Super-Resolution Enhanced Medical Image Diagnosis with Sample Affinity Interaction

Zhen Chen, Xiaoqing Guo, Peter Y. M. Woo, and Yixuan Yuan, *Member, IEEE*

***Abstract*— The degradation in image resolution harms the performance of medical image diagnosis. By inferring high-frequency details from low-resolution (LR) images, super-resolution (SR) techniques can introduce additional knowledge and assist high-level tasks. In this paper, we propose a SR enhanced diagnosis framework, consisting of an efficient SR network and a diagnosis network. Specifically, a Multi-scale Refined Context Network (MRC-Net) with Refined Context Fusion (RCF) is devised to leverage global and local features for SR tasks. Instead of learning from scratch, we first develop a recursive MRC-Net with temporal context, and then propose a recursion distillation scheme to enhance the performance of MRC-Net from the knowledge of the recursive one and reduce the computational cost. The diagnosis network jointly utilizes the reliable original images and more informative SR images by two branches, with the proposed Sample Affinity Interaction (SAI) blocks at different stages to effectively extract and integrate discriminative features towards diagnosis. Moreover, two novel constraints, sample affinity consistency and sample affinity regularization, are devised to refine the features and achieve the mutual promotion of these two branches. Extensive experiments of synthetic and real LR cases are conducted on wireless capsule endoscopy and histopathology images, verifying that our proposed method is significantly effective for medical image diagnosis.**

***Index Terms*— medical image diagnosis, super resolution, semantic consistency.**

I. Introduction

THE details of small pathologies and texture information around abnormalities are crucial to the diagnosis of clinical experts and computer-aided algorithms [1]. It is reported that a high definition colonoscopy can bring a 3.8% improvement to polyp detection compared with conventional resolution endoscopic examinations [2]. However, access to expensive high-end imaging equipment is limited in remote and impoverished areas where medical images have generally inferior spatial resolution. This could interfere with the early diagnosis of diseases that rely on reliable and accurate image interpretation. Single image super-resolution (SR) offers a feasible alternative to mitigate the resolution degradation by reconstructing high-resolution (HR) images from their corresponding low-resolution (LR) inputs. With the introduction of extra pixels, SR techniques compensate for the missing information of LR images. Recently, numerous deep learning based SR methods [3]–[11] have been investigated with various network structures or novel loss functions, and some studies have proven that pretrained SR networks promote the down-stream tasks, especially object detection [12], [13]. The elevation of SR models to detection tasks is readily achieved, as imperceptible tiny objects can be captured by networks after SR. However, promoting medical image diagnosis through SR techniques has not been well explored.

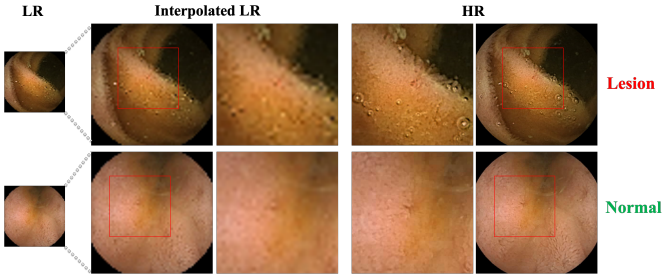

Fig. 1. The resolution degradation problem in medical image diagnosis. Examples are from CAD-CAP dataset [14]. Each row includes the LR image with $\mathbf{64} \times \mathbf{64}$ pixels, the $\mathbf{8}\times$ bicubic interpolated LR image and the HR image with $\mathbf{512} \times \mathbf{512}$ pixels from left to right.

We hypothesize that exploiting the detailed information introduced by SR techniques can improve the medical image diagnosis. The intuition behind this hypothesis is illustrated by wireless capsule endoscopy (WCE) images in Fig. 1. Small lesions depicted by LR images may not be perceived by clinical experts and algorithms. Even after interpolation, lesion and normal tissues exhibit similar confounding features at image space, which may confuse the diagnosis to misinterpret both samples as lesions. In contrast, HR images with more details are clear to distinguish the first sample as lesion and the second sample as the normal one. This comparison demonstrates that high-frequency details in image space can eliminate the ambiguity in inference and lead to more accurate and reliable diagnosis. However, previous studies directly applied trained SR networks as a pre-processing step [12], [13], [15], [16], which are far from satisfactory with two shortcomings.

The first problem is that state-of-the-art SR networks [4]–[7] require extremely high hardware resources. For example,

This work was supported by Hong Kong Research Grants Council (RGC) Early Career Scheme grant 21207420 (CityU 9048179), National Natural Science Foundation of China (62001410) and Shenzhen-Hong Kong Innovation Circle Category D Project SGDX2019081623300177 (CityU 9240008). *(Corresponding author: Yixuan Yuan)*

Z. Chen, X. Guo and Y. Yuan are with Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China (e-mail: {zchen.ee, xqguo.ee}@my.cityu.edu.hk, yxyuan.ee@cityu.edu.hk).

Peter Y. M. Woo is with Department of Neurosurgery, Kwong Wah Hospital, Hong Kong SAR, China (e-mail: wym307@ha.org.hk).

the enhanced deep super-resolution network (EDSR) contains 4.31×10^7 parameters and requires 1.65×10^{12} floating point operations (FLOPs) and 1.72 GB runtime memory when computing a single image with 128×128 resolution [5]. These burdensome SR models are unaffordable to assist the medical diagnosis, thus an efficient yet powerful SR network is indispensable for the diagnosis framework. Therefore, we propose a Multi-scale Refined Context Network (MRC-Net). In the MRC-Net, a MRC module is devised to extract both global structures at large-scale and local details at small-scale, with the Refined Context Fusion (RCF) to enhance the interaction of global and local paths. Consequently, MRC-Net is capable to extract abundant information with fewer layers and filters. To further improve the performance of MRC-Net instead of training from scratch, we propose a recursion distillation scheme. Specifically, we first train a recursive MRC-Net with a LSTM unit. The LSTM unit can harness all the features of previous iterations to correct the current inference errors, thereby achieving better SR reconstruction. After that, a single-forward MRC-Net is built and guided to learn the temporal knowledge of the trained recursive one under explicit supervision. In this way, the single-forward MRC-Net not only performs on par with the recursive one, but also requires less computation and runtime memory.

Another challenge of existing methods [12], [13], [15], [16] is that the original LR images and the SR images are not utilized comprehensively. These methods solely utilized the SR images for specific tasks, ignoring the original LR images and corresponding reliable information. In fact, SR is an inherently ill-posed task as various HR images can be degraded into the same LR image. SR methods may introduce artifacts to the reconstructed images, thereby leading to biased diagnosis. Therefore, we propose a novel diagnosis network, composed of two diagnosis branches, to jointly and collaboratively utilize the excessive information from SR images and the reliable information from original LR images. Instead of integrating the features of these two branches directly, a novel Sample Affinity Interaction (SAI) block is devised to exploit and assimilate discriminative features by investigating the relationship of samples, which results in a better interaction at semantic space. Besides, a sample affinity consistency is devised to constrain these two branches to maintain consistent yet effective semantic information towards diagnosis, and a sample affinity regularization is proposed to rectify the semantic information at low-level from the high-level one, thereby promoting the performance.

In this framework, MRC-Net is first trained on a specific SR dataset using the recursion distillation scheme for SR. To further exploit SR knowledge for diagnosis, we fine-tune the pretrained MRC-Net and diagnosis network in an end-to-end manner under diagnostic supervision. We summarize our contributions as follows:

- We propose a SR enhanced diagnosis framework for medical images, which is composed of an efficient SR network and a diagnosis network. To the best of our knowledge, this work represents the first effort to comprehensively joint super-resolution and medical diagnosis.
- We design an efficient MRC-Net for the SR task, with RCF to strengthen network capability. Moreover, to enable MRC-Net with the temporal knowledge, a recursion distillation scheme is devised to retain impressive SR performance with computation and runtime memory remarkably reduced.
- In the diagnosis network, we propose a novel SAI block to exploit and integrate features among two different resolution branches, utilizing the semantic relationship among samples. Additionally, two kinds of constraints, sample affinity consistency and sample affinity regularization, are devised to guide the multi-level information interaction.
- Extensive experiments on WCE and histopathology images prove the effectiveness of our approach, which outperforms state-of-the-art diagnosis algorithms on both synthetic and real LR images. The SR enhanced diagnosis framework significantly promotes the baseline diagnosis network with a 5.83% increase in accuracy, approaching the performance of HR images.

II. Related Works

A. Super-Resolution

In recent years, various deep learning based SR methods have been investigated. Dong *et al.* [3] proposed the super-resolution convolutional neural network (SRCNN) to predict SR images using three stacked convolutional layers. Kim *et al.* [4] introduced the global skip connection and made it possible to stack more layers for SR. EDSR [5] employed residual blocks [17] to further enhance the network capability. Moreover, a lightweight cascading residual network (CARN) [6] utilized dense skip connections to improve the residual blocks based structure. Instead of reconstruction in image space, a multi-level wavelet CNN (MWCNN) [7] conducted SR in wavelet domain to emphasize high-frequency components.

At the same time, targeted SR algorithms are proposed to meet the requirements of various medical fields [8]–[11]. Particularly, Zhao *et al.* [8] devised a channel splitting network with a dense branch and a residual branch to exploit hierarchical features for SR of magnetic resonance images. Li *et al.* [9] developed a two-stage SR network optimized by novel gradient sensitive loss and traditional mean square error (MSE) loss to super-resolve the arterial spin labeling. Khan *et al.* [10] adopted SR techniques to enhance the ultrasound imaging at the Hilbert domain. Instead of conventional single LR-HR pair, Mukherjee *et al.* [11] built multiple histopathology images with intermediate resolution, and trained a recurrent SR network with such a multi-resolution dataset.

B. Super-Resolution Aided High-Level Tasks

Since SR models can provide complementary information for low resolution images, several methods have been proposed to apply SR techniques to promote the performance of high-level tasks [12], [13]. To address the problem that tiny faces suffer from the missing of detailed information, Bai *et al.* [12] directly generated high-quality SR face regions from blurry small inputs to detect faces in the wild. Shermeyer *et al.* [13] explored the influence of various SR methods on object

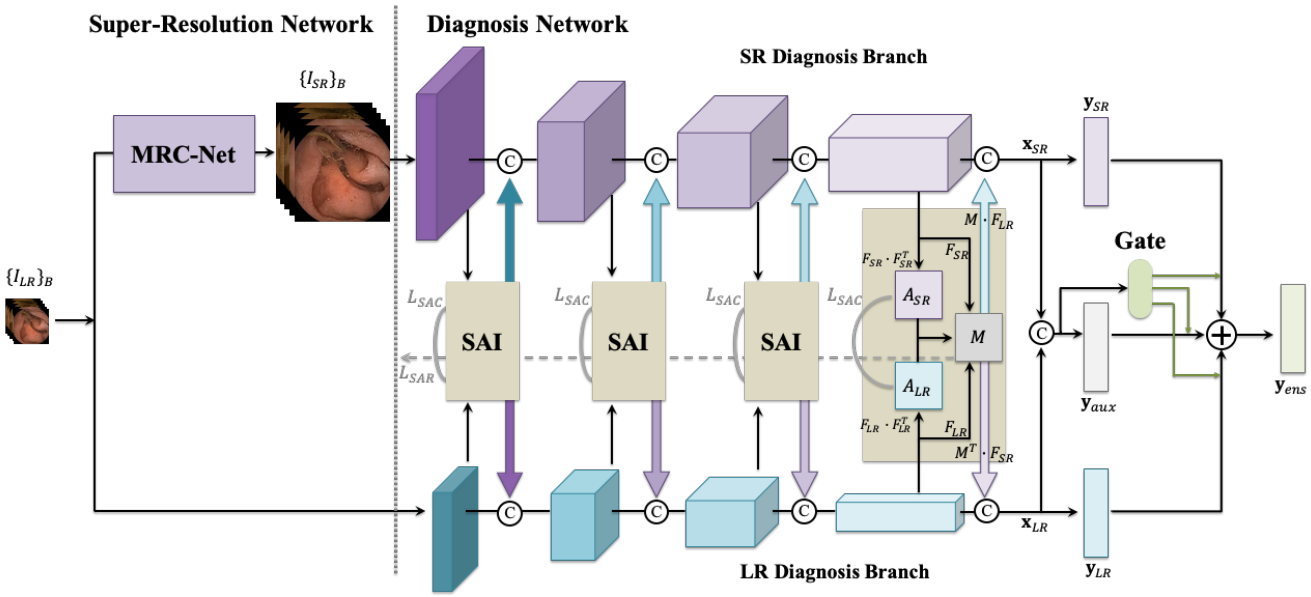

Fig. 2. The SR enhanced diagnosis framework, including a super-resolution network and a diagnosis network. The diagnosis network contains a SR diagnosis branch and a LR diagnosis branch. The Sample Affinity Interaction (SAI) blocks at multi-stage exchange the complementary semantic information of two branches. The gate mechanism integrates two branches and auxiliary prediction $\mathbf{y}_{aux}$ to produce the final diagnosis $\mathbf{y}_{ens}$.

detection in satellite imagery and proved that SR techniques improved the detection of tiny objects.

In healthcare applications, Srivastav *et al.* [15] introduced a progressive SR network to complement high-frequency features for the down-stream surgery task, which prompted surgeon pose estimation with a 6.5% increase in the percentage of correct key points. For the retinal SR, Mahapatra *et al.* [16] proposed a progressive generative adversarial network with a triplet loss to enable the stepwise improvement of image quality, which can promote the vasculature segmentation and microaneurysm detection. Instead of directly employing trained SR networks as a pre-processing step, we investigate a comprehensive diagnosis framework to leverage the information of both the SR images and the original LR ones.

III. Method

The SR enhanced diagnosis framework is composed of a SR network and a diagnosis network, as illustrated in Fig. 2. Particularly, the SR network employs a MRC-Net, which is optimized by the recursion distillation scheme to enhance the SR performance. The diagnosis network jointly considers the SR images and the original LR ones to generate the diagnostic predictions. In this section, we first introduce the MRC-Net structure and the recursion distillation scheme. Subsequently, we present the diagnosis network part, including the SAI block, the gate mechanism and tailored loss functions.

A. Super-Resolution Network

We propose the MRC-Net for the SR of medical images. As shown in Fig. 3, the MRC module extracts global and local features in two parallel paths, and utilizes the devised Refined Context Fusion (RCF) to conduct the efficient contextual information interaction. As appropriate larger network capacity improves the training process [18], the recursion distillation scheme is proposed to generate an enhanced MRC-Net. Specifically, we first train a recursive MRC-Net with LSTM mechanism, and then distill its temporal knowledge to a single-forward MRC-Net, thereby preserving performance with reduced resource demand. Compared with training from scratch, the distilled MRC-Net is strengthened with significant performance gain.

1) Recursive MRC-Net: We exemplify the architecture of MRC-Net with the recursive version in Fig. 3 (a). The MRC-Net first executes two groups of a 3×3 convolutional layer and two residual blocks [17] to convert the input I_{LR} into feature maps. After that, a convolutional LSTM unit is adopted to exploit and store the information of past feature maps. Specifically, a LSTM unit consists of an input gate $\mathbf{i}_t$, a forget gate $\mathbf{f}_t$, an output gate $\mathbf{o}_t$ and a cell state $\mathbf{C}_t$ [19], [20]. Provided with input features $\mathbf{X}$ and previous feedback $\mathbf{Y}_{t-1}$, temporal updates of states and gates are calculated as follows:

$$\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{W}_{xi} * \mathbf{X} + \mathbf{W}_{yi} * \mathbf{Y}_{t-1} + \mathbf{W}_{ci} \odot \mathbf{C}_{t-1} + \mathbf{b}_i\right) \\
\mathbf{f}_t &= \sigma\left(\mathbf{W}_{xf} * \mathbf{X} + \mathbf{W}_{yf} * \mathbf{Y}_{t-1} + \mathbf{W}_{cf} \odot \mathbf{C}_{t-1} + \mathbf{b}_f\right) \\
\mathbf{C}_t &= \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tanh\left(\mathbf{W}_{xc} * \mathbf{X} + \mathbf{W}_{yc} * \mathbf{Y}_{t-1} + \mathbf{b}_c\right) \\
\mathbf{o}_t &= \sigma\left(\mathbf{W}_{xo} * \mathbf{X} + \mathbf{W}_{yo} * \mathbf{Y}_{t-1} + \mathbf{W}_{co} \odot \mathbf{C}_t + \mathbf{b}_o\right) \\
\mathbf{H}_t &= \mathbf{o}_t \odot \tanh\left(\mathbf{C}_t\right)
\end{aligned} \tag{1}$$

where $*$ and $\odot$ denote the convolution operation and Hadamard product respectively, and σ and $\tanh$ are the Sigmoid and Tanh function respectively. $\mathbf{W}_{xi}$ represents parameters of a convolutional layer mapping from $\mathbf{X}$ to $\mathbf{i}_t$. Similarly, other $\mathbf{W}$ symbols also follow this rule. Since each forward is delivered with the same input, the recursive MRC-Net shares parameters during temporal iterations, without introducing extra parameters.

Then, the current output of LSTM unit, $\mathbf{H}_t$, is delivered

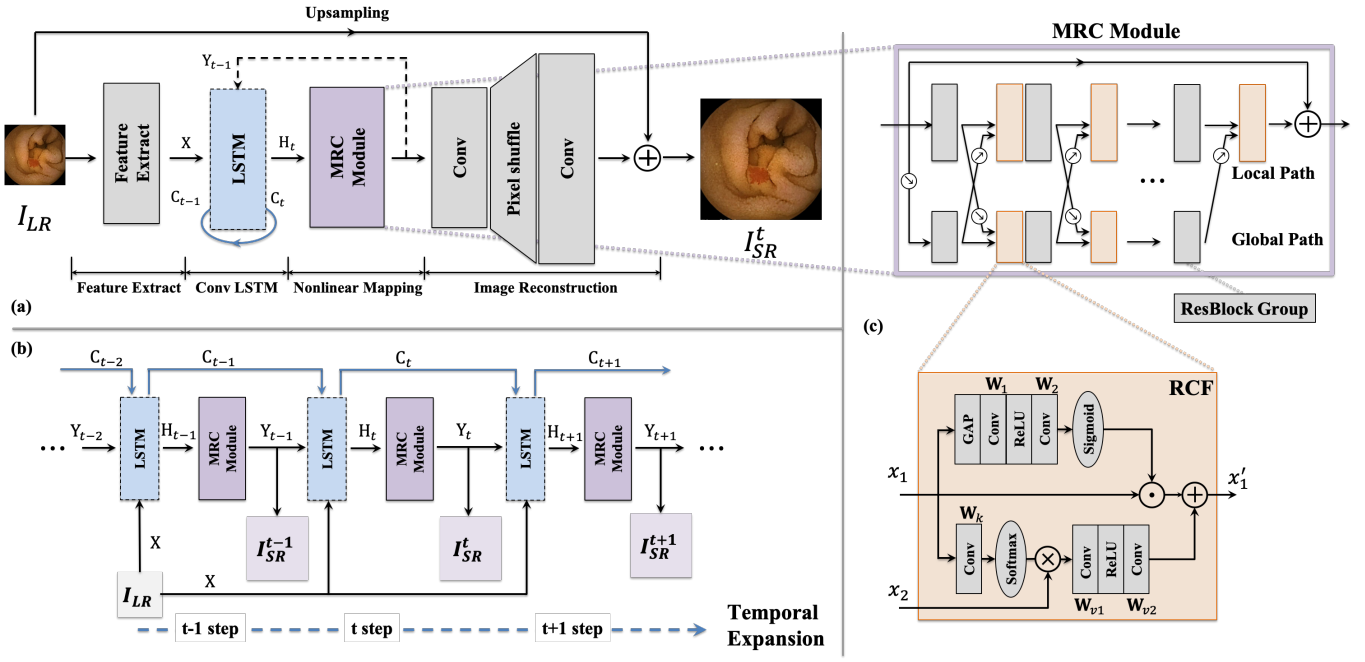

Fig. 3. (a) The recursive MRC-Net for SR. The single-forward MRC-Net deletes the feedback $\mathbf{Y}_{t-1}$ marked in a dashed line, and consequently the LSTM unit degenerates into convolutional layers without cell states. (b) The temporal expansion of the recursive MRC-Net, where feature extract and image reconstruction are omitted for simplicity. The learnable parameters of MRC module are shared during iterations. Provided with LR input $\boldsymbol{I_{LR}}$, the recursive MRC-Net utilizes $\mathbf{Y}_{t-1}$ and $\mathbf{C}_{t-1}$ to generate SR prediction $\boldsymbol{I}_{\boldsymbol{SR}}^{\boldsymbol{t}}$ at $\boldsymbol{t}$ step. Then, the feedback $\mathbf{Y}_{\boldsymbol{t}}$ and cell state $\mathbf{C}_{\boldsymbol{t}}$ are delivered to the recursive MRC-Ne at $\boldsymbol{t+1}$ step to enhance and rectify the previous SR output. (c) Multi-scale Refined Context (MRC) module and Refined Context Fusion (RCF).

into the MRC module, denoted as f, to exploit the significant features for SR recovery in a residual manner, as follows:

$$\mathbf{Y}_t = f(\mathbf{H}_t) + \mathbf{H}_t \tag{2}$$

The processed features $\mathbf{Y}_t$ will be sent back to the LSTM at the next $t+1$ step. The temporal expansion of the recursive MRC-Net is illustrated in Fig. 3 (b). The recursion mechanism progressively enhances the features and rectifies the potential mistakes made in previous iterations. Finally, we reconstruct features into image space with targeted resolution by a pixel shuffle layer [21]. A global skip connection with upsampling is adopted to accelerate the training convergence. The single-forward MRC-Net is composed of the same architecture as the recursive one, without the feedback of MRC module.

Multi-scale Refined Context (MRC) Module Since multi-scale feature maps provide different and complementary information of images [22], we devise the MRC module for the SR task. It consists of a local path with input feature maps and a parallel global path with the corresponding downsampled counterpart, as shown in Fig. 3 (c). Each path is constructed with repeated groups of residual blocks and the proposed RCF alternately. The stacked residual blocks in these two paths extract the features at different scales. With bilinear down/upsampling to adjust the resolution, RCF utilizes efficient global context [23] to refine and integrate the significant features of these two paths iteratively, as illustrated in Fig. 3 (c). Denote x_1 and x_2 as the feature maps of the current path and the other one, respectively, the refined output x_1' is computed as follows:

$$\begin{aligned} x_1' =& x_1 \odot \sigma\left(\mathbf{W}_2 * \delta\left(\mathbf{W}_1 * \mathbf{GAP}\left(x_1\right)\right)\right) \\ &+ \mathbf{W}_{v2} * \sigma(\mathbf{W}_{v1} * (\mathbf{Softmax}(\mathbf{W}_k * x_1) \times x_2)) \end{aligned} \tag{3}$$

where σ and δ are Sigmoid and PReLU function, respectively, $\mathbf{GAP}$ represents global average pooling, and each $\mathbf{W}$ represents the parameters of 1×1 conv layers. $\mathbf{Softmax}$ is conducted to normalize feature maps in width and height dimensions, and $\times$ stands for matrix multiplication in spatial dimension with necessary reshape as [23]. The first term adaptively modifies x_1 by a self-gate, while the second term rectifies x_2 using the spatial context of x_1. By adding these two terms together, RCF is able to utilize the context of x_1 and x_2 comprehensively, preserving the significant features for medical SR tasks [24]. In this way, MRC-Net can capture details in the local path and preserve structural information in the global path simultaneously, which is helpful to exploit appropriate information for better reconstruction.

2) Recursion Distillation Scheme: A novel recursion distillation scheme is proposed to enhance the efficient MRC-Net for super resolution of medical images. Particularly, we first train a recursive MRC-Net containing a LSTM unit, denoted as $f_r(\theta_1)$, where θ_1 represents its parameters. As previously introduced, the recursive MRC-Net contains temporal context. Then, we build a single-forward MRC-Net $f_s(\theta_2)$ to distill the knowledge of $f_r(\theta_1)$. $f_s(\theta_2)$ is initialized with the trained parameters of $f_r(\theta_1)$, as $\theta_2 := \theta_1$, and trained with the distill

Algorithm 1: The Recursion Distillation Scheme

Input : The recursive MRC-Net $f_r(\theta_1)$;
The training dataset $\{I^i_{LR}, I^i_{HR}\}^{1\leq i\leq |D|}$;
Weighted factors γ_d, γ_1, γ_2, γ_3 and γ_4;
Output: The distilled MRC-Net with single-forward $f_s(\theta_2)$;

Initialization: $\theta_2 := \theta_1$;
while $f_s(\theta_2)$ reaches convergence **do**
  $I^r_{SR}, F^r_{SR}, F^r_{LR}, C^r, H^r = f_r(I_{LR}; \theta_1)$ at the last step of recursion;
  $I^s_{SR}, F^s_{SR}, F^s_{LR}, C^s, H^s = f_s(I_{LR}; \theta_2)$;
  comptue SR supervision $L_{\mathrm{MAE}}(I^s_{SR}, I_{HR})$;
  compute the distill loss L_d as Eq. (4), containing the recursion knowledge from $f_r(\theta_1)$;
  minimize $L_{\mathrm{MAE}}(I^s_{SR}, I_{HR}) + \gamma_d L_d$;
end
while $f_s(\theta_2)$ reaches convergence **do**
  $I^s_{SR} = f_s(I_{LR}; \theta_2)$;
  minimize $L_{\mathrm{MAE}}(I^s_{SR}, I_{HR})$;
end
Obtain the distilled single-forward MRC-Net $f_s(\theta_2)$ with recursion knowledge.

loss L_d, as computed in MSE:

$$\begin{aligned} L_d =& \gamma_1 L_{\mathrm{MSE}}(\mathbf{C}^s, \mathbf{C}^r) + \gamma_2 L_{\mathrm{MSE}}(\mathbf{H}^s, \mathbf{H}^r) \\ &+ \gamma_3 L_{\mathrm{MSE}}(F^s_{LR}, F^r_{LR}) + \gamma_4 L_{\mathrm{MSE}}(F^s_{SR}, F^r_{SR}) \end{aligned} \tag{4}$$

where γ_1, γ_2, γ_3 and γ_4 are weights to balance different loss terms. Particularly, $L_{\mathrm{MSE}}(\mathbf{C}^s, \mathbf{C}^r)$ and $L_{\mathrm{MSE}}(\mathbf{H}^s, \mathbf{H}^r)$ are the explicit supervision from cell states and LSTM outputs, respectively. In addition, constraints on multi-resolution features, $L_{\mathrm{MSE}}(F^s_{LR}, F^r_{LR})$ and $L_{\mathrm{MSE}}(F^s_{SR}, F^r_{SR})$, are leveraged to assist the distillation procedure, where F_{LR} and F_{SR} are feature maps before and after the pixel shuffle layer, respectively. Considering the predictions of $f_r(\theta_1)$ may deviate from the ground-truth [25], we further fine-tune the distilled MRC-Net $f_s(\theta_2)$ with mean absolute error loss L_{MAE} on the training set. The procedures of recursion distillation scheme are summarized in Algorithm 1.

Different from existing distillation methods [26]–[28], $f_s(\theta_2)$ has the same architecture as the recursive teacher $f_r(\theta_1)$. Removing the recursive feedback, $f_s(\theta_2)$ demands less computation, runtime memory and inference time than $f_r(\theta_1)$. Through appropriate knowledge transfer, the distilled $f_s(\theta_2)$ preserves comparable performance with $f_r(\theta_1)$. In the following, the diagnosis framework will adopt the distilled single-forward MRC-Net $f_s(\theta_2)$, which is denoted as MRC-Net for simplicity. To avoid ambiguity, the recursive version $f_r(\theta_1)$ is named as recursive MRC-Net. The MRC-Net will be further fine-tuned under the diagnosis framework.

B. Diagnosis Network

As shown in Fig. 2, the diagnosis network includes two diagnosis branches progressively interacted through the proposed SAI blocks. We implement two branches with the same structure. These two branches exploit the SR and original LR images, and predict the diagnosis $\mathbf{y}_{SR}$ and $\mathbf{y}_{LR}$ separately, as well as an auxiliary prediction $\mathbf{y}_{aux}$ for each sample. Finally, a gate mechanism is utilized to integrate the final diagnosis $\mathbf{y}_{ens}$. Particularly, feature maps at different stages are delivered to the SAI blocks to exploit the semantic affinity among samples and exchange the rectified features to the opposite branch. Moreover, we devise the sample affinity consistency loss L_{SAC} and sample affinity regularization loss L_{SAR} to constrain the semantic information of two branches and provide complementary supervision for the diagnosis network.

1) Sample Affinity Interaction (SAI) Block: As the observation that semantically similar inputs tend to produce coherent activation patterns in a trained network [29], to leverage the knowledge among samples can potentially encourage the diagnosis network to extract discriminative features for medical diagnosis. Instead of directly integrating features of LR diagnosis branch and SR diagnosis branch, we propose the SAI block to exploit the semantic relationship of these two branches with sample affinity, enabling a better information interaction. Given a mini-batch input with B samples, denote the feature maps of a specific stage of the SR diagnosis branch as $\mathbf{F}_{SR} \in \mathbf{R}^{B\times C\times H\times W}$, where C is the number of channels, and H and W are the height and width of feature maps. The LR diagnosis branch generates the feature map $\mathbf{F}_{LR} \in \mathbf{R}^{B\times C\times H'\times W'}$ at the same stage. First, we reshape these features maps into two dimensions, as $\mathbf{F}_{SR} \in \mathbf{R}^{B\times CHW}$ and $\mathbf{F}_{LR} \in \mathbf{R}^{B\times CH'W'}$. The sample affinity $\mathbf{A}_{SR} \in \mathbf{R}^{B\times B}$ of the SR diagnosis branch is formulated in matrix multiplication as:

$$\mathbf{A}_{SR} = \mathbf{F}_{SR} \cdot \mathbf{F}^T_{SR} \tag{5}$$

Specifically, $\mathbf{A}_{SR}[i,j] = \mathbf{F}_{SR}[i,:] \cdot \mathbf{F}_{SR}[j,:]^T$ represents the semantic similarity between i-th sample and j-th sample, with a large value for samples within the same category or a small value for samples across different categories [29]. We also conduct the same process on $\mathbf{F}_{LR}$ to generate the sample affinity $\mathbf{A}_{LR}$ of the LR diagnosis branch as $\mathbf{A}_{LR} = \mathbf{F}_{LR} \cdot \mathbf{F}^T_{LR}$. A row-wise L_2 normalization is applied to both $\mathbf{A}_{SR}$ and $\mathbf{A}_{LR}$ before further operations.

Then, we calculate the mutual sample affinity of two branches by multiplying $\mathbf{A}_{SR}$ and $\mathbf{A}_{LR}$. Specifically, $\mathbf{M}_{L\to S} \in \mathbf{R}^{B\times B}$ represents the mutual sample affinity from the LR diagnosis branch to the SR diagnosis branch and $\mathbf{M}_{S\to L} \in \mathbf{R}^{B\times B}$ represents the reverse mapping, as follows:

$$\begin{aligned} \mathbf{M}_{L\to S} &= \mathbf{A}_{SR} \cdot \mathbf{A}^T_{LR} \\ \mathbf{M}_{S\to L} &= \mathbf{A}_{LR} \cdot \mathbf{A}^T_{SR} \end{aligned} \tag{6}$$

Note that these two matrices are transposed to each other, as $\mathbf{M}_{L\to S} = (\mathbf{M}_{S\to L})^T$. We apply $\mathbf{M}_{L\to S}$ to map $\mathbf{F}_{LR}$ of the LR diagnosis branch to the SR diagnosis branch, and apply the transposed affinity $\mathbf{M}_{S\to L}$ on $\mathbf{F}_{SR}$ in turn. The transferred features $\mathbf{F}_{L\to S}$ and $\mathbf{F}_{S\to L}$ are computed as follows:

$$\begin{aligned} \mathbf{F}_{L\to S} &= \mathbf{M}_{L\to S} \cdot \mathbf{F}_{LR} \\ \mathbf{F}_{S\to L} &= \mathbf{M}_{S\to L} \cdot \mathbf{F}_{SR} \end{aligned} \tag{7}$$

Finally, we reshape $\mathbf{F}_{L\to S}$ and $\mathbf{F}_{S\to L}$ back into 4-D tensors, as $\mathbf{F}_{L\to S} \in \mathbf{R}^{B\times C\times H'\times W'}$ and $\mathbf{F}_{S\to L} \in \mathbf{R}^{B\times C\times H\times W}$, and

conduct necessary up/downsampling to achieve compatible resolution with the opposite branch. After that, the features from the current branch and the affinity-rectified features from the opposite branch, e.g., $\mathbf{F}_{SR}$ and $\mathbf{F}_{L\to S}$, are concatenated in channel dimension and followed by a 1×1 bottleneck to generate compact feature maps for the next layer. Therefore, the SAI block utilizes the semantic relationship of samples and enables the information interaction between two branches.

Compared with previous spatial or channel affinity within each sample, which adjust features at the representation space [30]–[33], the proposed SAI block utilizes the sample affinity to model the relationship among samples, guaranteeing the consistency of different inputs at semantic space. In addition, the SAI block achieves better adaptation of the features among two branches, since mutual sample affinity exploits the bijective mapping at semantic space.

2) Diagnosis Ensemble with Gate Mechanism: A gate mechanism is utilized to produce the final diagnosis prediction $\mathbf{y}_{ens}$. The features of two branches, $\mathbf{x}_{SR}$ and $\mathbf{x}_{LR}$, are concatenated after global average pooling and then delivered to two separate convolutional layers to generate a gate vector $\mathbf{G}$ and an auxiliary prediction, $\mathbf{y}_{aux}$. The gate is calculated as $\mathbf{G}=\mathrm{Conv}_2(\mathrm{ReLU}(\mathrm{Conv}_1([\mathbf{x}_{SR},\mathbf{x}_{LR}])))$, where Conv_1 and Conv_2 are two 1×1 convolutional layers to shrink the channel dimension progressively. Given each sample, $\mathbf{G}\in\mathbf{R}^3$ represents the importance of $\mathbf{y}_{SR}$, $\mathbf{y}_{LR}$ and $\mathbf{y}_{aux}$. The final diagnosis $\mathbf{y}_{ens}$ is obtained by summation weighted with the importance score $\mathbf{G}$, as follows:

$$\mathbf{y}_{ens}=\left[\mathbf{y}_{SR},\,\mathbf{y}_{LR},\,\mathbf{y}_{aux}\right]\cdot\max(0,\tanh(\mathbf{G})) \tag{8}$$

where $\tanh$ adjusts the range of $\mathbf{G}$ and $\max$ operation omits the diagnosis branch with a negative score. In this way, the importance scores in $\mathbf{G}$ are controlled to $[0,1)$. As different $\mathbf{G}$ vectors are generated for each sample, the gate mechanism achieves a tailor-made ensemble strategy.

3) Loss Functions: To guide the multi-level interaction of two branches and promote the diagnosis performance, the proposed framework is optimized by a joint loss function, including the proposed sample affinity consistency loss, sample affinity regularization loss, cross entropy loss and rank loss.

Sample Affinity Consistency Loss Although the SR images and the original LR images in two branches have different representation spaces, these paired images share the identical diagnostic labels. To guarantee the consistent semantic information of two branches towards diagnosis, we propose the sample affinity consistency L_{SAC} to minimize the multi-stage differences between $\mathbf{A}_{SR}$ and $\mathbf{A}_{LR}$, as follows:

$$L_{\mathrm{SAC}}=\sum_{s=1}^{S}||\tau(\mathbf{A}_{SR}^{(s)})-\tau(\mathbf{A}_{LR}^{(s)})||_{\mathrm{F}} \tag{9}$$

where $||\cdot||_{\mathrm{F}}$ is the Frobenius norm of matrix to measure the affinity difference of two branches. S is the number of stages in the diagnosis network. τ is the temperature function to soften the affinity distribution in row-wise:

$$\tau(\mathbf{A}[i,:])=\frac{\exp\left(\mathbf{A}[i,:]/T\right)}{\sum_j^B\exp\left(\mathbf{A}[j,:]/T\right)} \tag{10}$$

where T is the temperature to control the distribution, as a larger T makes the distribution softer.

Sample Affinity Regularization Loss As deep layers extract more abstract features than shallow layers at semantic space, the supervision on attention maps from deep layers to shallow layers can lead an improvement in high-level tasks [34], [35]. Following this, we propose the sample affinity regularization L_{SAR} to constrain the mutual affinity matrix of shallow layer using the deep layer one. The L_{SAR} is defined as follows:

$$\begin{aligned}L_{\mathrm{SAR}}=\sum_{s=1}^{S-1}\frac{1}{2}(&||\tau(\mathbf{M}_{L\to S}^{(s+1)})-\tau(\mathbf{M}_{L\to S}^{(s)})||_{\mathrm{F}}\\&+||\tau(\mathbf{M}_{S\to L}^{(s+1)})-\tau(\mathbf{M}_{S\to L}^{(s)})||_{\mathrm{F}})\end{aligned} \tag{11}$$

where $\mathbf{M}^{(s)}$ represents the mutual sample affinity at the s-th stage. In this way, the high-quality mutual sample affinity of deep stages can refine the counterpart of shallow stages, and the improved affinity matrices of shallow stages in turn benefit the deeper stages.

Cross Entropy Loss We also employ the cross entropy loss L_{CE} on the final prediction $\mathbf{y}_{ens}$ as well as preliminary predictions $\mathbf{y}_{SR}$, $\mathbf{y}_{LR}$ and $\mathbf{y}_{aux}$:

$$L_{\mathrm{CE}}=\sum_{\mathbf{y}\in\{\mathbf{y}_{ens},\mathbf{y}_{SR},\mathbf{y}_{LR},\mathbf{y}_{aux}\}}L_{\mathrm{CE}}(\mathbf{y},\mathbf{t}) \tag{12}$$

where $\mathbf{t}$ represents the one-hot vector of category label.

Rank Loss To encourage the final diagnosis $\mathbf{y}_{ens}$ to provide a better diagnosis than preliminary ones, we introduce a rank loss L_{rank} to penalize the case that $\mathbf{y}_{ens}$ performs worse:

$$L_{\mathrm{rank}}=\sum_{\mathbf{y}\in\{\mathbf{y}_{SR},\mathbf{y}_{LR},\mathbf{y}_{aux}\}}\max\left\{0,\mathbf{y}(c)-\mathbf{y}_{ens}(c)+m\right\} \tag{13}$$

where c is the index of the correct category and $\mathbf{y}(c)$ is a scalar representing the predicted probability on the correct category. m is a margin between $\mathbf{y}_{ens}$ and $\mathbf{y}$, leading $\mathbf{y}_{ens}$ producing a higher probability for the correct category. We empirically set m as 0.05. In this way, $\mathbf{y}_{ens}$ integrates preliminary predictions to accomplish a more reliable diagnosis.

Finally, the total loss of the diagnosis network is defined as follows:

$$L=L_{\mathrm{CE}}+\lambda_1 L_{\mathrm{rank}}+\lambda_2 L_{\mathrm{SAC}}+\lambda_3 L_{\mathrm{SAR}} \tag{14}$$

where λ_1, λ_2 and λ_3 are trade-off weights to adjust the importance of loss components, which are empirically set as 1, 100 and 100, respectively. Through optimizing with Eq. (14), our SR enhanced diagnosis framework can obtain the remarkable performance.

IV. Experiments

A. Dataset and Settings

As the acquisition equipment for high-quality WCE and histopathology images is typically very expensive, SR techniques can process images acquired by low-quality devices into the same resolution with satisfactory quality [11], which are widely available and comparatively inexpensive. To prove the contribution of our work to this issue, we conducted experiments on these two kinds of images in synthetic and real LR scenarios.

TABLE I
SR PERFORMANCE AND RESOURCE DEMANDS ON $4\times$ ENDOSR DATATSET AND $2\times$ HISTOSR DATATSET. BEST AND SECOND BEST RESULTS ARE **HIGHLIGHTED** AND UNDERLINED.

$4\times$ **EndoSR**	PSNR	SSIM	Param (10^6)	FLOPs (10^{10})	Memory (10^2 MB)
Bicubic	34.350	0.9114	-	-	-
VDSR [4]	38.090	0.9500	**0.67**	35.0	24.3
CARN [6]	38.309	0.9520	1.11	5.17	4.51
EDSR* [5]	38.381	0.9531	1.52	6.50	**3.03**
MWCNN [7]	38.382	0.9527	24.9	54.3	8.59
MRC-Net	**38.586**	**0.9543**	0.78	**3.18**	4.35
$2\times$ **HistoSR**	PSNR	SSIM	Param (10^6)	FLOPs (10^{10})	Memory (10^2 MB)
Bicubic	28.399	0.8859	-	-	-
VDSR [4]	32.378	0.9481	**0.67**	4.92	3.42
CARN [6]	32.393	0.9477	0.96	1.78	1.44
EDSR* [5]	32.676	0.9502	1.37	2.53	**0.97**
MWCNN [7]	32.498	0.9486	24.9	7.64	1.21
MRC-Net	**32.754**	**0.9509**	0.71	**1.00**	1.35

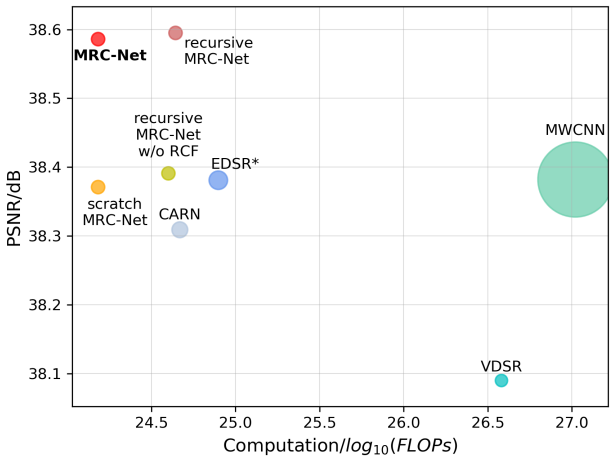

Fig. 4. The performance and resource demands of state-of-the-art SR networks on $4\times$ EndoSR dataset. The horizontal axis represents the computation, measured by logarithmic FLOPs for better observation. The vertical axis represents PSNR. The circle area represents the amount of parameters. A smaller circle at the top-left corner is considered to be a better SR network.

1) WCE Dataset: We evaluated the SR enhanced diagnosis framework on the CAD-CAP dataset [14], which consists of $1,800$ labeled WCE images, including 600 normal images, 600 inflammatory ones and 600 vascular lesion ones. Following the process to unify image resolution and quality [36], we resized the WCE images to 128×128 and applied a uniform circle mask to the dataset. Therefore, the processed CAD-CAP dataset serves as the synthetic LR case for diagnosis. Four-fold cross validation was adopted for this dataset. For the SR task, we built the Endoscopy SR (EndoSR) dataset with another $1,807$ unlabeled WCE images of the CAD-CAP dataset [14]. All images were first resized into 512×512 resolution as the HR images, which were further downsampled into the LR ones with a resolution of 128×128 using bicubic kernel. Consequently, the EndoSR dataset provided a $4\times$ zoom-in mapping for the SR task, with $1,507$ LR-HR pairs for training and 300 pairs for test.

2) Histopathological Dataset: For the histopathology images, we evaluated the SR enhanced diagnosis framework on a modified PCam dataset [37], [38], which consists of $178,240$ images with 96×96 pixels from lymph node section. These images were captured by $10\times$ objective microscopy, which serve as the diagnosis task of real LR images. Each image was annotated with a binary label, indicating the presence of metastatic tissue. The experiment was conducted in 3 non-overlap fold settings, with $160,416$ training images and $17,824$ test images in each evaluation. Furthermore, we built a Histopathology Super-Resolution (HistoSR) dataset using the high-quality H&E stained WSIs of Camelyon16 dataset [39]. Through random cropping and bicubic downsampling, the HistoSR dataset provided a $2\times$ zoom-in mapping from 96×96 patch to 192×192 patch, with $30,000$ training pairs and $5,000$ test pairs.

3) Experimental Settings: In our implementation, the MRC module was constructed with 4 successive groups of 4 residual blocks, and global and local paths individually contained 32 filters at each layer. In the diagnosis network, both the SR diagnosis branch and the LR diagnosis branch employed ResNet-18 structure [17]. The diagnosis framework was trained in two steps. Specifically, MRC-Net was first trained under the recursion distillation scheme on a specific SR dataset, optimized by Adam with the batch size of 4. The learning rate was initialized as 1×10^{-4} and halved after every 100 epochs until convergence. The weight decay was set as 1×10^{-5}. In the recursion distillation scheme, the number of recursion was empirically set as 3, and γ_1, γ_2, γ_3, γ_4 and γ_d are set as 2, 2, 2, 0.5 and 1, respectively. To integrate MRC-Net into the diagnosis framework, we conducted the SR with RGB channels. After that, the whole framework was optimized by the diagnostic supervision, optimized by Adam with the batch size of 8. The learning rate was initialized as 1×10^{-4} and halved after every 30 epochs, where the learning rate of MRC-Net was decreased by a factor of 0.01. For the consistency and regularization of sample affinity, we empirically set T as 1.

We evaluated SR networks from the perspective of performance and efficiency. Specifically, peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) were adopted to assess the reconstruction quality, while the number of parameters, FLOPs and runtime memory were calculated to measure the resource demand of SR networks. The diagnosis performance was evaluated by accuracy, F1 score, sensitivity, specificity, Cohen's Kappa (CK) and Matthews Correlation Coefficient (MCC). Since the CAD-CAP dataset contains three categories, binary metrics were first computed for each category and then averaged, which are called macro averaging.

B. Super-Resolution Experiments

1) SR on EndoSR and HistoSR: We compared our MRC-Net with state-of-the-art networks [4]–[7] on EndoSR and HistoSR dataset in Table I. As large-scale SR networks bring impractical resource overheads to the diagnosis framework, efficient SR networks with impressive performance are preferred in this work. We employed the EDSR baseline model [5], which is a smaller EDSR with the same topology. For the

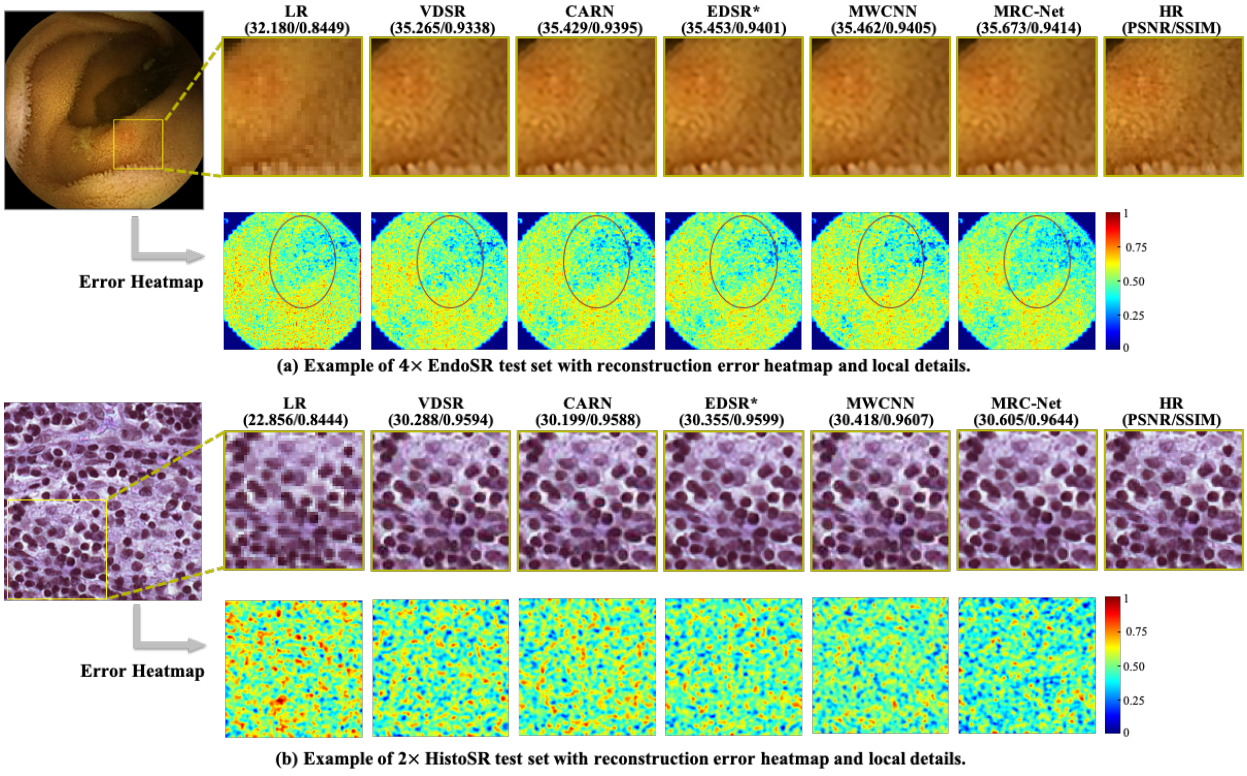

Fig. 5. Qualitative results of (a) $4\times$ EndoSR dataset; (b) $2\times$ HistoSR dataset. In each subfigure, the first row illustrates the LR, SR predictions of various SR networks and HR ground truth of the yellow box. The second row compares our MRC-Net with VDSR [4], CARN [6], EDSR* [5] and MWCNN [7]. For a better visualization, each point of the heatmap represents the normalized reconstruction error after 11×11 Gaussian filtering, with blue for low reconstruction errors and red for high errors.

TABLE II
ABLATION STUDY OF MRC-NET ON $4\times$ ENDOSR DATASET. BEST AND SECOND BEST RESULTS ARE **HIGHLIGHTED** AND UNDERLINED.

Config	PSNR	SSIM	Param. (10^6)	FLOPs (10^{10})	Memory (10^2 MB)
Stack	38.548	0.9534	1.76	5.04	7.02
Recursive	**38.595**	**0.9543**	0.79	5.04	7.02
- w/o LSTM	38.417	0.9531	0.69	4.47	6.22
- w/o RCF	38.391	0.9528	0.78	4.83	6.68
Scratch	38.371	0.9526	0.79	**3.18**	**4.35**
MRC-Net	38.586	**0.9543**	0.79	**3.18**	**4.35**
DRCN [40]	37.817	0.9485	1.78	557	94.8
DRRN [41]	37.959	0.9487	**0.30**	140	48.7
SRRFN [42]	38.342	0.9538	4.21	53.8	26.6

EndoSR dataset, MRC-Net achieves a superior performance, with PSNR of 38.586 dB and SSIM of 0.9543. Our MRC-Net outperforms the second best MWCNN [7] with a 0.204 dB advantage, and demands only 5.86% FLOPs and 3.13% parameters of MWCNN. In fact, MRC-Net requires the least computation among these state-of-the-art algorithms, with low amount of parameters as well as runtime memory. An evident comparison is illustrated in Fig. 4, where our MRC-Net at the top-left corner is considered to achieve the best trade-off between performance and efficiency among state-of-the-art networks [4]–[7].

Moreover, the consistent advantage of MRC-Net is also confirmed on the HistoSR dataset. The MRC-Net generates the best reconstruction with PSNR of 32.754 dB and SSIM of 0.9509, and requires the least 1.00×10^{10} FLOPs and the second least 0.71 million parameters. With the fact that MRC-Net only brings a 3% overhead of parameters to the entire framework, the efficient yet powerful MRC-Net is suitable for enhancing the diagnostic task.

Qualitative results of EndoSR and HistoSR dataset are elaborated in Fig. 5. The reconstructed details of attentive regions, prove that our MRC-Net can generate visual high-quality SR predictions, which are also more consistent with ground truth measured by PSNR and SSIM. To further compare with state-of-the-art networks, we visualized the heatmap of reconstruction errors, which were normalized by the maximum and minimum reconstruction errors of different methods on the same sample. Specifically, MRC-Net alleviates reconstruction errors apparently within the marked region in Fig. 5 (a), and generates significantly less red areas and more blue areas in Fig. 5 (b).

2) Ablation Study: To confirm the capability of the recursion distillation scheme and the RCF, we implemented the following networks on the EndoSR dataset, as shown in Table II.

- *Stack*: A network stacked with 3 MRC modules, where the LSTM unit degenerates into convolutional layers.
- *Recursive*: A 3-times recursive MRC-Net containing a MRC module. In addition, we also conducted the recursive MRC-Net without LSTM unit or RCF to verify the influence of LSTM unit and RCF.

TABLE III

THE DIAGNOSIS COMPARISON ON CAD-CAP AND PCAM DATASET. BEST AND SECOND BEST RESULTS ARE **HIGHLIGHTED** AND UNDERLINED.

CAD-CAP	Accuracy (%)	F1 score (%)	Sensitivity (%)	Specificity (%)	CK (%)	MCC (%)	Param (10^7)	FLOPs (10^9)
Fan *et al.* [43]	80.13 ± 0.36	79.93 ± 0.49	80.14 ± 0.39	90.07 ± 0.18	70.20 ± 0.39	70.32 ± 0.49	9.25	1.07
Jia *et al.* [44]	80.52 ± 0.46	80.47 ± 0.45	80.52 ± 0.48	90.26 ± 0.23	70.79 ± 0.69	70.88 ± 0.72	1.50	1.85
Yuan *et al.* [45]	91.15 ± 0.78	91.11 ± 0.77	91.15 ± 0.76	95.57 ± 0.40	86.72 ± 1.18	86.87 ± 1.28	0.06	1.23
Guo *et al.* [36]	$\underline{93.06} \pm 0.84$	$\underline{93.05} \pm 0.84$	$\underline{93.06} \pm 0.84$	$\underline{96.53} \pm 0.42$	$\underline{89.58} \pm 1.26$	$\underline{89.65} \pm 1.32$	0.23	2.72
Ours	$\mathbf{94.94} \pm 0.38$	$\mathbf{94.78} \pm 0.53$	$\mathbf{94.78} \pm 0.53$	$\mathbf{97.39} \pm 0.26$	$\mathbf{92.17} \pm 0.79$	$\mathbf{92.20} \pm 0.81$	2.60	54.3
PCam	Accuracy (%)	F1 score (%)	Sensitivity (%)	Specificity (%)	CK (%)	MCC (%)	Param (10^7)	FLOPs (10^9)
Araújo *et al.* [46]	85.93 ± 0.48	85.90 ± 0.51	88.18 ± 3.81	83.68 ± 4.69	71.86 ± 0.95	72.10 ± 0.59	0.93	0.04
ResNet-101 [17]	93.96 ± 0.31	93.96 ± 0.31	94.59 ± 2.02	93.33 ± 2.40	87.92 ± 0.62	87.98 ± 0.56	4.25	2.88
Veeling *et al.* [37]	$\underline{95.25} \pm 0.99$	$\underline{95.25} \pm 0.99$	93.90 ± 1.06	$\underline{95.96} \pm 1.78$	89.86 ± 1.74	89.89 ± 1.75	0.40	1.68
Kassani *et al.* [47]	95.05 ± 0.31	95.05 ± 0.31	$\underline{94.23} \pm 0.38$	95.87 ± 0.24	$\underline{90.10} \pm 0.61$	$\underline{90.11} \pm 0.61$	15.5	8.62
Ours	$\mathbf{97.11} \pm 0.04$	$\mathbf{97.03} \pm 0.04$	$\mathbf{96.78} \pm 0.14$	$\mathbf{97.28} \pm 0.12$	$\mathbf{94.06} \pm 0.07$	$\mathbf{94.06} \pm 0.07$	2.59	13.7

TABLE IV

THE WCE DIAGNOSIS COMPARISON WITH LR, BICUBIC INTERPOLATED LR AND SR INPUTS.

Input	Method	Accuracy (%)	F1 score (%)	CK (%)
LR	Fan *et al.* [43]	80.13 ± 0.36	79.93 ± 0.49	70.20 ± 0.39
	Jia *et al.* [44]	80.52 ± 0.46	80.47 ± 0.45	70.79 ± 0.69
	Yuan *et al.* [45]	91.15 ± 0.78	91.11 ± 0.77	86.72 ± 1.18
	Guo *et al.* [36]	93.06 ± 0.84	93.05 ± 0.84	89.58 ± 1.26
Bicubic ILR	Fan *et al.* [43]	80.25 ± 0.43	80.08 ± 0.56	70.37 ± 0.67
	Jia *et al.* [44]	80.73 ± 0.26	80.66 ± 0.30	71.09 ± 0.39
	Yuan *et al.* [45]	91.16 ± 0.80	91.13 ± 0.81	86.74 ± 1.20
	Guo *et al.* [36]	93.15 ± 0.40	93.15 ± 0.40	89.72 ± 0.59
SR	Fan *et al.* [43]	80.28 ± 0.26	80.13 ± 0.35	70.43 ± 0.31
	Jia *et al.* [44]	80.69 ± 0.74	80.68 ± 0.82	71.04 ± 1.11
	Yuan *et al.* [45]	91.37 ± 0.78	91.34 ± 0.79	87.05 ± 1.17
	Guo *et al.* [36]	93.27 ± 1.25	93.25 ± 1.21	89.91 ± 1.87
LR	Ours	$\mathbf{94.94} \pm 0.38$	$\mathbf{94.78} \pm 0.53$	$\mathbf{92.17} \pm 0.79$

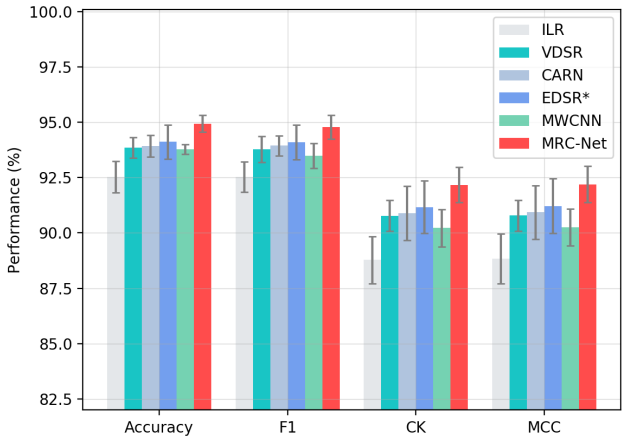

Fig. 6. Performance comparison of the SR enhanced diagnosis framework using various SR methods on CAD-CAP dataset.

- *Scratch*: A single-forward MRC-Net trained from scratch.
- *MRC-Net*: A single-forward MRC-Net optimized by the recursion distillation scheme.

In Table II, the recursive MRC-Net achieves the best performance, including 38.595 dB of PSNR and 0.9543 of SSIM. The recursive MRC-Net outperforms the stack one with 0.047 dB in PSNR, containing 45% parameters of the stack one. We conjecture that the recursion mechanism can leverage parameters more effectively than simply stacking more layers. As the recursive MRC-Net without LSTM unit reaches 38.417 dB PSNR, removing the LSTM unit leads a 0.178 dB PSNR decrease to the recursive MRC-Net, which performs worse than the stack one. The existing simple recursive SR networks [40]–[42] without LSTM unit, also lead to unsatisfactory SR performance on $4\times$ EndoSR dataset. This confirms that the LSTM unit is significant to exploit and rectify features of previous steps in the recursion mechanism. Moreover, the recursive MRC-Net without RCF achieves 38.391 dB in PSNR, which proves that RCF can integrate multi-scale features better with a 0.204 dB gain in PSNR. Compared with the 38.371% accuracy of training from scratch, the recursion distillation scheme brings MRC-Net a PSNR advantage of 0.215 dB. In this way, MRC-Net preserves the comparable performance of the recursive one, achieving 38.586 dB in PSNR with only 63% FLOPs and 62% runtime memory.

C. Diagnosis Experiments

1) WCE Experiment: We compared the diagnosis performance of our framework with state-of-the-art algorithms [36], [43]–[45] on the CAD-CAP dataset. As shown in Table III, our SR enhanced diagnosis framework achieves a superior performance with averaged accuracy of 94.94%, macro F1 score of 94.78% and CK of 92.17%, which outperforms the second best one [36] with an advantage of 1.88% in accuracy and 2.59% in CK. Moreover, our method reveals impressive yet balanced sensitivity of 94.78% and specificity of 97.39%. By introducing SR knowledge with affordable computation overhead, our diagnosis framework achieves more reliable predictions on WCE images. In Table IV, we further implemented state-of-the-art algorithms [36], [43]–[45] with interpolated LR images and SR images. Compared with the LR input baselines, these WCE diagnosis approaches with the interpolated LR images as input are improved with a marginal F1 increase of 0.15%, 0.19%, 0.02% and 0.10% for [43], [44], [45] and [36], respectively. Moreover, the best baseline method [36] with SR input achieves 93.27% in accuracy, worse than our framework with a 1.67% gap. This validates that simply applying resolution interpolation or SR network as a pre-processing can hardly improve the diagnosis, which also confirms the necessity of our tailor-made SR enhanced diagnosis framework.

We further replaced our MRC-Net with state-of-the-art SR methods [4]–[7] in our framework. In Fig. 6, the bicubic interpolation of LR input, denoted as ILR, serves as the baseline with accuracy of 92.52% to eliminate the impact

TABLE V

THE ABLATION STUDY OF SR AND SAI BLOCKS ON CAD-CAP DATASET.BEST AND SECOND BEST RESULTS ARE **HIGHLIGHTED** AND UNDERLINED.

Config	Accuracy (%)	F1 score (%)	Sensitivity (%)	Specificity (%)	CK (%)	MCC (%)
LR	89.11 ± 0.83	89.03 ± 0.80	89.11 ± 0.83	94.55 ± 0.42	83.66 ± 1.25	84.03 ± 1.36
HR	94.03 ± 0.20	94.00 ± 0.18	93.94 ± 0.11	97.00 ± 0.09	90.99 ± 0.27	91.05 ± 0.27
SR	93.11 ± 0.85	93.10 ± 0.84	93.11 ± 0.85	96.56 ± 0.43	89.67 ± 1.28	89.78 ± 1.37
LR+LR w/o SAI	90.74 ± 0.13	90.70 ± 0.21	90.74 ± 0.13	95.37 ± 0.06	86.11 ± 0.19	86.19 ± 0.18
LR+HR w/o SAI	94.11 ± 0.39	94.10 ± 0.39	94.11 ± 0.38	97.05 ± 0.19	91.16 ± 0.58	91.20 ± 0.56
LR+SR w/o SAI	93.64 ± 0.64	93.55 ± 0.64	93.64 ± 0.64	96.82 ± 0.32	90.33 ± 0.96	90.37 ± 0.94
LR+LR w/ SAI	92.45 ± 0.87	92.43 ± 0.85	92.48 ± 0.88	96.22 ± 0.44	88.66 ± 1.29	88.84 ± 1.27
LR+ILR w/ SAI	92.52 ± 0.71	92.53 ± 0.68	92.52 ± 0.71	96.26 ± 0.36	88.78 ± 1.07	88.83 ± 1.13
LR+HR w/ SAI	$\mathbf{95.15} \pm 0.24$	$\mathbf{95.08} \pm 0.19$	$\mathbf{95.17} \pm 0.21$	$\mathbf{97.58} \pm 0.11$	$\mathbf{92.72} \pm 0.30$	$\mathbf{92.75} \pm 0.29$
LR+SR w/ SAI (Ours)	$\underline{94.94} \pm 0.38$	$\underline{94.78} \pm 0.53$	$\underline{94.78} \pm 0.53$	$\underline{97.39} \pm 0.26$	$\underline{92.17} \pm 0.79$	$\underline{92.20} \pm 0.81$

TABLE VI

CONFUSION MATRIX OF DIAGNOSIS ON 4-FOLD CAD-CAP DATASET.

True \ Predicted	Normal	Inflammatory	Vascular lesion
Normal	**99.83**±0.29	0.00±0.00	0.17±0.29
Inflammatory	0.00±0.00	**94.33**±1.37	5.67±1.37
Vascular lesion	0.00±0.00	9.83±0.55	**90.17**±0.55

TABLE VII

PERFORMANCE AND RESOURCE ANALYSIS OF SAI BLOCKS ON CAD-CAP DATASET.

Config	Accuracy (%)	CK (%)	Param. (10^7)	FLOPs (10^{10})	Memory (10^2 MB)
Ours *w/o* SAI	93.64	90.33	**2.46**	**5.32**	**6.18**
Ours	**94.94**	**92.17**	2.60	5.43	6.50

of increasing the input resolution. In contrast, adopting SR networks improves the diagnostic performance of the entire framework, and our MRC-Net obtains the largest performance gain. This comparison further confirms that our MRC-Net is a better choice for the diagnosis framework than existing SR networks [4]–[7] from the perspective of down-stream tasks.

Furthermore, we demonstrated the confusion matrix of our framework in Table VI. Provided with the SR knowledge, our diagnosis framework can easily distinguish the normal images from abnormal ones with the sensitivity of 99.83%. Some mistakes happened between the inflammatory and vascular lesions since they may show very similar characteristics, including 9.83% vascular lesions are misjudged as the inflammatory ones and 5.67% in turn. In general, our framework produces accurate and unprejudiced predictions for WCE diagnosis.

2) Histopathology Experiment: We further conducted a consistent experiment on the PCam dataset. As shown in Table III, our diagnosis framework also achieves a superior performance on the real LR images captured at low-magnification, with the accuracy of 97.11% and F1 score of 97.03% as well as the balanced sensitivity of 96.78% and specificity of 97.28%. Besides, our diagnosis framework with both CK and MCC of 94.06%, outperforms the second best one [37] with a margin of 3.96% in CK and 3.85% in MCC. We also implemented the ResNet-101 [17] on PCam dataset, which is a upgraded version of the diagnosis branch. The performance gap between ResNet-101 and our framework, e.g., a 3.15% gap in accuracy, supports the fact that introducing SR knowledge is more effective than simply expanding the network. The comparison in Table III proves that the effectiveness of our SR enhanced diagnosis framework with a distinct advantage over state-of-the-art methods in the real LR case.

3) Ablation Study: To evaluate the proposed SR network and SAI blocks, we implemented the following networks on the CAD-CAP dataset, with results presented in Table V.

- *LR*, *HR* and *SR*: A diagnosis branch was trained and evaluated with 128×128 LR images, 512×512 HR images and super-resolved ones, respectively. For the *SR* case, the MRC-Net pretrained on EndoSR dataset served as a pre-processing step of the diagnosis branch, and the MRC-Net and diagnosis branch were jointly fine-tuned under the diagnostic supervision, which is consistent with previous work [48]. The *HR* case represents the ideal scenario without resolution degradation.
- *LR+LR w/o SAI*, *LR+HR w/o SAI* and *LR+SR w/o SAI*: Two diagnosis branches were trained and evaluated with corresponding inputs. The interaction of two branches was achieved by concatenation of feature maps.
- *LR+LR w/ SAI*, *LR+ILR w/ SAI*, *LR+HR w/ SAI* and *LR+SR w/ SAI*: Two diagnosis branches were trained and evaluated with corresponding inputs, employing the SAI blocks for interaction. The *LR+ILR w/ SAI* replaced the input of one branch in the *LR+LR w/ SAI* with the bicubic interpolated LR images. Note that the *LR+SR w/ SAI* is our SR enhanced diagnosis framework.

In Table V, the *LR* provides a baseline accuracy of 89.11%. The *SR* achieves accuracy of 93.11%, with an accuracy advantage of 4.00% over the *LR*. The SR knowledge promotes the diagnosis branch to alleviate the resolution degradation, approaching the *HR* with accuracy of 94.03%.

With the concatenation of feature maps from an additional LR diagnosis branch, the *LR* case is improved with 1.63% in accuracy, while the *HR* and *SR* cases are enhanced with only 0.08% and 0.53%, respectively. This indicates that the concatenation of multi-scale features cannot facilitate the medical diagnosis effectively. Furthermore, we replaced the concatenation with the proposed SAI blocks, which outperforms the concatenation interaction with a 1.04% and 1.30% accuracy increase in the *HR* and *SR* cases, respectively. Consequently, our SR enhanced diagnosis framework brings a 5.83% accuracy improvement to the baseline *LR*, and also outperforms the single-branch *HR* with 0.91% in accuracy. Note that the

TABLE VIII
ABLATION STUDY OF SAC AND SAR ON CAD-CAP DATASET.

SAC	SAR	Accuracy (%)	F1 (%)	CK (%)	MCC (%)
		94.06 ± 0.73	94.05 ± 0.73	91.08 ± 1.10	91.11 ± 1.15
	✓	94.52 ± 0.68	94.52 ± 0.65	91.78 ± 1.02	91.87 ± 1.05
✓		94.67 ± 0.59	94.67 ± 0.57	92.00 ± 0.88	92.04 ± 0.92
✓	✓	$\mathbf{94.94} \pm 0.38$	$\mathbf{94.78} \pm 0.53$	$\mathbf{92.17} \pm 0.79$	$\mathbf{92.20} \pm 0.81$

TABLE IX
THE IMPACT OF SR PART ON DIAGNOSIS TASK WITH SYNTHETIC LR WCE IMAGES AND REAL LR HISTOPATHOLOGY IMAGES.

Data	MRC-Net Config	Accuracy (%)	F1 (%)	CK (%)
WCE	*Scratch+Fine-tune*	93.33 ± 0.97	93.35 ± 0.95	90.03 ± 1.42
	Pretrain+Fixed	93.63 ± 1.03	93.64 ± 1.01	90.44 ± 1.55
	Pretrain+Fine-tune (Ours)	$\mathbf{94.94} \pm 0.38$	$\mathbf{94.78} \pm 0.53$	$\mathbf{92.17} \pm 0.79$
Histo.	*Scratch+Fine-tune*	95.73 ± 0.28	95.52 ± 0.27	91.46 ± 0.43
	Pretrain+Fixed	96.09 ± 0.16	96.07 ± 0.16	92.05 ± 0.33
	Pretrain+Fine-tune (Ours)	$\mathbf{97.11} \pm 0.04$	$\mathbf{97.03} \pm 0.04$	$\mathbf{94.06} \pm 0.07$

negligible performance differences between *LR+ILR w/ SAI* and *LR+LR w/ SAI* eliminate the impact of input resolution, and verify that the diagnosis improvement attributes to high-frequency details produced by SR. These ablation experiments confirm that our framework can effectively alleviate the resolution degradation problem in medical image diagnosis.

We further analyzed the trade-off between performance and resource overhead of SAI blocks. As shown in Table VII, SAI blocks significantly promote the diagnosis performance, with an increase of 1.30% in accuracy and 1.84% in CK. Correspondingly, these SAI blocks bring an overhead of only 5.69% parameters and 5.18% runtime memory to the diagnosis framework. In particular, the computation slightly increases with 2.07% FLOPs. Therefore, our designed SAI is efficient to be employed for practical diagnosis scenarios.

In our SR enhanced diagnosis framework, two kinds of constraints, sample affinity consistency (SAC) and sample affinity regularization (SAR), were developed to optimize the diagnosis task. Specifically, SAC loss in Eq. (9) introduces the semantic consistency between LR and SR diagnosis branches at multi-stage, and SAR loss in Eq. (11) minimizes the semantic gap among the mutual sample affinities of different layers. Therefore, we conducted ablation study on these two constraints in our framework when SAI blocks were kept, as shown in Table VIII. The baseline framework without any sample semantic constraint achieves the accuracy of 94.06%. Adding SAC and SAR individually can increase the accuracy by 0.61% and 0.46% respectively, and 0.88% when utilized together. Thus, imposing the reasonable constraints of sample semantics among different branches and layers can effectively promote the diagnosis performance of the entire framework.

V. DISCUSSION

A. Contribution of SR Network on Diagnosis

To investigate the contribution of the SR network to the diagnosis framework, we conducted the experiment on synthetic LR WCE images and real LR histopathology images. Specifically, we implemented the SR enhanced diagnosis framework with different optimization manners for MRC-Net. Note that the experimental results in Table IX share the same network structure, which ensures that the amount of parameters and the input resolution are consistent.

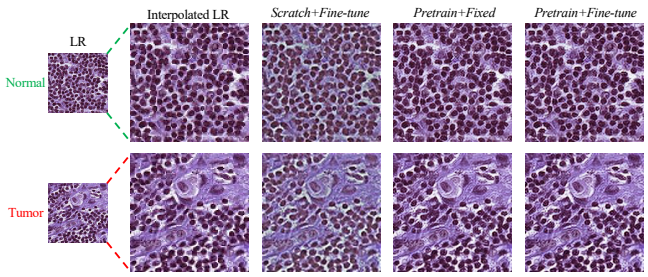

Fig. 7. The SR images of PCam dataset reconstructed by MRC-Net under different configurations.

- *Scratch + Fine-tune*: The MRC-Net was trained within the entire diagnosis framework on the diagnosis dataset from scratch, without the pretraining on the EndoSR dataset. Thus, this case does not introduce SR knowledge.
- *Pretrain + Fixed*: The MRC-Net was trained on the SR dataset, and then fixed when training the entire diagnosis framework on the diagnosis dataset. Thus, the SR knowledge obtained by pretraining MRC-Net on SR dataset is directly utilized in the diagnosis task without adaptation.
- *Pretrain + Fine-tune*: The MRC-Net was first trained on the SR dataset, and then fine-tuned on the diagnosis dataset using the gradients from the diagnostic supervision of the SR enhanced diagnosis framework. This case is actually our method in Table III.

For the synthetic LR WCE case in Table IX, the *Scratch + Fine-tune* achieves 93.33% accuracy and 90.03% CK, and the *Pretrain + Fine-tune* case increases the accuracy and CK to 94.94% and 92.17% respectively, which indicates the SR knowledge introduced by pretraining of MRC-Net can bring an increase of 1.61% in accuracy and 2.14% in CK. Compared with the *Pretrain + Fixed* with 93.63% accuracy and 90.44% CK, fine-tuning MRC-Net of the *Pretrain + Fine-tune* case further transfers the introduced SR knowledge to serve the entire framework better, with a gain of 1.31% in accuracy and 1.73% in CK. As a price, the SR performance of MRC-Net on the EndoSR dataset drops by 0.452 dB PSNR. The consistent diagnosis improvement is also observed on the real LR histopathology images, as in 4^{th} to 6^{th} rows of Table IX. Moreover, to verify the effectiveness of SR knowledge from synthetic SR dataset on real LR images, we provide both normal and tumor SR reconstruction images of PCam dataset under different MRC-Net configurations in Fig. 7. The nearest interpolated LR image shows very limited information and rough structure. As *Scratch + Fine-tune* hasn't been optimized on HistoSR dataset, *Scratch + Fine-tune* with a global skip connection can only produce a low-quality reconstruction, which represents the case without SR knowledge. When MRC-Net has been pretrained on HistoSR dataset, both *Pretrain + Fixed* and *Pretrain + Fine-tune* generate improved SR images with abundant texture and distinct structure, which reveal the SR knowledge of synthetic SR dataset is beneficial to super-resolve the real LR images. In this way, the SR knowledge is the basis of the diagnosis improvement, and the

TABLE X
THE IMPACT OF SCALE FACTOR ON SR AND DIAGNOSIS PERFORMANCE.

Data	Scale Factor	SR		Diagnosis	
		PSNR	SSIM	Accuracy (%)	F1 (%)
WCE	$2\times$	**39.925**	**0.9756**	94.07 ± 0.90	94.09 ± 0.89
	$4\times$	38.586	0.9543	**94.94**± 0.38	**94.78**± 0.53
Histo.	$2\times$	**32.754**	**0.9509**	**97.11**± 0.04	**97.03**± 0.04
	$4\times$	30.314	0.8628	95.85 ± 0.26	95.67 ± 0.58

joint of the pretrained MRC-Net and the diagnosis network benefits the framework with a large margin.

B. Impact of SR Scale Factor on Diagnosis

The scale factor of SR task is a hyper-parameter in our SR enhanced diagnosis framework, and inappropriate scale factors may degrade the diagnostic performance. To investigate the impact of SR scale factor on both WCE and histopathology diagnosis tasks, we first implemented MRC -Net with different scale factors on SR dataset, and then optimized the SR part and diagnosis part jointly for the diagnosis dataset. The input size of LR image remains the same under $2\times$ and $4\times$ scale factors, i.e., 128×128 for WCE case and 96×96 for histopathology case. We presented the SR performance of MRC-Net and the diagnostic performance of the entire framework at different scale factors in Table X. In the case of WCE images, MRC-Net achieves impressive SR reconstruction at both $2\times$ and $4\times$, with PSNR of 39.925 dB and 38.586 dB, respectively. Therefore, $4\times$ scale factor is recommended for the WCE case with more SR knowledge than the $2\times$ one, which further promotes the diagnosis with a margin of 0.87% in accuracy.

However, achieving SR with larger scale factors is difficult, which may result in the inferior quality of SR reconstruction. As the structure and texture details of histopathology images are much more complex, the $4\times$ SR on HistoSR dataset with PSNR of only 30.314 dB is not reliable enough to support the diagnosis task. In contrast, the high-quality $2\times$ SR can introduce reliable knowledge to enhance the diagnosis. The accuracy difference of 0.74% indicates the scale factor of $2\times$ is more suitable for the histopathology case.

Generally, the scale factor has two effects on the performance of the diagnosis framework. On one hand, a larger scale factor can provide more knowledge to improve the diagnostic performance more. On the other hand, it may also produce more conspicuous distortion and artifacts containing unreliable information, which interferes the subsequent diagnosis task. The selection of the scale factor should give priority to the reliability of SR task in medical image analysis field.

VI. CONCLUSION

To address the resolution degradation problem in medical image diagnosis, we propose a super-resolution enhanced diagnosis framework. Specifically, MRC-Net with RCF is devised to efficiently leverage the global and local features for SR reconstruction, and the recursion distillation scheme can promote MRC-Net using the temporal knowledge derived from a recursive one. To better exploit the information from both original LR images and the reconstructed SR ones, we propose the SAI block to exchange features with the semantic relationship among samples, as well as the consistency and regularization on sample affinity to guide the multi-scale information interaction. Extensive experiments on synthetic and real LR images confirm the effectiveness and efficiency of our framework, outperforming state-of-the-art methods by a large margin.

REFERENCES

[1] J. Ma, J. Yu, S. Liu, L. Chen, X. Li, J. Feng, Z. Chen, S. Zeng, X. Liu, and S. Cheng, "Pathsrgan: Multi-supervised super-resolution for cytopathological images using generative adversarial network," *IEEE Trans. Med. Imaging*, 2020.

[2] V. Subramanian, J. Mannath, C. Hawkey, and K. Ragunath, "High definition colonoscopy vs. standard video endoscopy for the detection of colonic polyps: a meta-analysis," *Endoscopy*, vol. 43, no. 06, pp. 499–505, 2011.

[3] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2015.

[4] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016, pp. 1646–1654.

[5] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPR Workshops*, 2017, pp. 136–144.

[6] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *ECCV*, 2018, pp. 252–268.

[7] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-cnn for image restoration," in *CVPR Workshops*, 2018, pp. 773–782.

[8] X. Zhao, Y. Zhang, T. Zhang, and X. Zou, "Channel splitting network for single mr image super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5649–5662, 2019.

[9] Z. Li, Q. Liu, Y. Li, Q. Ge, Y. Shang, D. Song, Z. Wang, and J. Shi, "A two-stage multi-loss super-resolution network for arterial spin labeling magnetic resonance imaging," in *MICCAI*. Springer, 2019, pp. 12–20.

[10] S. Khan, J. Huh, and J. C. Ye, "Deep learning-based universal beamformer for ultrasound imaging," in *MICCAI*. Springer, 2019, pp. 619–627.

[11] L. Mukherjee, H. D. Bui, A. Keikhosravi, A. Loeffler, and K. W. Eliceiri, "Super-resolution recurrent convolutional neural networks for learning with multi-resolution whole slide images," *J. Biomed. Opt.*, vol. 24, no. 12, p. 126003, 2019.

[12] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," in *CVPR*, 2018, pp. 21–30.

[13] J. Shermeyer and A. Van Etten, "The effects of super-resolution on object detection performance in satellite imagery," in *CVPR Workshops*, 2019, pp. 0–0.

[14] X. Dray, C. Li, J.-C. Saurin, F. Cholet, G. Rahmi, J. Le Mouel, C. Leandri, S. Lecleire, X. Amiot, J.-M. Delvaux *et al.*, "Cad-cap: une base de données française à vocation internationale, pour le développement et la validation d'outils de diagnostic assisté par ordinateur en vidéocapsule endoscopique du grêle," *Endoscopy*, vol. 50, no. 03, p. 000441, 2018.

[15] V. Srivastav, A. Gangi, and N. Padoy, "Human pose estimation on privacy-preserving low-resolution depth images," in *MICCAI*. Springer, 2019, pp. 583–591.

[16] D. Mahapatra, B. Bozorgtabar, and R. Garnavi, "Image super-resolution using progressive generative adversarial networks for medical image analysis," *Comput. Med. Imaging Graph.*, vol. 71, pp. 30–39, 2019.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[18] S. Arora, N. Cohen, and E. Hazan, "On the optimization of deep networks: Implicit acceleration by overparameterization," in *ICML*, 2018.

[19] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NIPS*, 2015, pp. 802–810.

[20] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *CVPR*, 2018, pp. 2482–2491.

[21] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, 2016, pp. 1874–1883.

[22] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020.

[23] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *ICCV*, 2019, pp. 0–0.

[24] Z. Chen, X. Guo, C. Yang, B. Ibragimov, and Y. Yuan, "Joint spatial-wavelet dual-stream network for super-resolution," in *MICCAI*. Springer, 2020, pp. 184–193.

[25] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *CVPR*, 2017, pp. 4133–4141.

[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[27] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *ICLR*, 2015.

[28] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *ICLR*, 2017.

[29] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *ICCV*, 2019, pp. 1365–1374.

[30] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[32] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[33] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *CVPR*, 2019, pp. 3146–3154.

[34] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, "Learning lightweight lane detection cnns by self attention distillation," in *ICCV*, 2019, pp. 1013–1021.

[35] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," in *ICCV*, 2019, pp. 8040–8049.

[36] X. Guo and Y. Yuan, "Triple anet: Adaptive abnormal-aware attention network for wce image classification," in *MICCAI*. Springer, 2019, pp. 293–301.

[37] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," in *MICCAI*. Springer, 2018, pp. 210–218.

[38] Kaggle Histopathologic Cancer Detection, https://www.kaggle.com/c/histopathologic-cancer-detection, last accessed: 2020/06/14.

[39] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[40] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *CVPR*, 2016, pp. 1637–1645.

[41] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *CVPR*, 2017, pp. 3147–3155.

[42] J. Li, Y. Yuan, K. Mei, and F. Fang, "Lightweight and accurate recursive fractal network for image super-resolution," in *ICCV Workshops*, 2019, pp. 0–0.

[43] S. Fan, L. Xu, Y. Fan, K. Wei, and L. Li, "Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images," *Phys. Med. Biol.*, vol. 63, no. 16, p. 165001, 2018.

[44] X. Jia and M. Q.-H. Meng, "A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images," in *EMBC*. IEEE, 2016, pp. 639–642.

[45] Y. Yuan, W. Qin, B. Ibragimov, B. Han, and L. Xing, "Riis-densenet: Rotation-invariant and image similarity constrained densely connected convolutional network for polyp detection," in *MICCAI*. Springer, 2018, pp. 620–628.

[46] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PloS one*, vol. 12, no. 6, 2017.

[47] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Classification of histopathological biopsy images using ensemble of deep learning networks," *arXiv preprint arXiv:1909.11870*, 2019.

[48] M. Mostofa, S. N. Ferdous, B. S. Riggan, and N. M. Nasrabadi, "Joint-srvdnet: Joint super resolution and vehicle detection network," *IEEE Access*, vol. 8, pp. 82 306–82 319, 2020.