scientific reports

OPEN

An efficient dual-branch framework via implicit self-texture enhancement for arbitrary-scale histopathology image super-resolution

Minghong Duan[1,2], Linhao Qu[1,2], Zhiwei Yang[2,3], Manning Wang[1,2], Chenxi Zhang[1,2]✉ & Zhijian Song[1,2]✉

High-quality whole-slide scanning is expensive, complex, and time-consuming, thus limiting the acquisition and utilization of high-resolution histopathology images in daily clinical work. Deep learning-based single-image super-resolution (SISR) techniques provide an effective way to solve this problem. However, the existing SISR models applied in histopathology images can only work in fixed integer scaling factors, decreasing their applicability. Though methods based on implicit neural representation (INR) have shown promising results in arbitrary-scale super-resolution (SR) of natural images, applying them directly to histopathology images is inadequate because they have unique fine-grained image textures different from natural images. Thus, we propose an Implicit Self-Texture Enhancement-based dual-branch framework (ISTE) for arbitrary-scale SR of histopathology images to address this challenge. The proposed ISTE contains a feature aggregation branch and a texture learning branch. We employ the feature aggregation branch to enhance the learning of the local details for SR images while utilizing the texture learning branch to enhance the learning of high-frequency texture details. Then, we design a two-stage texture enhancement strategy to fuse the features from the two branches to obtain the SR images. Experiments on publicly available datasets, including TMA, HistoSR, and the TCGA lung cancer datasets, demonstrate that ISTE outperforms existing fixed-scale and arbitrary-scale SR algorithms across various scaling factors. Additionally, extensive experiments have shown that the histopathology images reconstructed by the proposed ISTE are applicable to downstream pathology image analysis tasks.

High-resolution (HR) whole slide images (WSIs) contain rich cellular morphology and pathological patterns, and they are the gold standard for clinical diagnosis and the basis for automated histopathology image analysis tasks, including segmentation and classification[1–4]. However, the acquisition and utilization of digital WSIs remain limited in the daily clinical workflow[4,5]. On the one hand, HR digital WSIs are typically obtained through sophisticated and costly whole-slide scanning equipment, which is often difficult to access in remote and underserved regions. On the other hand, acquiring HR digital WSIs involves using dedicated micro-cameras within the whole slide scanner to capture image fragments from different local regions of the specimen, which are then stitched together to form a complete image depicting the entire specimen[6]. Such a digital process is highly time-consuming[4,5]. Furthermore, HR digital WSIs are very large, often reaching gigapixels, which places additional demands on clinical funding support, professional training, ample data storage, and efficient data management[2,7]. Therefore, if it is possible to scan low-resolution (LR) histopathology images with cheaper devices while designing algorithms that can produce WSIs maintaining high quality, the digitization process could be accelerated, and the clinical application of automated techniques to analyze histopathology images could be promoted[4,5,8].

[1]Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China. [2]Shanghai Key Laboratory of Medical Image Computing and Computer Assisted Intervention, Shanghai 200032, China. [3]Academy for Engineering and Technology, Fudan University, Shanghai 200433, China. ✉email: chenxizhang@fudan.edu.cn; zjsong@fudan.edu.cn

Super-resolution (SR) algorithms based on deep learning can accurately map a single LR image to an HR image[10,14,17–25]. Recently, deep learning-based methods have been widely applied in histopathology image SR. Most approaches construct a large dataset of LR-HR image pairs to train neural networks in an end-to-end manner. The trained neural networks can generate HR images with input LR images. For example, Mukherjee et al.[10] utilized a convolutional neural network with an upsampling layer to produce SR images. Chen et al.[12] proposed a spatial wavelet dual-stream network to perform the SR image generation. As shown in Fig. 1a, although these previous methods demonstrate promising performance, they can only be trained and tested at fixed integer scales as they rely on up-sampling modules such as learnable deconvolution or pixel shuffle[10,12]. If different scaling factors are required, the network would need to be retrained for each specific scale. However, in clinical pathological diagnosis, doctors usually need to continuously zoom in and out of sections at different scaling factors, so the applicability of these models is greatly limited. This highlights the importance of arbitrary-scale SR models for histopathology imaging. Once trained, such a model could perform SR at multiple scales without the need for retraining. Furthermore, it enables scaling at any magnification, including non-integer scaling factors. This capability not only assists doctors in observing and analyzing histopathology images at various scales, leading to more accurate diagnoses, but also better meets clinical needs for images at different magnifications. Unfortunately, to our knowledge, no existing arbitrary-scale SR model is specifically designed for histopathology images.

Recently, inspired by implicit neural representation (INR)[26–28], some studies have pioneered arbitrary-scale SR for natural images[15,29]. For example, Chen et al.[15] proposed the local implicit image function (LIIF), which represents 2D images as latent code through an encoder and maps the input coordinates and corresponding latent variables to RGB values through the decoding function based on the multilayer perceptron (MLP), enabling image SR at arbitrary scales. As shown in Fig. 1b, although these methods can be directly applied to histopathology images, they do not account for the unique texture characteristics of histopathology images, resulting in sub-optimal performance. As shown in Fig. 1d, histopathology images contain a large amount of fine-grained cell morphology and repetition, unlike natural images. Better reconstructing the unique texture characteristics at arbitrary scales is essential for histopathology image SR.

Motivated by the observation above, we propose an efficient dual-branch framework based on implicit self-texture enhancement (ISTE) for arbitrary-scale SR of histopathology images to better deal with its special texture. Figure 1c briefly illustrates the overall framework of ISTE. Specifically, ISTE consists of a feature aggregation branch and a texture learning branch. In the feature aggregation branch, we introduce the Local Feature Interaction (LFI) module, which is designed to enhance feature interaction within local regions and to focus the framework's attention on discriminative local details such as the morphology and structure of cell nuclei. In the texture learning branch, we propose the Texture Learner (TL), aiming to enhance the learning of high-frequency texture information, including details like intercellular gaps and tissue texture fragments. After that, we design a two-stage texture enhancement strategy for these two branches, where the first stage is feature-based texture enhancement, and the second stage is spatial domain-based texture enhancement. Considering that histopathology images contain many similar cell morphologies and periodic texture patterns, we assume that these similar regions can assist each other in reconstruction in the feature space, so we design the self-texture fusion (STF) module to accomplish feature-based texture enhancement. The main idea is to retrieve the texture information from the texture learning branch and transfer it to the feature aggregation branch for information fusion and enhancement. For spatial domain-based texture enhancement, we decode the features of the two branches into RGB values in the spatial domain using the local pixel decoder (LPD) and the local texture

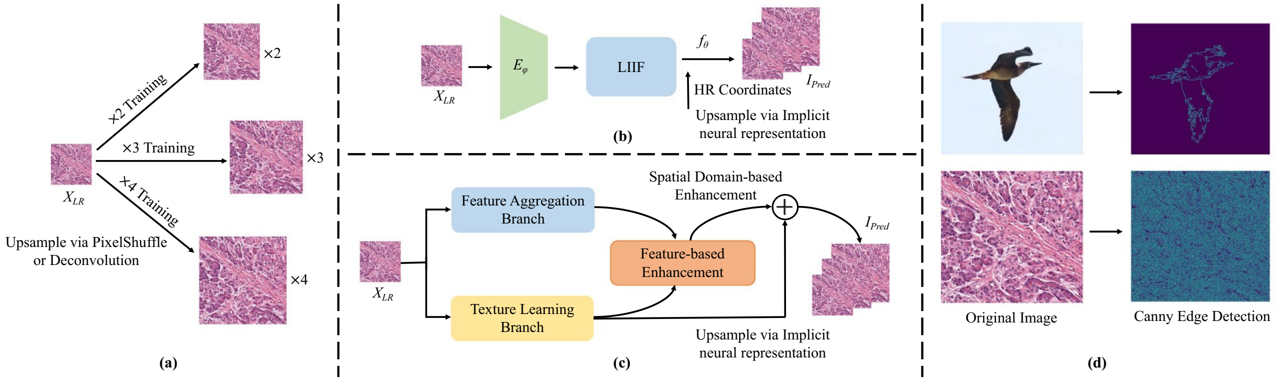

×2
×2 Training
×3 Training
×3
×4 Training
×4
X_{LR}
Upsample via PixelShuffle or Deconvolution
(a)
X_{LR}
E_φ
LIIF
f_θ
I_{Pred}
HR Coordinates
Upsample via Implicit neural representation
(b)
Feature Aggregation Branch
Spatial Domain-based Enhancement
Feature-based Enhancement
I_{Pred}
X_{LR}
Texture Learning Branch
Upsample via Implicit neural representation
(c)
Original Image
Canny Edge Detection
(d)

Fig. 1. Motivation of our ISTE. (**a**) Existing SR methods for histopathology images[6,9–14] can only achieve fixed integer-scale SR and need to retrain the model to achieve different scaling factors; (**b**) Existing SR algorithms based on implicit neural networks for natural images (exemplified by LIIF[15]) perform SR directly in the spatial domain, and lack attention and enhancement of image texture information; (**c**) ISTE is an efficient dual-branch framework based on implicit self-texture enhancement for arbitrary-scale histopathology image SR. ISTE further enhances its performance through feature-based and spatial domain-based texture enhancement; (**d**) We use the Canny operator[16] to extract texture from both natural and histopathology images. It is evident that, in contrast to natural images, histopathology images contain a large amount of fine-grained cell morphology and arrangement information, and they tend to have richer texture information.

decoder (LTD), respectively, and perform information fusion in the spatial domain. These two decoders are based on implicit neural networks[15], thus enabling image SR at arbitrary scales. Extensive experiments on three public datasets have shown that ISTE performs better than existing fixed-scale and arbitrary-scale SR algorithms at multiple scales and helps to improve downstream task performance. Overall, the contributions of this paper are as follows:

- We introduce ISTE, an efficient dual-branch framework based on implicit self-texture enhancement for arbitrary-scale SR of histopathology images. ISTE recovers the texture details from the low resolution image through feature-based texture enhancement and spatial domain-based texture enhancement.
- The proposed ISTE achieves state-of-the-art performance at various scaling factors on three public datasets, and we demonstrate the effectiveness of the proposed texture enhancement strategy through a series of ablation experiments.
- The histopathology images reconstructed by ISTE are shown to be effective for two downstream tasks in pathology image analysis: gland segmentation and cancer detection. The performance of these tasks can be improved by using the reconstructed images.

Related works

Deep learning-based super-resolution methods for natural images

Single-image super-resolution (SISR) refers to recovering an HR image from an LR image or an LR image sequence, which is a classical low-level computer vision task with a wide range of applications[19–25]. Deep neural networks can achieve accurate mapping from LR images to HR images due to their powerful fitting ability. Thus, they have become the mainstream approach in current SR studies. Numerous methods based on convolutional neural networks (CNNs) have been proposed for natural image SR, including SRCNN[30], EDSR[17], and RDN[31]. To further improve the performance of SR, some methods utilized residual modules[32,33], densely connected modules[34,35], and other blocks[36,37] for the design of the CNNs. Subsequently, a series of SR methods based on attention mechanism have emerged, such as channel attention[38,39], self-attention (IPT[40], SwinIR[41]), and non-local attention[42,43]. However, these methods can only be trained and tested at a fixed integer scale, and need to be retrained for new scaling factors.

In recent years, implicit neural representation (INR) has been proposed as a continuous data representation for various tasks in computer vision[26–28]. INR uses a neural network (usually a coordinate-based MLP) to establish a mapping between coordinates and their signal values, which allows continuous and efficient modeling of 2D image signals. This approach has been widely used in research on arbitrary-scale SR[15,29,44–46]. For example, Chen et al.[15] first applied INR to the SR algorithm and proposed the local implicit image function (LIIF) for arbitrary-scale SR. Lee et al.[29] proposed the local texture estimator (LTE), which transforms coordinates into the fourier domain information to enhance the representation of the local implicit function. Chen et al.[44] proposed the local implicit transformer (LIT) to enhance the local implicit function's focus on the context of the target reconstruction region. Fu et al.[45] introduced the local mixed implicit network (LMI), which considers multiple independent point coordinates and features to learn the spatial texture information of real-world images in a mix manner. Although these methods can be directly applied to histopathology images for continuous scale super-resolution, they fail to recover the special textures of the histopathology images effectively.

Deep learning-based super-resolution methods for pathological images

In recent years, deep learning-based SR algorithms have been widely used in pathological images to improve imaging resolution[6,9–14,47,48]. Upadhyay et al.[9] developed a generative adversarial network that integrated the tasks of pathological image SR and surgical smoke removal into a single framework. Mukherjee et al.[10] implemented SR image generation using a CNN with an up-sampling layer and augmented the outputs using the K-nearest neighbor algorithm. Chen et al.[12] accomplished the SR task through a spatial wavelet dual-stream network incorporating a refined context fusion module. Xie et al.[47] proposed the multi-features extraction module and the multi-scale selective fusion method to better extract and fuse multi-scale features for super-resolution. Li et al.[14] employed a multi-scale CNN-based generative adversarial network for SR image generation and introduced a curriculum learning training strategy. Wu et al.[6] incorporated a magnification classification branch into the SR network, improving SR performance through multi-task learning. These studies demonstrate the promise of using SR to enhance pathological image resolution in resource-limited settings. However, they still have some limitations. For instance, they restrict training and testing to specific scaling factors, and the resultant SR outputs still leave room for refinement. We attribute this primarily to a lack of adequate consideration for the unique textural characteristics of pathological images. In this paper, we introduce ISTE as a solution to address these challenges, aiming to achieve high-quality arbitrary-scale SR of pathological images.

Methods

Problem formulation and framework overview

Given a set of N pairs of corresponding LR images and HR images $\left\{X_{LR}^{i}, Y_{HR}^{i}\right\}_{i=1}^{N}$, the objective is to find the optimal parameters $\hat{\theta}$ of the SR model F_{θ}:

$$\hat{\theta} = \arg_{\theta} \min \frac{1}{N} \sum_{i=1}^{N} L\left(F_{\theta}\left(X_{LR}^{i}\right), Y_{HR}^{i}\right) \tag{1}$$

where X_{LR}^{i} is a LR image and Y_{HR}^{i} is its corresponding ground truth, and L is the L1 loss function to measure the difference between the ground-truth and the generated SR images. Figure 2 shows the overall framework of the proposed ISTE. We first utilize the backbone of SwinIR[41] as the encoder to perform feature pre-extraction on the input LR image X_{LR} and then input the pre-extracted feature F_{LR} into the upper feature aggregation branch and lower texture learning branch of ISTE, respectively. In the feature aggregation branch, we input the feature F_{LR} into the local feature interactor (LFI) to enhance the interaction of features in the local region and obtain feature F_{LFI}, which helps to improve the model's ability to focus on local details in the image. In the texture learning branch, we input the feature F_{LR} into the texture learner (TL) to enhance the learning of high-frequency information and extract the feature F_{TL}. Then we design a two-stage texture enhancement strategy for these two branches, where the first stage is feature-based texture enhancement, and the second stage is spatial domain-based texture enhancement. In the first stage, we designed the self-texture fusion (STF) module to leverage the interaction of similar regions of the pathological images in the feature space, thereby accomplishing feature-based texture enhancement to assist in reconstruction. In the second stage, we decode the F_{STF} from the STF module to obtain the image I_{LPD} through the local pixel decoder (LPD). Simultaneously, we decode the F_{TL} from the TL module to obtain the image I_{LTD} through the local texture decoder (LTD). Subsequently, we perform spatial summation of I_{LTD} and I_{LPD}, obtaining the final reconstructed HR image I_{Pred}. The purpose of the second stage is to fully utilize the features F_{TL} learned by the texture learner and decode them into the spatial domain for texture enhancement.

Local feature interactor

We propose the LFI module to enhance the interaction of features within local regions, thereby capturing the correlation of features within local regions to improve the model's focus on local details such as the morphology and structure of cell in the histopathology image. As shown in Fig. 3, the size of the feature map F_{LR} is $h \times w \times 64$, and we denote each vector of F_{LR} as $F_{LR}^{j} (j = 1, 2, \ldots, h \times w)$. The LFI first assigns a window of size 3×3 to each vector of F_{LR}, and the eight neighboring vectors in the window around F_{LR}^{j} form a set $F_{N}^{j} = \left\{ F_{N_i}^{j} \mid i = 3, 4, \ldots, 10 \right\}$. The average pooling result of the vectors within a window is denoted as F_{P}^{j}. The feature map F_{LFI} output by the LFI is calculated through self-attention so that each point on the feature map incorporates local features while paying more attention to itself. We denote each vector of F_{LFI} as $F_{LFI}^{j} (j = 1, 2, \ldots, h \times w)$, and it is calculated as follows:

$$F_{LFI}^{j} = \sum_{i=1}^{10} \frac{\exp\left(\left(Q_{LR}^{j} \right)^{T} K_{i}^{j} \right)}{\sqrt{d} \Sigma_{i=1}^{10} \exp\left(\left(Q_{LR}^{j} \right)^{T} K_{i}^{j} \right)} V_{i}^{j} \tag{2}$$

where Q_{LR}^{j} is the query mapped linearly from F_{LR}^{j}, K_{1}^{j} is the key mapped linearly from F_{LR}^{j}, V_{1}^{j} is the value mapped linearly from F_{LR}^{j}, K_{2}^{j} is the key mapped linearly from F_{P}^{j}, V_{2}^{j} is the value mapped linearly from F_{P}^{j}, $\left\{ K_{i}^{j} \mid i = 3, 4, \ldots, 10 \right\}$ is the key mapped linearly from F_{N}^{j}, $\left\{ V_{i}^{j} \mid i = 3, 4, \ldots, 10 \right\}$ is the value mapped

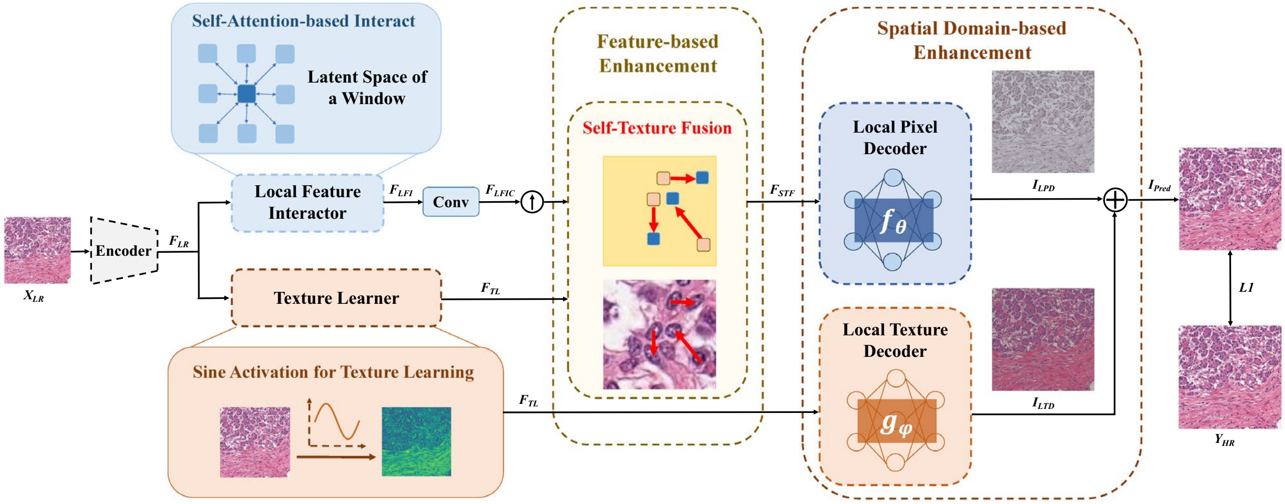


Fig. 2. Workflow of our ISTE. The LR image X_{LR} is input into the encoder to get the pre-extracted feature map F_{LR} first. In the feature aggregation branch, we input the feature F_{LR} into the local feature interactor and a convolutional layer to obtain F_{LFIC}. In the texture learning branch, we input the feature F_{LR} into the texture learner to obtain the texture feature F_{TL}. Then the feature maps from the two branches are input to the self-texture fusion module to accomplish feature-based enhancement. Finally, the enhanced feature F_{STF} output from the STF module and the texture feature F_{TL} output from the texture learner are decoded into RGB values respectively, and added up to accomplish spatial domain-based texture enhancement.

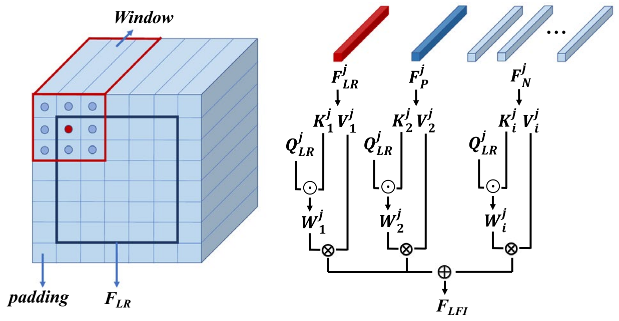


Fig. 3. Local feature interactor.

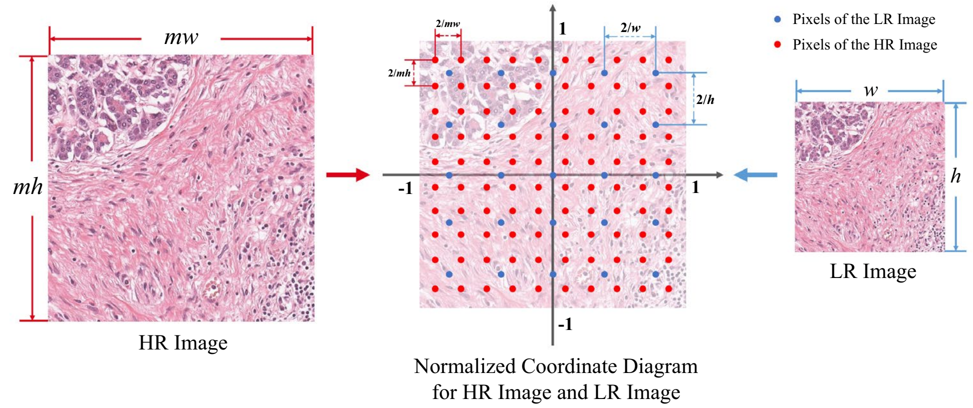


Fig. 4. Illustration of coordinate normalization. The red dots represent the pixels of the HR image, with coordinates denoted as (X', Y'). The blue dots represent the pixels of the LR image, with coordinates denoted as (*X*, *Y*). After coordinate normalization, each pixel of the HR image has a corresponding nearest neighbor pixel in the LR image.

linearly from F^j_N, and *d* is the dimension of these vectors. The parameters used by each window are shared in the self-attention calculation.

Texture learner

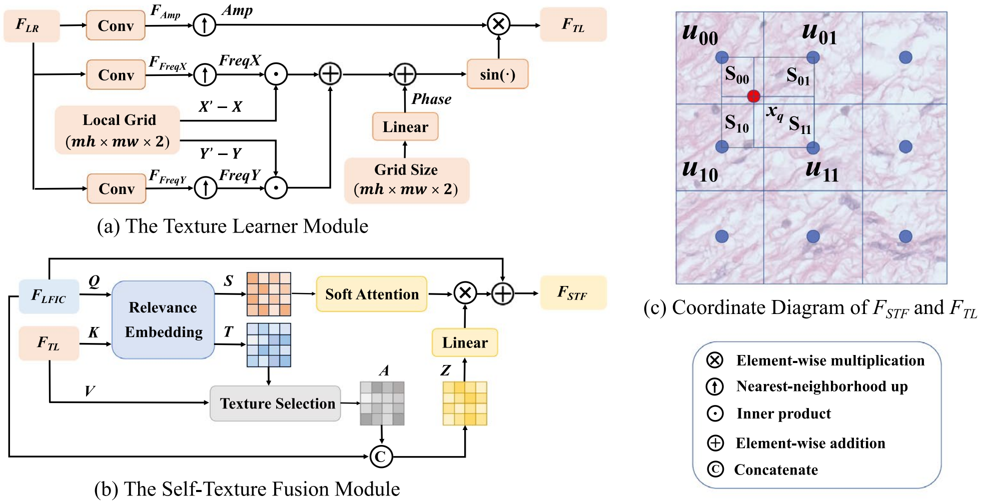
Inspired by LTE[29], we propose the TL module for learning high-frequency texture information in histopathology images. We employ sine activation to effectively enhance implicit neural representations for learning high-frequency texture details in the images, thereby mitigating spectral bias issues stemming from the ReLU activation functions[26]. As shown in Fig. 4, we normalize each pixel's 2D coordinate $(X', Y') = \left\{\left(\mathrm{x}'_i, \mathrm{y}'_j\right) \mid i = 1, 2, \ldots, mw, j = 1, 2, \ldots, mh\right\}$ in the continuous HR image domain and the 2D coordinate $(X, Y) = \left\{\left(\mathrm{x}_i, \mathrm{y}_j\right) \mid i = 1, 2, \ldots, mw, j = 1, 2, \ldots, mh\right\}$ nearest to (X', Y') in the continuous LR image domain between −1 and 1, where *m* represents the scaling factor. The local grid is defined as $(X' - X, Y' - Y)$. Each HR image pixel has a corresponding closest pixel in the LR image. As shown in Fig. 5a, the TL module firstly outputs three feature maps $F_{Amp} \in h \times w \times 256$, $F_{FreqX} \in h \times w \times 256$ and $F_{FreqY} \in h \times w \times 256$ through three convolutional layers respectively, and predicts the feature maps $Amp \in mh \times mw \times 256$, $FreqX \in mh \times mw \times 256$ and $FreqY \in mh \times mw \times 256$ corresponding to each pixel coordinate of the HR image through nearest-neighbor interpolation. Then we use linear projection based on an MLP and Sigmoid activation function to map $(2/\mathrm{mw}, 2/\mathrm{mh})$ to a 256-dimensional feature vector *Phase* to simulate the effect of texture fragment offset when the image scaling factor changes. The output of the TL module is calculated as follows:

$$F_{TL} = Amp \otimes \mathrm{Sin}(FreqX \odot \left(X' - X\right) + FreqY \odot \left(Y' - Y\right) + Phase) \tag{3}$$

where ⊗ represents element-wise multiplication and ⊙ represents inner product operation.

Self-texture fusion module for feature-based enhancement

Inspired by SRNTT[49] and T2 Net[50], we propose the STF module based on cross-attention, which aims to globally retrieve texture features from F_{TL} that are most similar to F_{LFIC} and to fuse these retrieved features with



(a) The Texture Learner Module

(b) The Self-Texture Fusion Module

(c) Coordinate Diagram of F_{STF} and F_{TL}

Fig. 5. (**a**) Texture learner; (**b**) Self-texture fusion module; (**c**) Coordinate diagram of F_{STF} and F_{TL} for the local pixel decoder and local texture decoder.

F_{LFIC}, thus completing the feature-based texture enhancement. As shown in Fig. 5b, we use the features sampled from F_{LFIC} by nearest-neighbor interpolation as the query (Q) and use F_{TL} as the key (*K*) and value (*V*) of the cross-attention module. To retrieve the texture features that are most relevant to the feature F_{LFIC}, we first compute the similarity matrix *R* of *Q* and *K*, where each element $r_{i,j}$ of *R* is computed according to Eq. (4):

$$r_{i,j} = \left\langle \frac{q_i}{\|q_i\|}, \frac{k_j}{\|k_j\|} \right\rangle \tag{4}$$

where q_i represents an element of *Q*, and k_j represents an element of *K*. Then we obtain the coordinate index matrix *T* with the highest similarity to q_i in *K*. An element in *T* is $t_i = \arg\max_j (r_{i,j})$, and t_i represents the position coordinates of the texture feature k_j with the highest similarity to q_i in F_{TL}. We select the feature vector a_i with the highest similarity to each element in *Q* from *V* according to the coordinate index matrix *T* to obtain the retrieved texture feature *A*, which can be represented by $a_i = v_{t_i}$, where a_i is an element in *A* and v_{t_i} represents the element at the t_i-th position in *V*. To fuse the retrieved texture feature *A* with the feature F_{LFIC}, we first concatenate F_{LFIC} with *A* and obtain the aggregated feature *Z* through an MLP, where $Z = MLP(Concat(F_{LFIC}, A))$. Finally, we calculate the soft attention map *S*, where an element s_i in *S* represents the confidence of each element a_i in the retrieved texture feature *A*, and $s_i = \max_j (r_{i,j})$. F_{STF} is calculated as Eq. (5):

$$F_{STF} = F_{LFIC} \oplus Z \otimes S \tag{5}$$

where $\langle \cdot \rangle$ represents inner product operation, $\| \cdot \|$ represents the square root operation, and $\oplus$ represents element-wise summation.

Spatial domain-based enhancement

In spatial domain-based texture enhancement, we decode the texture feature F_{TL} directly into the spatial domain I_{LTD} and add it to I_{LPD}, which is reconstructed from F_{FLIC} using the LPD, to obtain the final output I_{Pred}. First, we utilize the LPD to decode the feature F_{STF} into the RGB value I_{LPD}. We parameterize the LPD as an MLP f_θ. As shown in Fig. 5c, u_t denotes the coordinates of F_{LR}, while x_q denotes the coordinates of both F_{STF} and F_{TL}. We denote the upper-left, upper-right, lower-left, and lower-right coordinates of an arbitrary point x_q as $u_t (t \in 00, 01, 10, 11)$. The RGB value at coordinate x_q in the HR image decoded by the LPD can be represented as Eq. (6), where *c* consists of two elements, 2/*mh* and 2/*mw*, which represent the sizes of each pixel in I_{LPD}, and θ is the parameter of the MLP f_θ. Similarly, we calculate the RGB values of the texture information I_{LTD} at coordinate x_q via Eq. (7), where the LTD is parameterized as an MLP g_φ. We use the LTD to decode the texture features into the spatial domain texture information I_{LTD} and add it to the I_{LPD} via Eq. (8) for spatial domain texture enhancement to obtain the final prediction I_{Pred}, where φ is the parameter of the MLP g_φ. $S_t (t \in 00, 01, 10, 11)$ is the area of the rectangular region between x_q and u_t, and the weights are normalized by $S = \sum_{t \in \{00,01,10,11\}} S_t$.

$$I_{LPD} = \sum_{t \in \{00,01,10,11\}} \frac{S_t}{S} \cdot f_\theta \left(F_{STF}, x_q - u_t, c \right) \tag{6}$$

		In-distribution						Out-of-distribution			
		×2		×3		×4		×6		×8	
Dataset	Methods	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
TCGA	Bicubic	32.98±0.962	0.9353±0.0127	28.12±0.858	0.8070±0.0271	25.63±0.844	0.6874±0.0345	23.05±0.873	0.5354±0.0401	21.64±0.913	0.4606±0.0438
	EDSR[17]	36.14±0.962	0.9709±0.0063	31.16±0.914	0.9010±0.0183	28.01±0.840	0.8074±0.0278	–	–	–	–
	RDN[31]	36.72±0.962	0.9732±0.0059	31.76±0.911	0.9076±0.0171	28.52±0.807	0.8190±0.0258	–	–	–	–
	SwinIR[41]	36.73±0.971	0.9731±0.0058	31.77±0.895	0.9094±0.0167	28.83±0.813	0.8258±0.0251	–	–	–	–
	SRMFENet[47]	36.26±0.941	0.9713±0.0061	31.42±0.910	0.9033±0.0180	28.42±0.836	0.8140±0.0268	–	–	–	–
	Li et al.[14]	34.61±0.842	0.9580±0.0073	29.89±0.816	0.8725±0.0188	26.57±0.769	0.7358±0.0280	–	–	–	–
	SWD-Net[12]	36.76±0.965	0.9734±0.0058	31.73±0.914	0.9074±0.0172	28.85±0.864	0.8219±0.0260	–	–	–	–
	LIIF[15]	36.92±0.957	0.9742±0.0055	31.99±0.911	0.9110±0.0163	29.08±0.866	0.8275±0.0251	25.55±0.829	0.6641±0.0349	23.72±0.859	0.5609±0.0398
	LTE[29]	36.99±0.975	0.9748±0.0056	31.98±0.908	0.9109±0.0164	29.11±0.866	0.8280±0.0250	25.52±0.823	0.6617±0.0349	23.67±0.853	0.5580±0.0398
	LMI[45]	36.66±0.940	0.9726±0.0057	31.81±0.907	0.9093±0.0166	28.93±0.864	0.8251±0.0254	25.50±0.836	0.6657±0.0350	23.69±0.860	0.5614±0.0394
	ITSRN[46]	36.69±0.939	0.9728±0.0057	31.73±0.905	0.9078±0.0169	28.87±0.868	0.8231±0.0258	25.49±0.839	0.6665±0.0350	23.74±0.864	0.5626±0.0392
	LIT[44]	36.68±0.956	0.9733±0.0057	31.43±0.901	0.9077±0.0164	28.49±0.842	0.8220±0.0246	25.39±0.828	0.6669±0.0339	23.71±0.857	0.5635±0.0387
	ISTE(ours)	**37.76±1.034**	**0.9796±0.0050**	**32.06±0.914**	**0.9124±0.0163**	**29.19±0.867**	**0.8307±0.0247**	**25.61±0.821**	**0.6674±0.0342**	**23.76±0.856**	**0.5637±0.0395**
TMA	Bicubic	28.54±2.890	0.8931±0.0474	25.25±2.932	0.7708±0.1004	23.43±2.915	0.6735±0.1407	21.50±2.868	0.5647±0.1839	20.44±2.849	0.5123±0.2042
	EDSR[17]	30.54±2.792	0.9370±0.0272	26.38±2.880	0.8228±0.0782	24.94±2.884	0.7652±0.1014	–	–	–	–
	RDN[31]	31.02±2.705	0.9422±0.0255	28.07±2.930	0.8749±0.0571	25.97±2.869	0.8027±0.0884	–	–	–	–
	SwinIR[41]	31.20±2.747	0.9438±0.0247	28.18±2.939	0.8773±0.0563	26.26±2.954	0.8092±0.0868	–	–	–	–
	SRMFENet[47]	30.87±2.812	0.9399±0.0262	27.80±2.916	0.8689±0.0593	25.86±2.892	0.7965±0.0911	–	–	–	–
	Li et al.[14]	29.50±2.754	0.9211±0.0334	26.09±2.801	0.8207±0.0779	24.06±2.770	0.7206±0.1211	–	–	–	–
	SWD-Net[12]	31.18±2.832	0.9430±0.0251	28.06±2.946	0.8746±0.0574	26.09±2.934	0.8024±0.0894	–	–	–	–
	LIIF[15]	30.76±±2.562	0.9422±0.0253	27.84±2.794	0.8745±0.0572	25.87±2.858	0.7990±0.0908	23.50±2.886	0.6751±0.1425	22.05±2.874	0.5954±0.1741
	LTE[29]	31.26±2.834	0.9434±0.0250	28.19±2.949	0.8784±0.0558	26.22±2.975	0.8077±0.0875	23.73±2.958	0.6806±0.1409	22.17±2.926	0.5974±0.1738
	LMI[45]	31.25±2.831	0.9437±0.0248	28.05±2.936	0.8775±0.0558	26.15±2.965	0.8052±0.0880	23.64±2.941	0.6793±0.1404	22.12±2.907	0.5961±0.1732
	ITSRN[46]	30.64±2.233	0.9430±0.0250	27.66±2.551	0.8729±0.0579	25.82±2.694	0.8014±0.0895	23.46±2.771	0.6796±0.1399	22.05±2.792	**0.5997±0.1716**
	LIT[44]	31.10±2.837	0.9422±0.0254	27.93±2.944	0.8715±0.0587	25.90±2.957	0.7940±0.0928	23.41±2.933	0.6661±0.1458	21.95±2.898	0.5868±0.1770
	ISTE(ours)	**31.27±2.828**	**0.9444±0.0243**	**28.23±2.954**	**0.8809±0.0547**	**26.46±2.979**	**0.8160±0.0842**	**23.86±2.963**	**0.6851±0.1393**	**22.19±2.931**	0.5965±0.1742

continued

Dataset	Methods	In-distribution						Out-of-distribution			
		×2		×3		×4		×6		×8	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
HistoSR	Bicubic	27.43±3.322	0.8585±0.0496	23.88±3.394	0.6999±0.0936	22.01±3.498	0.5770±0.1243	19.95±3.654	0.4259±0.1678	18.89±3.683	0.3529±0.1898
	EDSR[17]	31.53±3.185	0.9407±0.0243	27.81±3.261	0.8588±0.0559	25.76±3.218	0.7820±0.0853	–	–	–	–
	RDN[31]	31.50±3.199	0.9396±0.0243	27.92±3.258	0.8611±0.0554	25.89±3.307	0.7853±0.0825	–	–	–	–
	SwinIR[41]	31.51±3.213	0.9397±0.0243	27.89±3.167	0.8624±0.0551	25.90±3.213	0.7870±0.0822	–	–	–	–
	SRMFENet[47]	31.39±3.203	0.9383±0.0246	27.75±3.238	0.8566±0.0565	25.67±3.242	0.7774±0.0841	–	–	–	–
	Li et al.[14]	28.98±3.133	0.9024±0.0360	25.34±3.117	0.7843±0.0750	23.50±3.164	0.6893±0.0992	–	–	–	–
	SWD-Net[12]	31.49±3.216	0.9393±0.0243	27.87±3.253	0.8595±0.0559	25.78±3.268	0.7810±0.0841	–	–	–	–
	LIIF[15]	31.56±3.212	0.9399±0.0243	28.03±3.270	0.8639±0.0549	25.93±3.310	0.7862±0.0820	22.94±3.498	0.6279±0.1195	20.87±3.821	0.4889±0.1598
	LTE[29]	31.58±3.244	0.9403±0.0242	28.03±3.286	0.8647±0.0545	25.93±3.317	0.7872±0.0816	22.95±3.500	0.6298±0.1192	20.89±3.815	0.4909±0.1588
	LMI[45]	31.42±3.159	0.9388±0.0245	27.82±3.247	0.8594±0.0556	25.72±3.256	0.7780±0.0831	22.73±3.475	0.6178±0.1203	20.73±3.783	0.4811±0.1603
	ITSRN[46]	31.50±3.190	0.9395±0.0245	27.94±3.257	0.8621±0.0553	25.83±3.291	0.7835±0.0823	22.90±3.515	0.6327±0.1164	20.90±3.844	0.4924±0.1555
	LIT[44]	31.43±3.205	0.9387±0.0246	27.84±3.247	0.8599±0.0556	25.73±3.281	0.7798±0.0828	22.84±3.515	0.6317±0.1168	20.86±3.849	0.4933±0.1554
	ISTE(ours)	**31.65±3.252**	**0.9410±0.0239**	**28.14±3.299**	**0.8673±0.0540**	**26.05±3.327**	**0.7909±0.0813**	**23.01±3.508**	**0.6331±0.1186**	**20.94±3.828**	**0.s4948±0.1586**

Table 1. Quantitative comparisons on the TMA, TCGA, and HistoSR datasets. The best results are indicated in bold.

$$I_{LTD} = \sum_{t \in \{00,01,10,11\}} \frac{S_t}{S} \cdot g_{\varphi}\left(F_{TL}\right) \tag{7}$$

$$I_{Pred} = I_{LPD} + I_{LTD} \tag{8}$$

Experiments

In this section, we introduce the datasets, implementation details, and compare our ISTE with other SR methods. Finally, we conduct a series of ablation studies to validate the effectiveness of each component in the proposed ISTE.

Datasets

In terms of experimental data, this paper utilize three publicly available datasets: (1) Tissue Microarray (TMA) dataset: Following Li et al.[14], we experimented on the TMA dataset to validate our method. The TMA dataset, a widely used public dataset in pancreatic cancer research[51,52], was scanned by an Aperio AT digital pathology scanner (Leica Biosystems, Wetzlar, Germany) at a magnification of 0.504 µm/pixel and contains 573 WSIs (average 3850 × 3850 pixels each). We randomly selected 460 WSIs as the training set, 57 WSIs as the validation set, and 56 WSIs as the test set. (2) Histopathology Super-Resolution (HistoSR) dataset: Following Chen et al.[12], we conducted experiments on the Histopathology Super-Resolution (HistoSR) dataset, which is built on the high-quality H&E stained WSIs of the Camelyon16 dataset[53]. The HistoSR dataset contains HR images with a patch size of 192 × 192 through random cropping. The training set comprises 30,000 HR patches, while the test set consists of 5000 HR patches. (3) TCGA Lung Cancer dataset: The TCGA lung cancer dataset[54] comprises 1054 WSIs (average 100,000 × 100,000 pixels each) from The Cancer Genome Atlas (TCGA) data center[55]. We selected five slides from this dataset and cut them into 400 sub-images with a size of 3072 × 3072. We randomly selected 320 sub-images as the training set, 40 as the validation set, and 40 as the test set.

Implementation details and evaluation metrics

Following previous SR methods based on implicit neural representation[15,29], we used the patches with the size of 48 × 48 as the input for training. We first randomly sampled the scaling factor m in a uniform distribution U(1, 4) and cropped patches with the size of $48m$ × 48 m from the ground truth HR images in a batch, where m represents the scaling factor. Following Li et al.[14], we resized the patches to 48 × 48 via bicubic downsampling and did a Gaussian blur to simulate degradation since it is difficult to acquire authentically downsampled images at arbitrary scales through scanners. The size of the Gaussian kernel was set to 1/2 of the scaling factor m. We sampled 48^2 pixels from the corresponding cropped patches to form RGB-Coordinate pairs. We utilized the deep learning toolbox Pytorch to implement ISTE and Adam as the optimizer, setting the initial learning rate to

	×1.5		×2.4		×3.3		×4.2		×5.1	
TCGA	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**
LIIF[15]	42.95±0.938	0.9962±0.0010	34.60±0.940	0.9532±0.0096	30.08±0.858	0.8777±0.0197	27.92±0.832	0.8018±0.0266	26.64±0.821	0.7285±0.0315
LTE[29]	43.34±0.951	0.9968±0.0009	34.61±0.943	0.9532±0.0096	30.08±0.858	0.8775±0.0197	27.93±0.832	0.8017±0.0266	26.62±0.814	0.7267±0.0316
LMI[45]	42.24±0.909	0.9951±0.0012	34.47±0.933	0.9522±0.0097	29.96±0.855	0.8759±0.0200	27.79±0.832	0.7999±0.0269	26.57±0.829	0.7300±0.0317
ITSRN[46]	42.38±0.913	0.9954±0.0012	34.41±0.928	0.9517±0.0098	29.93±0.854	0.8744±0.0203	27.73±0.834	0.7970±0.0274	26.54±0.833	0.7287±0.0319
LIT[44]	42.39±0.919	0.9954±0.0011	34.22±0.938	0.9515±0.0098	29.56±0.848	0.8732±0.0196	27.46±0.826	0.7976±0.0262	26.36±0.824	0.7308±0.0308
ISTE(ours)	**44.46±0.895**	**0.9982±0.0006**	**34.91±0.985**	**0.9568±0.0094**	**30.14±0.859**	**0.8791±0.0196**	**28.02±0.834**	**0.8053±0.0263**	**26.71±0.815**	**0.7312±0.0309**
	×1.5		×2.4		×3.3		×4.2		×5.1	
TMA	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**
LIIF[15]	32.47±2.401	0.9614±0.0145	29.45±2.675	0.9182±0.0368	26.63±2.773	0.8438±0.0705	25.17±2.789	0.7804±0.0983	24.31±2.850	0.7246±0.1222
LTE[29]	32.70±2.744	0.9611±0.0144	29.82±2.891	0.9203±0.0360	26.93±2.891	0.8487±0.0686	25.48±2.877	0.7889±0.0952	24.60±2.931	0.7321±0.1197
LMI[45]	32.77±2.744	0.9612±0.0144	29.83±2.887	0.9204±0.0360	26.90±2.885	0.8483±0.0686	25.39±2.871	0.7863±0.0956	24.49±2.917	0.7304±0.1194
ITSRN[46]	32.34±2.056	0.9610±0.0143	29.36±2.404	0.9189±0.0367	26.58±2.603	0.8453±0.0698	25.11±2.657	0.7817±0.0974	24.26±2.725	0.7285±0.1200
LIT[44]	32.76±2.743	0.9616±0.0147	29.68±2.895	0.9172±0.0373	26.71±2.890	0.8400±0.0722	25.23±2.871	0.7751±0.1006	24.26±2.913	0.7169±0.1252
ISTE(ours)	**32.80±2.745**	**0.9620±0.0156**	**29.85±2.900**	**0.9206±0.0358**	**27.00±2.889**	**0.8531±0.0667**	**25.67±2.874**	**0.7970±0.0919**	**24.80±2.925**	**0.7400±0.1167**
	×1.5		×2.4		×3.3		×4.2		×5.1	
HistoSR	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**
LIIF[15]	34.72±2.967	0.9722±0.0104	29.86±3.245	0.9111±0.0360	26.38±3.241	0.8281±0.0651	24.61±3.262	0.7678±0.0871	23.51±3.330	0.6922±0.1052
LTE[29]	34.78±3.025	0.9726±0.0104	29.87±3.265	0.9116±0.0358	26.38±3.249	0.8292±0.0646	24.60±3.272	0.7688±0.0868	23.50±3.341	0.6930±0.1050
LMI[45]	34.58±2.918	0.9718±0.0105	29.70±3.195	0.9089±0.0365	26.26±3.205	0.8238±0.0658	24.63±3.227	0.7598±0.0878	23.33±3.314	0.6833±0.1054
ITSRN[46]	34.60±2.943	0.9719±0.0105	29.76±3.218	0.9084±0.0371	26.33±3.226	0.8240±0.0664	24.63±3.249	0.7650±0.0874	23.43±3.335	0.6947±0.1036
LIT[44]	34.58±2.992	0.9714±0.0107	29.67±3.229	0.9071±0.0374	26.25±3.227	0.8215±0.0668	24.65±3.246	0.7615±0.0878	23.38±3.328	0.6915±0.1039
ISTE(ours)	**34.83±3.033**	**0.9728±0.0104**	**29.96±3.273**	**0.9131±0.0354**	**26.45±3.253**	**0.8317±0.0641**	**24.66±3.246**	**0.7720±0.0863**	**23.56±3.338**	**0.6962±0.1046**

Table 2. Quantitative comparisons at non-integer scales. The best results are indicated in bold.

Dataset	Methods	FID score						
		×1.5	×2	×2.4	×3	×3.3	×4	×4.2
TCGA	LIIF[15]	0.39	1.23	1.74	2.92	4.99	17.58	24.08
	LTE[29]	0.36	1.22	1.74	2.96	5.05	17.22	24.23
	LMI[45]	0.40	1.28	1.73	2.91	5.29	18.23	24.52
	ITSRN[46]	0.39	1.25	1.71	3.33	5.46	19.21	26.04
	LIT[44]	0.41	1.27	1.87	3.49	5.35	18.40	25.55
	ISTE(ours)	**0.30**	**1.07**	**1.66**	**2.86**	**4.91**	**16.45**	**22.83**
TMA	LIIF[15]	2.47	3.63	4.45	6.11	8.41	17.14	19.90
	LTE[29]	1.93	3.15	3.89	5.39	7.41	15.40	18.41
	LMI[45]	1.95	3.11	3.90	5.34	7.34	15.18	17.97
	ITSRN[46]	2.36	3.52	4.40	6.14	7.79	15.77	18.79
	LIT[44]	2.00	3.19	3.94	5.33	7.18	15.98	19.68
	ISTE(ours)	**1.88**	**2.77**	**3.42**	**4.74**	**6.47**	**13.53**	**16.72**
HistoSR	LIIF[15]	2.07	9.24	18.50	39.00	50.45	76.69	84.89
	LTE[29]	2.13	9.54	18.99	39.05	51.18	77.06	85.24
	LMI[45]	2.14	9.14	18.49	37.83	49.78	76.22	84.23
	ITSRN[46]	2.08	9.32	17.99	40.53	51.38	79.98	87.85
	LIT[44]	2.16	8.96	17.96	37.89	50.45	76.29	85.81
	ISTE(ours)	**2.04**	**8.92**	**17.92**	**37.82**	**49.40**	**75.45**	**83.76**

Table 3. Comparisions of FID scores. The best results are indicated in bold.

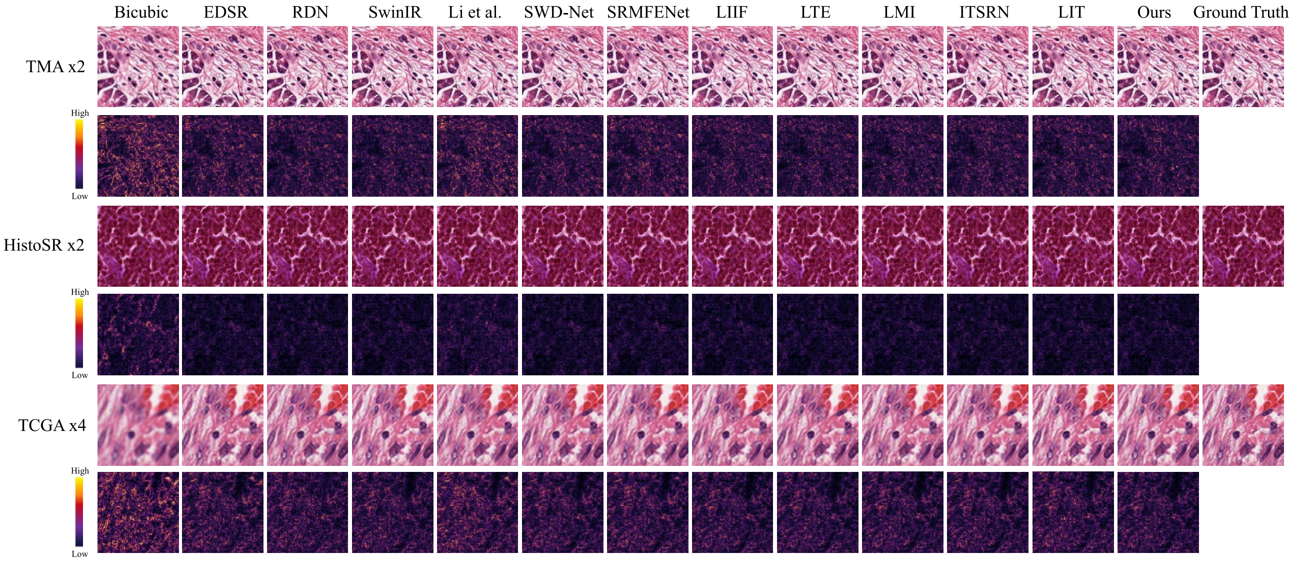


Fig. 6. Visual comparison with error maps of different methods on the TMA, HistoSR, and TCGA datasets. The error map represents the absolute error value between the reconstructed images and the ground truth. The brighter the color, the greater the error.

0.0001 and epochs to 1000. We employed structure similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR) to evaluate the quality of reconstructed images. The PSNR and SSIM are given by:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(I_{Pred}^{i}, Y_{HR}^{i}\right) \tag{9}$$

$$PSNR = 10 \times \log\left(\frac{255^2}{MSE}\right) \tag{10}$$

$$\text{SSIM}\left(I_{Pred}, Y_{HR}\right) = \frac{\left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{11}$$

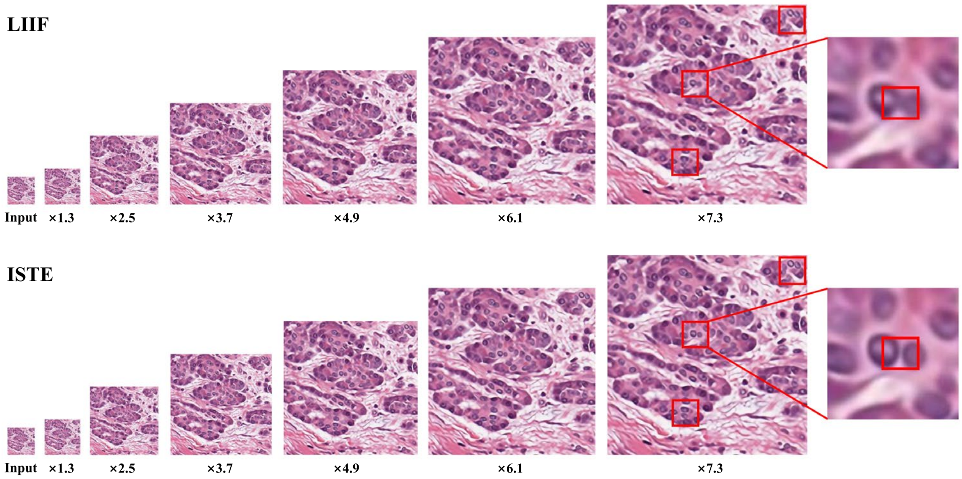


Fig. 7. Comparison of LIIF (upper row) and our ISTE (lower row) at non-integer scales.

Model							×2		×3		×4	
Dual-Branch	Single-Branch	TL	LFI	STF	LTD	LPD	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
×	✓	✓	×	×	✓	×	37.45±1.041	0.9778±0.0053	32.02±0.910	0.9115±0.0163	29.14±0.866	0.8290±0.0248
×	✓	×	✓	×	×	✓	37.44±1.032	0.9778±0.0053	32.01±0.910	0.9115±0.0163	29.14±0.866	0.8290±0.0248
✓	×	✓	✓	✓	×	✓	37.63±1.041	0.9789±0.0052	32.04±0.912	0.9120±0.0163	29.17±0.867	0.8302±0.0248
✓	×	✓	✓	×	✓	✓	37.66±1.037	0.9791±0.0051	32.04±0.913	0.9121±0.0163	29.17±0.867	0.8301±0.0248
✓	×	×	✓	✓	✓	✓	37.64±1.039	0.9790±0.0051	32.04±0.913	0.9121±0.0163	29.17±0.867	0.8301±0.0248
✓	×	✓	✓	✓	✓	✓	37.61±1.037	0.9788±0.0052	32.04±0.911	0.9121±0.0163	29.18±0.867	0.8303±0.0248
✓	×	✓	✓	✓	✓	✓	**37.76±1.034**	**0.9796±0.0050**	**32.06±0.914**	**0.9124±0.0163**	**29.19±0.867**	**0.8307±0.0247**

Table 4. Ablation study on the TCGA dataset. The best results are indicated in bold.

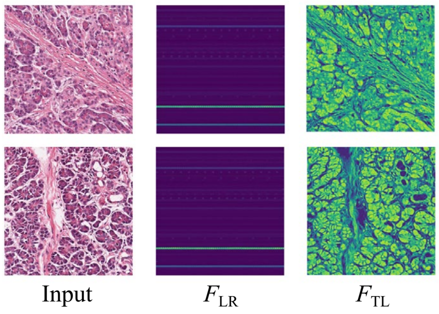


Fig. 8. Feature map visualization for the texture learner. F_{LR} represents the feature map input to the texture learner and F_{TL} represents the feature map output from the texture learner.

where I_{Pred} and Y_{HR} are the generated image and the ground truth image, respectively. *i* represents the index of the *i*-th pixel of the image, and *N* is the total number of the pixels in the image. μ_x, σ_x and σ_{xy} are the mean standard deviation and covariance, respectively.

Comparison with previous methods

We compared the performance of ISTE with state-of-the-art SR methods in both the pathological image domain: SWD-Net[12], SRMFENet[47] and Li et al.[14], and the natural image domain: Bicubic, EDSR[17], RDN[31], SwinIR[41], LIIF[15], LTE[29], LMI[45], ITSRN[46] and LIT[44], where the latter five are arbitrary-scale SR methods. For a fair comparison, the encoder used for arbitrary-scale SR methods is SwinIR[41] without the last upsampling layer.

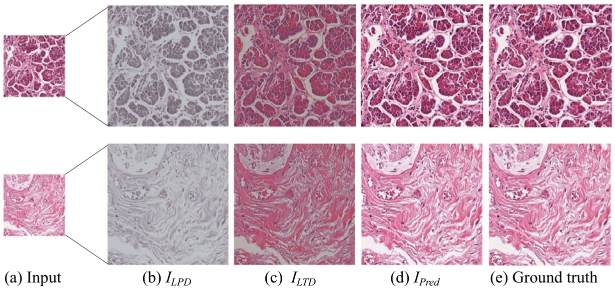


Fig. 9. (**a**) Input LR image; (**b**) Pixel information decoded from the LPD; (**c**) Texture information decoded from the LTD; (**d**) Output of the spatial domain-based enhancement; (**e**) Ground truth.

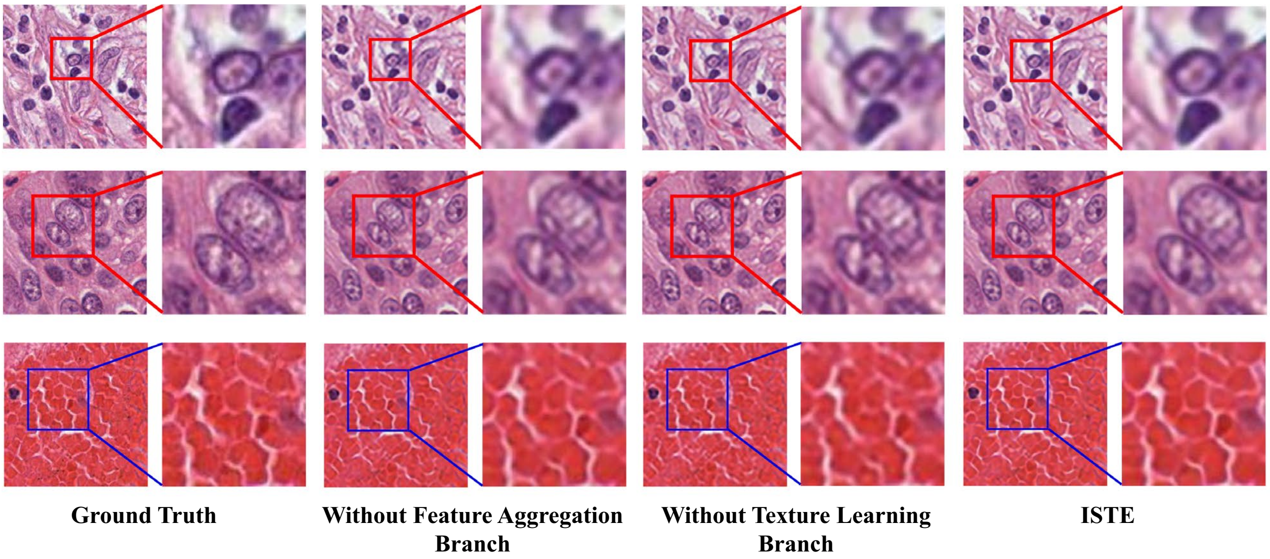

Ground Truth
Without Feature Aggregation Branch
Without Texture Learning Branch
ISTE

Fig. 10. Qualitative analysis of ablation experiments with the feature aggregation branch and the texture learning branch.

Quantitative results

We compared our ISTE with previous SR methods at five scaling factors of ×2, ×3, ×4, ×6, and ×8. As shown in Table 1, our ISTE achieved the highest performance in terms of PSNR and SSIM metrics at each scaling factor on the HistoSR and TCGA datasets. Although the SSIM metric for our method at ×8 scale is slightly lower than that of LTE[29] by 0.0009 on the TMA dataset, it outperforms the comparison methods in PSNR metrics at all scaling factors and in SSIM metrics at the other scaling factors. To substantiate our results, we evaluate the significant difference between our ISTE and other methods using paired Student's t-tests. Our ISTE method shows statistically significant differences compared to the comparison methods in almost all cases, with a p-value smaller than 0.001. The only exception is in the HistoSR dataset at the ×2 scale, where the significance test with EDSR on the SSIM metric yields a p-value slightly greater than 0.001 but still smaller than 0.05. It is worth noting that our method still demonstrates a statistically significant improvement over EDSR. To further assess the advantages of our method over other arbitrary-scale SR Methods, we present comparative results in Table 2 for ISTE, LIIF[15], LTE[29], LMI[45], ITSRN[46] and LIT[44] at non-integer scaling factors. Our method demonstrates superior performance in terms of both PSNR and SSIM metrics. We also provide the Frechet Inception Distance (FID) score metric to evaluate the perceptual quality of images generated by different methods, as shown in Table 3. It can be observed that our method outperforms the comparative methods in terms of FID. The results indicate that the textures of images generated by our method are more realistic, yielding perceptual effects superior to those of other arbitrary-scale SR methods.

Qualitative results

Figure 6 shows the visual results and absolute error maps of different methods on the TCGA datasets at the scale of ×4, TMA datasets at the scale of ×2, and HistoSR datasets at the scale of ×2. The proposed ISTE performs better in restoring texture information, closely approximating the ground truth. Based on the brightness levels

Experiment	F1		ObjDice		ObjHausdorff	
	Test A	Test B	Test A	Test B	Test A	Test B
Bicubic	0.71	0.85	0.83	0.88	133.73	109.21
HR U-Net	0.84	0.88	0.89	0.92	100.57	84.64
SISR	0.92	0.93	0.94	0.95	77.74	65.81
Original high resolution	0.95	0.93	0.96	0.96	66.70	61.17

Table 5. Gland segmentation on the GlaS dataset under different experimental settings.

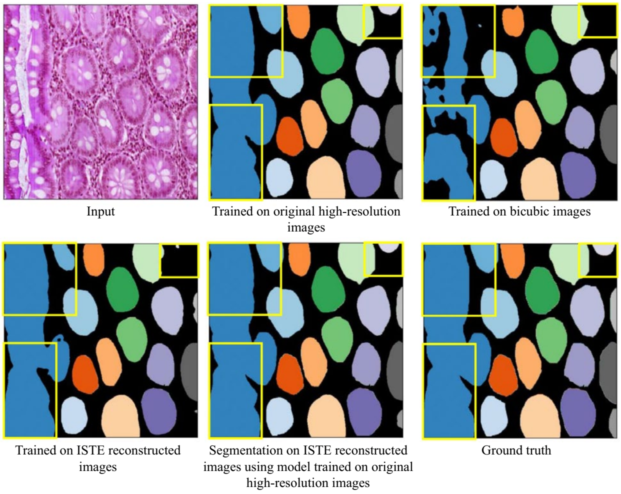

Input
Trained on original high-resolution images
Trained on bicubic images
Trained on ISTE reconstructed images
Segmentation on ISTE reconstructed images using model trained on original high-resolution images
Ground truth

Fig. 11. Qualitative evaluation of UNet for gland segmentation on the GlaS dataset[57] with different experiment setups.

in the absolute error maps, it is observable that our method's error maps contain more dark regions, indicating more minor errors in the reconstructed results compared to other methods. Figure 7 shows an example of a comparison of LIIF and our ISTE at non-integer scales. It can be seen that ISTE achieves arbitrary-scale SR with clear cell structure and texture. As shown in the red box, two cells are connected due to blurring in the image generated by LIIF while they are still separated in the image generated by ISTE at the scale of ×7.3. Please refer to supplementary figures for more comparisons.

Ablation study

To validate the effectiveness of each module in our proposed method, including the LFI, TL, STF, and LTD, we designed several variant networks for ablation experiments at scaling factors of ×2, ×3, and ×4 on the TCGA dataset, as shown in Table 4. To substantiate our results, we evaluate the significance of the differences between our proposed method and other variant networks using paired Student's t-tests. $P < 0.001$ was considered as a statistically significant level. We observe statistically significant differences with p-values smaller than 0.001 in all cases.

Evaluation of the local feature interactor

For the features obtained from the encoder F_{LR}, the LFI module enhances feature interaction within local regions. To investigate the effectiveness of the LFI module, we conducted an ablation experiment by directly removing the LFI module from the ISTE framework. As shown in Table 4, all metrics improve across all scaling factors when using the LFI.

Evaluation of the texture learner

The TL module is employed to enhance the learning of high-frequency textures in histopathology images. To investigate the effectiveness of this module, we conducted an ablation experiment by replacing the module with a convolutional layer. As shown in Table 4, it can be seen that after ablating the TL module, all metrics become worse at all scaling factors. To better illustrate the role of the TL module, we visualized the features input to and output from the TL, denoted as F_{LR} and F_{TL}, respectively, in Fig. 8. Compared to F_{LR}, the output feature map F_{TL} from the TL module contains richer texture information.

Evaluation of the self-texture fusion module

The STF module globally retrieves texture features that are most similar to F_{LFIC} in F_{TL} and fuses the retrieved features to F_{LFIC}. We designed a variant network without the STF module to evaluate its effectiveness.

Experiment	Accuracy	F1 score
Original	86.17%	0.8507
Low resolution	58.11%	0.2929
Bicubic	77.09%	0.7419
LIIF	80.54%	0.7721
ISTE(ours)	**81.15%**	**0.7816**

Table 6. The performance promotion using different SR methods in cancer detection. The best results are indicated in bold.

Dataset	Encoder	×2		×3		×4	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
TCGA	EDSR	36.75±0.955	0.9730±0.0058	31.92±0.911	0.9098±0.0167	29.00±0.870	0.8257±0.0254
	RDN	37.05±0.974	0.9751±0.0055	31.99±0.910	0.9110±0.0164	29.09±0.869	0.8281±0.0251
	SwinIR	**37.76±1.034**	**0.9796±0.0050**	**32.06±0.914**	**0.9124±0.0163**	**29.19±0.867**	**0.8307±0.0247**
TMA	EDSR	31.23±2.833	0.9432±0.0251	28.15±2.947	0.8775±0.0562	26.18±2.965	0.8062±0.0880
	RDN	31.22±2.787	0.9436±0.0248	28.16±2.929	0.8787±0.0555	26.22±2.958	0.8079±0.0872
	SwinIR	**31.27±2.828**	**0.9444±0.0243**	**28.23±2.954**	**0.8809±0.0547**	**26.46±2.979**	**0.8160±0.0842**

Table 7. The performance of ISTE with different encoders.

Specifically, we first take the feature F_{LFIC} obtained from the feature aggregation branch of the framework and decode it directly through the LPD to obtain I'_{LPD}. Then, we take the feature F_{TL} obtained from the texture learning branch and decode it through the LTD to obtain I'_{LTD}. We sum I'_{LPD} and I'_{LTD} to get the output I'_{Pred} of the variant network. As shown in Table 4, all metrics become worse at all scaling factors after ablating the STF module.

Evaluation of the texture decoder for spatial domain-based enhancement

The feature F_{STF} is decoded into the pixel information I_{LPD} by the LPD in the spatial domain. To accomplish spatial domain-based texture enhancement in the subsequent stage, LTD is employed to decode texture features F_{TL} directly into texture information I_{LTD} in the spatial domain, and we sum I_{LTD} and I_{LPD} to obtain I_{Pred}. To demonstrate the effectiveness of the designed spatial domain-based enhancement strategy, we removed the LTD from the ISTE framework and used only the pixels decoded by the LPD for the final prediction. The results in Table 4 suggest that incorporating spatial domain-based texture enhancement leads to improved results. To better illustrate the effectiveness of the spatial domain-based enhancement, we visualized the pixel information decoded by the LPD and the texture information decoded by the LTD in Fig. 9. It can be seen that the texture information I_{LTD} decoded from the LTD reveals clear outlines and texture features of the tissue cells and has more vibrant colors. This further illustrates the importance of LTD for spatial domain-based enhancement.

Evaluation of the dual-branch architecture

To further assess the effectiveness of the feature aggregation branch and texture learning branch in the proposed framework, we designed two single-branch variant networks: (1) retaining only the TL and LTD in the ISTE framework, which represents the ablation of the feature aggregation branch, and (2) retaining only the LFI and LPD in the ISTE framework, representing the ablation of the texture learning branch. As shown in Table 4, the proposed dual-branch architecture ISTE outperforms both single-branch variants, demonstrating the effectiveness of the feature aggregation and texture learning branches. Additionally, we provide visual comparisons in Fig. 10 to further validate the effectiveness of the proposed dual-branch framework. As shown in the first row, when the feature aggregation branch is removed, the reconstructed images show the loss of cellular boundaries. In the second and third rows, when the texture learning branch is removed, the model struggles to recover high-frequency details, such as intercellular gaps. In contrast, the complete dual-branch ISTE framework successfully reconstructs both cellular structures and intercellular gaps, further illustrating the effectiveness of the feature aggregation branch in capturing local details and the texture learning branch in reconstructing high-frequency textures.

Discussion

Applications in downstream pathology image analysis tasks

It is important to evaluate whether the images generated by the proposed ISTE in this paper can be used for pathology image analysis tasks. We demonstrate experimentally that ISTE effectively enhances the performance of two downstream tasks: gland segmentation and cancer detection. First, for gland segmentation, we trained and tested the state-of-the-art segmentation model U-Net[56] on the Glas dataset from the MICCAI 2015 Gland Segmentation Challenge[57]. The Glas dataset includes a training set and two test sets, Test A and Test B. The training set contains 85 labeled images, Test A contains 60 labeled images, and Test B contains 20 labeled images.

Query	Method	Params	Mem (GB)	Time (ms)
48 × 48	LIIF	11.4M	1.5	129.5
	LTE	11.5M	1.5	150.0
	LMI	11.1M	1.5	193.6
	ITSRN	11.7M	1.5	158.7
	LIT	16.1M	2.1	159.1
	ISTE (ours)	12.5M	1.8	407.3
96 × 96	LIIF	11.4M	1.6	99.5
	LTE	11.5M	1.5	97.8
	LMI	11.1M	1.9	126.9
	ITSRN	11.7M	1.7	122.2
	LIT	16.1M	3.7	158.6
	ISTE (ours)	12.5M	2.5	205.4
192 × 192	LIIF	11.4M	1.8	85.9
	LTE	11.5M	1.7	87.5
	LMI	11.1M	3.2	95.1
	ITSRN	11.7M	2.2	106.4
	LIT	16.1M	10.3	93.1
	ISTE (ours)	12.5M	12.7	104.5

Table 8. Comparisons of computational consumption for different methods.

Method	Params	×2		×3		×4	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
LIIF	11.4	36.92±0.957	0.9742±0.0055	31.99±0.911	0.9110±0.0163	29.08±0.866	0.8275±0.0251
LTE	11.5	36.99±0.975	0.9748±0.0056	31.98±0.908	0.9109±0.0164	29.11±0.866	0.8280±0.0250
LIIF*	13.2	36.91±0.960	0.9741±0.0056	31.98±0.908	0.9110±0.0163	29.08±0.866	0.8273±0.0251
LTE*	13.3	37.03±0.982	0.9751±0.0055	31.99±0.908	0.9111±0.0163	29.11±0.867	0.8283±0.0249
ISTE(ours)	12.5	**37.76±1.034**	**0.9796±0.0050**	**32.06±0.914**	**0.9124±0.0163**	**29.19±0.867**	**0.8307±0.0247**

Table 9. Comparisons of ISTE with LIIF and LTE after increasing the number of parameters.

We performed ×4 downsampling on the HR images to generate LR images using bicubic interpolation. We compared segmentation results under the following settings: (1) Original high-resolution: Train U-Net on the original HR GlaS dataset for segmentation of original high-resolution images; (2) SISR: Directly employing U-Net trained on the original HR GlaS dataset for segmentation of the reconstructed images produced by our ISTE; (3) HR U-Net: Train U-Net on the reconstructed images produced by our ISTE for segmentation of original HR images; (4) Bicubic: Train U-Net on LR images obtained by bicubic interpolation for segmentation of original HR images. Table 5 shows the quantitative test results, where larger values indicate better performance for the F1 score and Object Dice score, while smaller values indicate better performance for object Hausdorff distance. It can be seen that the U-Net model trained on the reconstructed images from our ISTE performs better than the U-Net model trained on the LR image dataset, showing higher F1 scores and object Dice scores, as well as lower object Hausdorff distances. In particular, when evaluated on the Test B dataset, our results for segmentation of reconstructed images using U-Net trained on the original HR GlaS training set are close to those for segmentation of the original HR image, both with an F1 score of 0.93. Figure 11 shows representative results for different experimental setups, and we observe that the U-Net trained on LR images produced the worst results, it not only failed to detect small glands but also produced poor segmentation results for large glands. In contrast, the U-Net trained on the reconstructed images effectively outlined the boundaries of the macro glands and detected the tiny glands. Compared to using LR images for training, utilizing the generated SR images can improve segmentation accuracy when evaluating.

To further evaluate the contribution of our ISTE to the cancer detection task, we conducted tumor recognition on the PCam dataset[58]. The PCam dataset comprises 262,144 color images for training and 32,768 images for testing, with each image annotated with a binary label indicating the presence of metastatic tissue. We performed ×2 downsampling on HR images of the test set to generate LR images using bicubic interpolation. We chose ResNet-50[59] as the classifier and trained it on the original PCam dataset. We compared classification results across the following settings: (1) Original: Directly employing trained ResNet-50 model to test on the original HR images in the test set; (2) Low resolution: Directly employing trained ResNet-50 model to test on the LR images of the test set; (3) Bicubic: Directly employing the trained ResNet-50 model to test on the bicubic interpolated images of the test set; (4) LIIF: Directly employing trained ResNet-50 model to test on the images generated by LIIF from the LR test set images; (5) ISTE: Directly employing trained ResNet-50 model to test on the images generated by our ISTE from the LR test set images. Table 6 illustrates the diagnostic performance with different experiment setups. By introducing additional prior knowledge, our ISTE leads to a performance improvement, resulting in a 4.06% accuracy increase compared to the Bicubic method. These results indicate that ISTE can improve classification performance by recovering more distinctive details.

The impact of different encoders on ISTE

We studied the impact of different encoders on the performance of ISTE using the TCGA and TMA datasets. We conducted a comparison using three different encoders: RDN[31], EDSR[17], and SwinIR[41]. As shown in Table 7, ISTE with the SwinIR encoder achieved the best performance. Compared to EDSR[17] and RDN[31] which use convolutional neural networks, SwinIR[41] integrated with the Swin Transformer block can more effectively handle long-range dependencies, which is crucial for capturing subtle texture variations in histopathology images. Specifically, for histopathology images with fine textures and complex structures, SwinIR is able to capture these details more accurately and provides stronger feature representation capabilities.

Computational consumption analysis for ISTE

Finally, we compared the computational consumption of our ISTE with other arbitrary-scale SR methods using an NVIDIA RTX 3090 with 24GB of memory. All models used SwinIR[41] as the encoder. We employed LR images with the size of 96×96 as input, computing 48×48, 96×96, and 192×192 output pixels for each query. As shown in Table 8, our model has a slightly longer runtime and consumes relatively more memory than the other SR models and does not have a clear advantage in terms of lightweight design. To further demonstrate that the reconstruction performance of our method comes from the network design rather than an increase in the number of parameters, we added a simple number of swin transformer blocks to the internal encoders of the two baseline models, LTE and LIIF, without modifying the network after the encoders. This modification resulted in a higher number of parameters than our ISTE. We then compared them on the TCGA dataset. As shown in Table 9, our method still achieves higher PSNR and SSIM. LIIF* and LTE* represent the models with increased parameters. This indicates that our network design is effective, and we will continue to work towards developing more computationally efficient models in the future.

Conclusion

In this work, we propose an innovative dual-branch framework ISTE based on implicit self-texture enhancement for arbitrary-scale histopathology image super-resolution. ISTE consists of a feature aggregation branch and a texture learning branch. We employ the feature aggregation branch to enhance the relevance of features in the local region while utilizing the texture learning branch to improve the learning of high-frequency texture details. We then design a two-stage texture enhancement strategy to fuse the features from the two branches to obtain SR images, where the first stage is feature-based texture enhancement and the second stage is spatial domain-based texture enhancement. Extensive experiments on publicly available datasets show that ISTE outperforms existing fixed-scale and arbitrary-scale SR methods across multiple scaling factors. Further experiments indicate that our method can enhance performance on two downstream tasks. In the future, we will continue to work on computationally efficient models and integrate the proposed SR models with existing diagnostic networks to improve diagnostic performance.

Data availability

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Received: 19 October 2024; Accepted: 13 May 2025
Published online: 29 May 2025

References

1. Gilbertson, J. R. et al. Primary histologic diagnosis using automated whole slide imaging: a validation study. *BMC Clin. Pathol.* **6**, 1–19 (2006).
2. Pantanowitz, L. et al. Review of the current state of whole slide imaging in pathology. *J. Pathol. Inform.* **2**, 36 (2011).
3. Weinstein, R. S. et al. An array microscope for ultrarapid virtual slide processing and telepathology. design, fabrication, and validation study. *Hum. Pathol.* **35**, 1303–1314 (2004).
4. Wilbur, D. C. Digital cytology: current state of the art and prospects for the future. *Acta Cytol.* **55**, 227–238 (2011).
5. Ghaznavi, F., Evans, A., Madabhushi, A. & Feldman, M. Digital imaging in pathology: whole-slide imaging and beyond. *Annu. Rev. Pathol.* **8**, 331–359 (2013).
6. Wu, X., Chen, Z., Peng, C. & Ye, X. Mmsrnet: Pathological image super-resolution by multi-task and multi-scale learning. *Biomed. Signal Process. Control* **81**, 104428 (2023).
7. Nielsen, P. S. et al. Virtual microscopy: an evaluation of its validity and diagnostic performance in routine histologic diagnosis of skin tumors. *Hum. Pathol.* **41**, 1770–1776 (2010).
8. Madabhushi, A. & Lee, G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med. Image Anal.* **33**, 170–175 (2016).
9. Upadhyay, U. & Awate, S. P. A mixed-supervision multilevel gan framework for image quality enhancement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 556–564 (Springer, 2019).
10. Mukherjee, L., Keikhosravi, A., Bui, D. & Eliceiri, K. W. Convolutional neural networks for whole slide image superresolution. *Biomed. Opt. Express* **9**, 5368–5386 (2018).
11. Juhong, A. et al. Super-resolution and segmentation deep learning for breast cancer histopathology image analysis. *Biomed. Opt. Express* **14**, 18–36 (2023).
12. Chen, Z., Guo, X., Yang, C., Ibragimov, B. & Yuan, Y. Joint spatial-wavelet dual-stream network for super-resolution. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, 184–193 (Springer, 2020).

13. Shahidi, F. Breast cancer histopathology image super-resolution using wide-attention GAN with improved Wasserstein gradient penalty and perceptual loss. *IEEE Access* **9**, 32795–32809 (2021).
14. Li, B., Keikhosravi, A., Loeffler, A. G. & Eliceiri, K. W. Single image super-resolution for whole slide image using convolutional neural networks and self-supervised color normalization. *Med. Image Anal.* **68**, 101938 (2021).
15. Chen, Y., Liu, S. & Wang, X. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8628–8638 (2021).
16. Canny, J. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 679–698 (1986).
17. Lim, B., Son, S., Kim, H., Nah, S. & Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 136–144 (2017).
18. Jia, Y., Chen, G. & Chi, H. Retinal fundus image super-resolution based on generative adversarial network guided with vascular structure prior. *Sci. Rep.* **14**, 22786 (2024).
19. Shi, W. *et al.* Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part III 16*, 9–16 (Springer, 2013).
20. Thornton, M. W., Atkinson, P. M. & Holland, D. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. *Int. J. Remote Sens.* **27**, 473–491 (2006).
21. Zou, W. W. & Yuen, P. C. Very low resolution face recognition problem. *IEEE Trans. Image Process.* **21**, 327–340 (2011).
22. Zhang, Y., Zhou, P. & Chen, L. Dual-branch feature encoding framework for infrared images super-resolution reconstruction. *Sci. Rep.* **14**, 9379 (2024).
23. Hu, L., Hu, L. & Chen, M. Edge-enhanced infrared image super-resolution reconstruction model under transformer. *Sci. Rep.* **14**, 15585 (2024).
24. Li, G., Cui, Z., Li, M., Han, Y. & Li, T. Multi-attention fusion transformer for single-image super-resolution. *Sci. Rep.* **14**, 10222 (2024).
25. Wang, L., Li, X., Tian, W., Peng, J. & Chen, R. Lightweight interactive feature inference network for single-image super-resolution. *Sci. Rep.* **14**, 11601 (2024).
26. Sitzmann, V., Martel, J., Bergman, A., Lindell, D. & Wetzstein, G. Implicit neural representations with periodic activation functions. *Adv. Neural. Inf. Process. Syst.* **33**, 7462–7473 (2020).
27. Tancik, M. et al. Fourier features let networks learn high frequency functions in low dimensional domains. *Adv. Neural. Inf. Process. Syst.* **33**, 7537–7547 (2020).
28. Mildenhall, B. et al. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **65**, 99–106 (2021).
29. Lee, J. & Jin, K. H. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1929–1938 (2022).
30. Dong, C., Loy, C. C., He, K. & Tang, X. Learning a deep convolutional network for image super-resolution. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, 184–199 (Springer, 2014).
31. Zhang, Y., Tian, Y., Kong, Y., Zhong, B. & Fu, Y. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2472–2481 (2018).
32. Cavigelli, L., Hager, P. & Benini, L. Cas-cnn: A deep convolutional neural network for image compression artifact suppression. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 752–759 (IEEE, 2017).
33. Kim, J., Lee, J. K. & Lee, K. M. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1646–1654 (2016).
34. Wang, X. *et al.* Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (2018).
35. Zhang, Y., Tian, Y., Kong, Y., Zhong, B. & Fu, Y. Residual dense network for image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 2480–2495 (2020).
36. Chen, Y. & Pock, T. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1256–1272 (2016).
37. Deng, X., Zhang, Y., Xu, M., Gu, S. & Duan, Y. Deep coupled feedback network for joint exposure fusion and image super-resolution. *IEEE Trans. Image Process.* **30**, 3098–3112 (2021).
38. Niu, B. *et al.* Single image super-resolution via a holistic attention network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 191–207 (Springer, 2020).
39. Zhang, Y. *et al.* Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301 (2018).
40. Chen, H. *et al.* Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12299–12310 (2021).
41. Liang, J. *et al.* Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1833–1844 (2021).
42. Liu, D., Wen, B., Fan, Y., Loy, C. C. & Huang, T. S. Non-local recurrent network for image restoration. *Adv. Neural Inf. Process. Syst.* **31** (2018).
43. Mei, Y., Fan, Y. & Zhou, Y. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3517–3526 (2021).
44. Chen, H.-W. *et al.* Cascaded local implicit transformer for arbitrary-scale super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18257–18267 (2023).
45. Fu, H. *et al.* Continuous optical zooming: A benchmark for arbitrary-scale image super-resolution in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3035–3044 (2024).
46. Yang, J., Shen, S., Yue, H. & Li, K. Implicit transformer network for screen content image continuous super-resolution. *Adv. Neural. Inf. Process. Syst.* **34**, 13304–13315 (2021).
47. Xie, L. *et al.* Shisrcnet: Super-resolution and classification network for low-resolution breast cancer histopathology image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 23–32 (Springer, 2023).
48. Ma, J. et al. Stsrnet: Self-texture transfer super-resolution and refocusing network. *IEEE Trans. Med. Imaging* **41**, 383–393 (2021).
49. Zhang, Z., Wang, Z., Lin, Z. & Qi, H. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7982–7991 (2019).
50. Feng, C.-M., Yan, Y., Fu, H., Chen, L. & Xu, Y. Task transformer network for joint mri reconstruction and super-resolution. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, 307–317 (Springer, 2021).
51. Drifka, C. R. et al. Highly aligned stromal collagen is a negative prognostic factor following pancreatic ductal adenocarcinoma resection. *Oncotarget* **7**, 76197 (2016).
52. Drifka, C. R. et al. Periductal stromal collagen topology of pancreatic ductal adenocarcinoma differs from that of normal and chronic pancreatitis. *Mod. Pathol.* **28**, 1470–1480 (2015).
53. Bejnordi, B. E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).

54. Li, B., Li, Y. & Eliceiri, K. W. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328 (2021).
55. Weinstein, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
56. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241 (Springer, 2015).
57. Sirinukunwattana, K. et al. Gland segmentation in colon histology images: The glas challenge contest. *Med. Image Anal.* **35**, 489–502 (2017).
58. Veeling, B. S., Linmans, J., Winkens, J., Cohen, T. & Welling, M. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, 210–218 (Springer, 2018).
59. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant 82372097, 82072021, and 62471149).

Author contributions

M.D. designed the methodology, conducted the experiments, and wrote the manuscript. L.Q. contributed to writing the manuscript. Z.Y., M.W., C.Z., and Z.S. revised the manuscript critically for important intellectual content. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.Z. or Z.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.