{r})} \right) \right) \right\|_2, \end{aligned} \tag{6}$$

where N_l denotes the number of VGG layers, and $\phi_l(\cdot)$ extracts features from the l-th VGG layer, $\mu(\cdot)$ and $\sigma(\cdot)$ represent the mean and variance of the features, respectively.

To help keep more detailed and local structural information, we further introduce an identity loss $\mathcal{L}_{id}$ into the DSN optimization. Instead of trading off between image structure and domain, the identity loss considers both local mapping and global distributions by maintaining structural information rather than changing domain representation. Let $I_s^{(\mathrm{g})}$ and $I_t^{(\mathrm{g})}$ denote the images generated by inputting two identical source domain images I_s and two identical target domain images I_t into the DSN, respectively. The identity loss measures the difference between I_s and $I_s^{(\mathrm{g})}$ for the source domain, and between I_t and $I_t^{(\mathrm{g})}$ for the target domain:

$$\begin{aligned} \mathcal{L}_{id} = & \lambda_{id1} \left(||I_s^{(\mathrm{g})} - I_s||_2 + ||I_t^{(\mathrm{g})} - I_t||_2 \right) \\ & + \lambda_{id2} \sum_{l=0}^{N_l} \left\| \phi_l(I_s^{(\mathrm{g})}) - \phi_l(I_s) \right\|_2 + \left\| \phi_l(I_t^{(\mathrm{g})}) - \phi_l(I_t) \right\|_2. \end{aligned} \tag{7}$$

It should be noted that all images are resized to have the same dimension during the training process.

Overall, the total DSN loss can be expressed as:

$$\mathcal{L}_{DSN} = \lambda_s \mathcal{L}_s + \lambda_d \mathcal{L}_d + \mathcal{L}_{id}. \tag{8}$$

Empirically, we set $\lambda_s = 7$, $\lambda_d = 10$, $\lambda_{id1} = 30$ and $\lambda_{id2} = 5$ across the experiments.

B. GAN-based super-resolution

Using the pseudo (LR, HR) image pairs synthesized by the DSN, a GAN-based super-resolution network (SRN) is trained to model the HR reconstruction. We now detail the SRN framework, as illustrated in Fig. 2(b,d,e).

Generator. The generator consists of an encoder, a bottleneck, and a decoder (see Fig. 2(d)). The encoder extracts features from the input LR images, the bottleneck incorporates dense connections and deeper networks for deep feature fusion, and the decoder merges features from both the encoder and bottleneck to generate HR images. This process emphasizes the restoration of both large-scale contour structures and fine-grained texture details.

Encoder. We employ the residual-in-residual dense block (RRDB) [14] for effective feature extraction, which first extracts features from the LR image, yielding a low-dimensional feature representation f_0. To capture multi-scale feature information, we then employ convolutional modules that progressively downsample the features while expanding their dimensionality. This process yields multi-scale features $f_i = E_i(f_{i-1}), i \in \{1, \cdots, N\}$, where $E_i(\cdot)$ represents a convolutional layer with a kernel size of 4, a stride of 2, and padding of 1.

Bottleneck. Previous studies [26], [27], [28] have shown the effectiveness of integrating dense connections and deeper networks in SR tasks. To enhance deep feature extraction and fusion, we use cascaded RRDB blocks within a high-dimensional space. This method facilitates the integration of high-dimensional features both before and after deepening the network with dense connections. Meanwhile, residual blocks can help reduce computational overhead and address gradient vanishing issues commonly associated with deeper networks. The proposed method is designed to promote the optimization of critical visual features and facilitate the extraction of detailed features, particularly lines and curves.

Decoder. A decoder with progressive fusion is used to merge the encoder and bottleneck features for HR MR image generation. This decoder utilizes convolutional layers with a kernel size of 3 and padding of 1 for upsampling, along with an efficient sub-pixel convolution operation (pixshuffle) at each layer. Skip connections are used to retain the large-scale low-dimensional information obtained from the encoder and perform multi-scale fusion with the small-scale high-dimensional information generated by the bottleneck, thereby increasing the amount of information obtained by the encoder. This allows the model to simultaneously focus on both the large-scale contour structures and small-scale texture details, resulting in more detailed and accurate MR image reconstruction.

Discriminator. Inspired by [29], we adopt an attention U-net discriminator architecture to specifically target and improve poorly generated textures during adversarial training. This architecture includes attention blocks (AB) and concentration blocks (CB), building upon a U-Net structure, as shown in Fig. 2(e). To boost training stability and tailor the discriminator more effectively for image SR tasks, we apply a spectral normalization regularization operation as described in [30].

SRN loss. The SRN loss comprises the generator loss and the discriminator loss. We denote the generator and discriminator as $G(\cdot)$ and $D(\cdot)$, respectively. Specifically, to preserve the structural integrity of MR images and enhance their texture details, we introduce the multi-layer perceptual loss $\mathcal{L}_{mper}$ into the generator. The pre-trained VGG19 is used to extract image features for the multi-layer perceptual loss, with layers $l \in \{2, 7, 9, 12, 17\}$ selected and weighted at $\tau_l = \{0.1, 0.1, 1.0, 1.0, 1.0\}$. We further employ the 1-norm distance loss $\mathcal{L}_1$ into the generator to evaluate the distance between the generated HR image and the ground truth. The multi-layer perceptual loss and content distance loss are as follows:

$$\mathcal{L}_{mper} = \sum_{l \in \{2,7,9,12,17\}} \tau_l \left\| \phi_l(X_{\mathrm{HR}}^{(\mathrm{r})}) - \phi_l(\hat{X}_{\mathrm{HR}}^{(\mathrm{g})}) \right\|_2, \tag{9}$$

$$\mathcal{L}_1 = \mathbb{E}_{\hat{y}_{\mathrm{LR}}^{(\mathrm{g})}} \left\| G \left(\hat{y}_{\mathrm{LR}}^{(\mathrm{g})} \right) - X_{\mathrm{HR}}^{(\mathrm{r})} \right\|_1. \tag{10}$$

Therefore, the discriminator loss, enhanced generator loss, and the total SRN loss are calculated as follows:

$$\begin{aligned}\mathcal{L}_D = &-\mathbb{E}_{X_{\mathrm{HR}}^{(\mathrm{r})}}\left[\log\left(D\left(X_{\mathrm{HR}}^{(\mathrm{r})}, \hat{X}_{\mathrm{HR}}^{(\mathrm{g})}\right)\right)\right] \\ &- \mathbb{E}_{X_{\mathrm{HR}}^{(\mathrm{g})}}\left[\log\left(1 - D\left(\hat{X}_{\mathrm{HR}}^{(\mathrm{g})}, X_{\mathrm{HR}}^{(\mathrm{r})}\right)\right)\right],\end{aligned} \tag{11}$$

$$\begin{aligned}\mathcal{L}'_G = &-\mathbb{E}_{X_{\mathrm{HR}}^{(\mathrm{r})}}\left[\log\left(1 - D\left(X_{\mathrm{HR}}^{(\mathrm{r})}, \hat{X}_{\mathrm{HR}}^{(\mathrm{g})}\right)\right)\right] \\ &- \mathbb{E}_{X_{\mathrm{HR}}^{(\mathrm{g})}}\left[\log\left(D\left(\hat{X}_{\mathrm{HR}}^{(\mathrm{g})}, X_{\mathrm{HR}}^{(\mathrm{r})}\right)\right)\right],\end{aligned} \tag{12}$$

$$\mathcal{L}_G = \eta_1 \mathcal{L}_{mper} + \eta_2 \mathcal{L}_1 + \eta_3 \mathcal{L}'_G, \tag{13}$$

$$\mathcal{L}_{SRN} = \mathcal{L}_G + \mathcal{L}_D, \tag{14}$$

where η_1, η_2 and η_3 are hyperparameters for adjusting different loss terms. We set $\eta_1 = 1$, $\eta_2 = 1$, and $\eta_3 = 0.1$ across the experiments.

III. Experiments

We utilized two unpaired HR/LR MR image datasets to assess the performance of our proposed method. Specifically, a private HR dataset (i.e., $\mathbf{X}_{\mathrm{HR}}^{(\mathrm{r})}$) from our institute served as the reference to guide the super-resolution of the LR images ($\mathbf{Y}_{\mathrm{LR}}^{(\mathrm{r})}$), which were sourced from the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. In addition, we collected paired LR images corresponding to our private HR scans to serve as ground truth for validating the down-sampling ability of the proposed DSN module. Further details on these datasets are provided below.

A. Experimental setup

ADNI LR image dataset. We include a total of 41 T1-weighted 1.5T MRI scans from ADNI Phase 1 (ADNI1) for super-resolution. Although the original resolution of these images, $256 \times 256 \times 160$, is not low compared to clinical MR scans, this study aims to achieve a higher sampling density to demonstrate the efficacy of the proposed SR framework. In addition, ADNI1 data is generally considered to have lower image quality compared to the later ADNI phases (i.e., ADNIGO/2 and ADNI 3). Therefore, it motivates us to use ADNI1 images as real LR data and apply our proposed methods to them for both image denoising and SR tasks. From these scans, we extract $3,520$ sagittal images from 22 subjects used for training (as real LR), and $3,040$ images from 19 subjects for testing. Detailed data acquisition protocols can be accessed at `adni.loni.usc.edu`.

Private HR/LR paired image dataset. Four volunteers underwent both HR and LR MRI scanning on a 3.0T GE scanner at our site using a 3D T1-w MP-RAGE sequence. These collected real HR images serve as source references for ADNI LR SR, while our private LR images are used exclusively for DSN validation. The HR scans feature a slice thickness/gap of 1/0 mm, spatial resolution of $1.0 \times 1.0 \times 1.0$ mm, TE of 3.444 ms, TR of 2737.62 ms, and a scanning time of 6 min 48s. The same sequence is used for the LR scanning with a reduced TE of 1.66 ms and scanning time of 4 min 28s. The resolutions of the private HR and LR MRI are $512 \times 512 \times 388$ and $256 \times 256 \times 388$, respectively. As HR and LR MRI share the same field of view (FOV), HR is defined by higher sampling density, whereas LR by lower sampling density. The private HR dataset is divided into the training and testing sets with a ratio of $8:2$, resulting in $1,241$ images for training and 311 images for testing.

Evaluation metrics. In the absence of ground truth (e.g., real ADNI HR images are not available), natural image quality evaluator (NIQE) [31] and blind/referenceless image spatial quality evaluator (BRISQUE) [32] are employed to assess the performance of our proposed DDASR framework. Here, NIQE evaluates image quality by analyzing the structure and content to simulate the perceptual quality of the human visual system. BRISQUE quantifies image quality by analyzing statistical features within the spatial domain of the image.

When ground truth is present (e.g., the private LR images are available for comparing the down-sampling performance), we use the peak signal-to-noise ratio (PSNR) [33] and the structural similarity index (SSIM) [34] to evaluate the performance of our proposed DSN (Sec.III-C). Specifically, PSNR quantifies image quality by comparing the signal-to-noise ratio between the real and synthetic images, while SSIM evaluates similarity from three key aspects: luminance, contrast, and structure.

Comparison with prior work. We compare the proposed SR framework with a benchmark method Bicubic, and three state-of-the-art methods including ESRGAN [14], RealSR [35], and RealESRGAN [5]. We further compare our proposed DSN framework with the state-of-the-art DS approaches utilized in ESRGAN, RealSR, and RealESRGAN. Specifically, ESRGAN employs the k-space DS method, RealSR utilizes a degradation framework to estimate blur kernels, and RealESRGAN adopts a high-order degradation modeling method.

To ensure a fair comparison, all compared methods were retrained on our dataset using the same model architectures and optimal hyperparameters as described in their papers. The mean and standard deviation computed across all testing samples are reported for each evaluation metric.

Implementation details. Our proposed model is trained on 2 NVIDIA GeForce RTX 3090 GPUs with 48 GB memory using the PyTorch framework. The DSN is trained for 300,000 iterations with a batch size of 3 and a hidden dimension of 512. For the SRN, a cosine annealing with warm restarts learning rate optimization strategy is employed with a minimum learning rate of 1e-7, and the total number of iterations is set to 250,000.

B. Comparison with the state-of-the-arts

We apply the proposed SR framework on the real ADNI LR data, acknowledged to have lower image quality. Non-reference metrics (NIQE and BRISQUE) are used to evaluate the performance. As illustrated in Table I, our method significantly outperforms others in terms of the NIQE score and achieves the second-best BRISQUE index score. Among the compared SR methods, Bicubic upsampling yields the

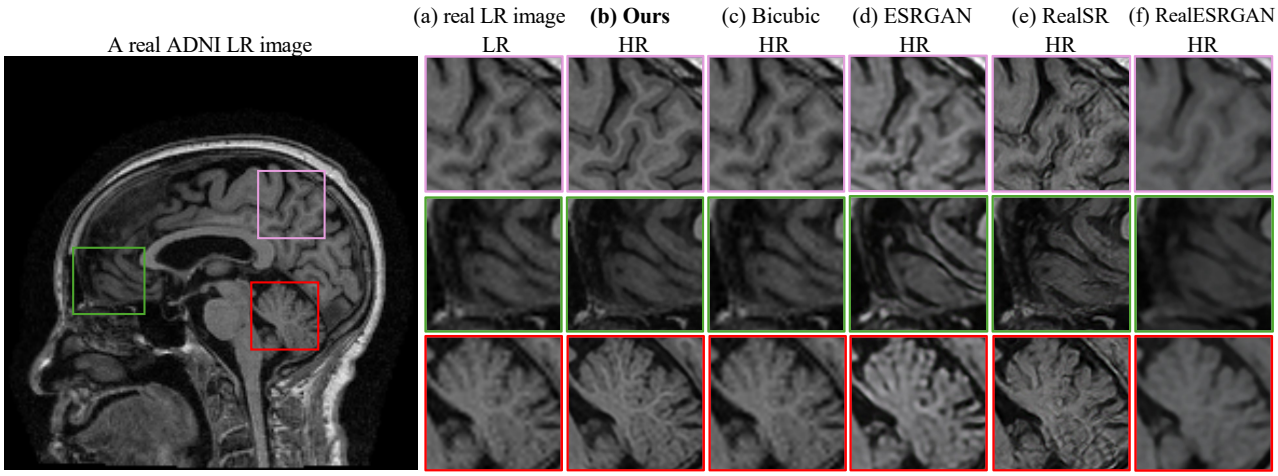

Fig. 3. Visual comparison of the SR results from different methods.

TABLE I
PERFORMANCE COMPARISON OF ADNI LR IMAGES SR.

Methods	$\text{NIQE}_{\text{std}} \downarrow$	$\text{BRISQUE}_{\text{std}} \downarrow$
Real LR	5.18±0.89	68.56±6.22
Bicubic	4.91±0.71	61.31±3.91
ESRGAN	6.00±0.24	60.87±3.92
RealSR	5.34±0.16	57.03±2.13
RealESRGAN	5.01±0.16	**40.43±9.43**
DDASR (**ours**)	**4.25±0.15***	44.67±3.21

Best results are in **bold**. The second-best results are underlined. Results with '$*$' indicate that our model is significantly better ($p < 0.05$) than the suboptimal method when using the independent t-test.

poorest BRISQUE scores, likely because it simply interpolates the image without considering quality improvement. ESRGAN shows the worst NIQE scores, potentially because it does not take the domain gap into consideration.

We further perform a visual comparison of several SR results. As shown in Fig. 3, our proposed method effectively enhances MR image quality by simultaneously preserving structural integrity and recovering biologically significant texture information. RealSR delivers good visual outcomes in certain regions but introduces serious structural alterations and unnecessary texture in many other regions. Although RealESRGAN achieves the best BRISQUE scores, it significantly compromises texture details and suffers from artifacts. Notably, it expands the cortical surface thickness, reduces the depth of brain sulci, and results in unclear boundaries between the cortex and medulla, which may hinder the detection of small lesions.

C. Ablation study

We conducted a series of ablation studies to evaluate the effectiveness of the proposed SR framework, the DSN and SRN modules, as well as the key components designed for the SRN. Below, we describe the details of each ablation design and present the corresponding results.

DSN: more realistic synthesized LR image.

We compare our proposed DSN framework with the DS approaches utilized in RealSR and RealESRGAN, as well as with the k-space method employed by ESRGAN. PSNR and SSIM are used to quantitatively measure the differences between the synthesized LR images and the ground truth (i.e., the real LR images from our site).

As shown in Table II and Fig. 4, our proposed DSN achieves the best performance in terms of PSNR and real-synthetic image distance (i.e., heatmaps in the bottom panel of Fig. 4) and yields the second-best SSIM score. Although k-space obtains the highest SSIM, it actually generates many artifact texture details (see the top panel of Fig. 4(c) and corresponding heatmap). Our DSN outperforms both RealSR and RealESRGAN across all metrics. RealESRGAN produces the poorest PSNR and SSIM scores, possibly due to its down-sampling strategy being optimized for natural images.

TABLE II
PERFORMANCE COMPARISON OF THE DOWN-SAMPLING METHODS ON THE REAL PRIVATE LR IMAGES.

Methods	$\text{PSNR}_{\text{std}} \uparrow$	$\text{SSIM}_{\text{std}} \uparrow$
K-space	20.52±4.57	**0.71±0.10**
RealESRGAN	19.98±7.53	0.46±0.21
RealSR	20.79±4.37	0.64±0.08
Ours	**21.65±4.21***	0.68±0.09

Best results are in **bold**. The second-best results are underlined. Results with '$*$' indicate that our DSN is significantly better ($P < 0.05$) than the suboptimal method when using the independent t-test.

Super-resolution network evaluation. We feed the down-sampled LR images generated by each of the benchmark methods into our SRN to examine whether our proposed SR method surpasses the performance of the others. The evaluation is conducted on the ADNI data, employing NIQE and BRISQUE metrics for performance comparison. As shown in Table III, our SRN demonstrates superior performance than the other SR methods regardless of the synthetic LR inputs. This demonstrates the generalizability and stability of our

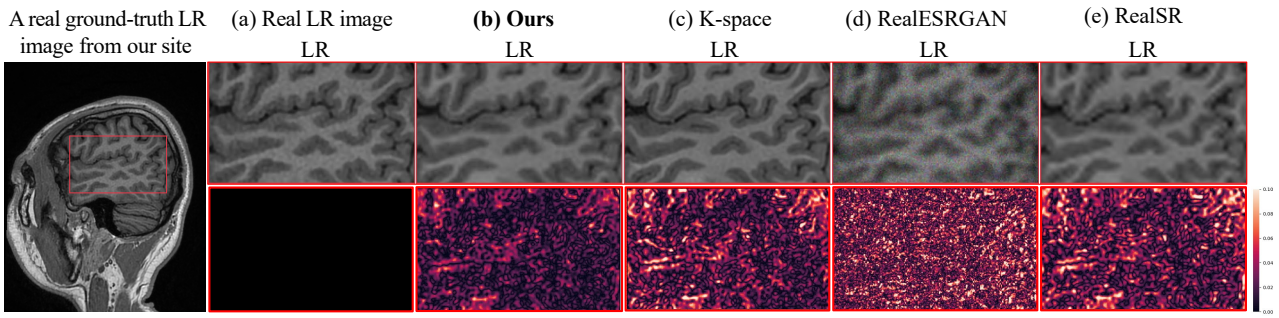

Fig. 4. Comparison of the down-sampling methods. A lower heatmap value indicates a smaller distance between the synthesized LR image and the real LR image.

proposed SRN.

TABLE III
PERFORMANCE EVALUATION OF THE PROPOSED SRN ON ADNI LR IMAGE SR.

Methods	$\text{NIQE}_{\text{std}} \downarrow$	$\text{BRISQUE}_{\text{std}} \downarrow$
Bicubic	4.91±0.71	61.31±3.91
Bicubic$^{\#}$	4.33±0.14	52.12±3.83
ESRGAN	6.00±0.24	60.87±3.92
ESRGAN$^{\#}$	5.08±0.55	103.23±4.30
RealSR	5.34±0.16	57.03±2.13
RealSR$^{\#}$	4.75±0.16	54.87±4.32
RealESRGAN	5.01±0.16	40.46±9.43
RealESRGAN$^{\#}$	4.51±0.14	35.35±2.83

All method utilize their specific downsampling strategy. Methods marked with superscript # employ our proposed SRN module, while those without the superscript use their own SRN modules.

Evaluation of GAN key components. We further evaluated the effectiveness of the key components designed for the GAN. We accordingly remove the multi-layer perceptual loss ($\mathcal{L}_{mper}$) from the generator and replace the attention U-Net architecture with the relativistic discriminator (D(RA)) [36]. Ablation experiments are conducted to super-resolve ADNI LR images, utilizing NIQE and BRISQUE metrics for performance comparison. The results, presented in Table IV, indicate that removing the multi-layer perceptual loss leads to a 17.9% reduction in NIQE and a 15.4% reduction in BRISQUE. We also observe that the attention-based U-Net discriminator demonstrates significant advantages, achieving the best results across both evaluation metrics (13.2% BRISQUE and 14.5% NIQE reductions).

IV. CONCLUSION

We propose a novel domain-distance adapted super-resolution framework for MR brain images SR. Our approach combines a transformer-based domain adaptation network, an encoding-decoding generator architecture, an attention UNet-based discriminator for SR, and multi-layer perceptual loss to preserve brain structural and texture information effectively. Experimental results demonstrate the superiority of our proposed model compared to state-of-the-art SR methods in both quantitative evaluation and perceptual quality assessments. This indicates that our proposed framework is a reliable solution for restoring outdated MRI images and has the potential to enhance clinical MR images. The proposed approach has wide applicability to various medical imaging tasks and can assist medical professionals in obtaining more accurate and detailed diagnostic information from 1.5T MR brain images. Therefore, our work has noteworthy contributions to the field of medical imaging and could have a positive impact on patient care.

TABLE IV
ABLATION STUDY OF THE MULTI-PERCEPTUAL LOSS AND THE ATTENTION U-NET MECHANISM ON ADNI LR IMAGE SR.

Methods	$\text{NIQE}_{\text{std}} \downarrow$	$\text{BRISQUE}_{\text{std}} \downarrow$
w/o $\mathcal{L}_{mper}$	5.01±0.21	51.54±3.22
w/ $\mathcal{L}_{mper}$	**4.25±0.15***	**44.67±3.21***
D(RA)	4.87±0.23	50.58±4.44
D(Attention U-Net)	**4.25±0.15***	**44.67±3.21***

Best results are in **bold**. Results with '$*$' indicate that the key component shows significant improvement (t-test $p < 0.05$).

ACKNOWLEDGMENT

This work is supported partly by the National Natural Science Foundation of China (62103116, 62102115), the Natural Science Foundation of Heilongjiang Province (LH2022F016), the Fundamental Research Funds for the Central Universities (3072024GH2604), the Shandong Provincial Natural Science Foundation (2022HWYQ-093), Heilongjiang Province basic research business fees for provincial higher education institutions (2023-KYYWF-0217), Harbin Medical University Cancer Hospital Climbing program (PDYS2024-10).

For private HR data, informed consent has been obtained from all participants and the research ethics committees or institutional review boards approved the data collection. ADNI data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

REFERENCES

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[2] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[3] B. Sahiner, A. Pezeshk, L. M. Hadjiiski, X. Wang, K. Drukker, K. H. Cha, R. M. Summers, and M. L. Giger, "Deep learning in medical imaging and radiation therapy," *Medical physics*, vol. 46, no. 1, pp. e1–e36, 2019.

[4] J. D. Rudie, T. Gleason, M. J. Barkovich, D. M. Wilson, A. Shankaranarayanan, T. Zhang, L. Wang, E. Gong, G. Zaharchuk, and J. E. Villanueva-Meyer, "Clinical assessment of deep learning–based super-resolution for 3d volumetric brain mri," *Radiology: Artificial Intelligence*, vol. 4, no. 2, p. e210059, 2022.

[5] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1905–1914, 2021.

[6] Y. Wei, S. Gu, Y. Li, R. Timofte, L. Jin, and H. Song, "Unsupervised real-world image super resolution via domain-distance aware training," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13 385–13 394, 2021.

[7] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

[8] C. Zhao, B. E. Dewey, D. L. Pham, P. A. Calabresi, D. S. Reich, and J. L. Prince, "Smore: a self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning," *IEEE transactions on medical imaging*, vol. 40, no. 3, pp. 805–817, 2020.

[9] C. Liu and D. Sun, "On bayesian adaptive video super resolution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 346–360, 2013.

[10] B. Huang, H. Xiao, W. Liu, Y. Zhang, H. Wu, W. Wang, Y. Yang, Y. Yang, G. W. Miller, T. Li *et al.*, "Mri super-resolution via realistic downsampling with adversarial learning," *Physics in Medicine & Biology*, vol. 66, no. 20, p. 205004, 2021.

[11] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo, "Unsupervised degradation representation learning for blind super-resolution," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10 581–10 590, 2021.

[12] W. Zhou, Q. Jiang, Y. Wang, Z. Chen, and W. Li, "Blind quality assessment for image superresolution using deep two-stream convolutional networks," *Information Sciences*, vol. 528, pp. 205–218, 2020.

[13] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *Inter. conf. on machine learning*, pp. 1747–1756, 2016.

[14] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," *The European Conference on Computer Vision Workshops (ECCVW)*, 2018.

[15] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] Z.-S. Liu, W.-C. Siu, and Y.-L. Chan, "Photo-realistic image super-resolution via variational autoencoders," *IEEE Transactions on Circuits and Systems for video Technology*, vol. 31, no. 4, pp. 1351–1365, 2020.

[17] D. Chira, I. Haralampiev, O. Winther, A. Dittadi, and V. Liévin, "Image super-resolution with deep variational autoencoders," *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pp. 395–411, 2023.

[18] K. C. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "Glean: Generative latent bank for large-factor image super-resolution," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14 245–14 254, 2021.

[19] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Srdiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[20] M. W. Gondal, B. Schölkopf, and M. Hirsch, "The unreasonable effectiveness of texture transfer for single image super-resolution," *Computer Vision–ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part V 15*, pp. 80–97, 2019.

[21] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," *Proceedings of the IEEE international conference on computer vision*, pp. 4491–4500, 2017.

[22] D. Han, R. Yu, S. Li, J. Wang, Y. Yang, Z. Zhao, Y. Wei, and S. Cong, "Mr image harmonization with transformer," *2023 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 2448–2453, 2023.

[23] Y. Deng, F. Tang, W. Dong, C. Ma, X. Pan, L. Wang, and C. Xu, "Stytr2: Image style transfer with transformers," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11 326–11 336, 2022.

[24] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," *International conference on machine learning*, pp. 4055–4064, 2018.

[25] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6881–6890, 2021.

[26] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

[27] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," *Proceedings of the European conference on computer vision (ECCV)*, pp. 286–301, 2018.

[28] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," *Proceedings of the IEEE international conference on computer vision*, pp. 4799–4807, 2017.

[29] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[30] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *International Conference on Learning Representations*, 2018.

[31] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[32] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[33] D. Poobathy and R. M. Chezian, "Edge detection operators: Peak signal to noise ratio based comparison," *Int J Image, Graphics and Signal Processing*, vol. 10, pp. 55–61, 2014.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[35] X. Ji, Y. Cao, Y. Tai, C. Wang, J. Li, and F. Huang, "Real-world super-resolution via kernel estimation and noise injection," *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 466–467, 2020.

[36] Z. Wei, Y. Huang, Y. Chen, C. Zheng, and J. Gao, "A-esrgan: Training real-world blind super-resolution with attention u-net discriminators," *Pacific Rim International Conference on Artificial Intelligence*, pp. 16–27, 2023.