EndoL2H: Deep Super-Resolution for Capsule Endoscopy

Yasin Almalioglu, Abdulkadir Gokce, Kagan Incetan, Muhammed Ali Simsek, Kivanc Ararat, Richard J. Chen, Nicholas J. Durr, Faisal Mahmood, Mehmet Turan*

Abstract—Wireless capsule endoscopy is the preferred modality for diagnosis and assessment of small bowel disease. However, the poor resolution is a limitation for both subjective and automated diagnostics. Enhanced-resolution endoscopy has shown to improve adenoma detection rate for conventional endoscopy and is likely to do the same for capsule endoscopy. In this work, we propose and quantitatively validate a novel framework to learn a mapping from low-to-high resolution endoscopic images. We use conditional adversarial networks and spatial attention to improve the resolution by up to a factor of 8×. Our quantitative study demonstrates the superiority of our proposed approach over Super-Resolution Generative Adversarial Network (SRGAN) and bicubic interpolation. For qualitative analysis, visual Turing tests were performed by 16 gastroenterologists to confirm the clinical utility of the proposed approach. Our approach is generally applicable to any endoscopic capsule system and has the potential to improve diagnosis and better harness computational approaches for polyp detection and characterization. Our code and trained models are available at [https://github.com/akgokce/EndoL2H](https://github.com/akgokce/EndoL2H).

Index Terms—Capsule Endoscopy, Super-Resolution, Conditional Generative Adversarial Network, Spatial Attention Network, Visual Turing

I. INTRODUCTION

MINALLY invasive capsule endoscopy has become the preferred diagnostic modality for small bowel disease. Despite successful clinical adoption, capsule endoscopy is known to have a limited diagnostic yield due to interpretation errors. These errors can be attributed to a variety of different factors including passive motion, post-procedural assessment, sparse sampling of the organ due to mechanical or power limitations, and low resolution images due to capsule camera limitations \[1\]. Several randomized controlled trials have shown the benefit of high resolution endoscopy for invasive endoscopic procedures \[2\]. A recent assessment of five independent studies found that the incremental yield of high definition colonoscopy for the detection of any polyp was 3.8% \[3\]. Recent success in deep learning has enabled the automated and objective analysis of endoscopy including depth estimation \[4\], \[5\], polyp detection and characterization. However, recent work has also shown that low resolution can affect the diagnostic performance of such algorithms \[6\]. Therefore, there is a clear need for methods that can enhance resolution of capsule endoscope for both subjective and objective analysis.

Enhancing the resolution of images by increasing the size of the optics and the sensor array is not always a feasible solution since practical cost and critical space considerations are prohibitive for many endoscopic applications. To address this issue, the computer vision community has been developing a collection of algorithms known as super-resolution, which

![Fig. 1: System overview.](image-url)
are used for generating high-resolution images from lower-resolution imaging systems. Enhanced image quality will likely lead to better disease detection, region segmentation, 3D reconstruction, visual odometry, etc. The ability to overcome fundamental resolution limits using super-resolution algorithms has shown significant progress and capability in numerous areas of medical imaging. Super-resolution (SR) is a highly challenging process of obtaining images with high resolution (HR) from low resolution (LR) camera outputs. Deep learning-based super-resolution approaches have demonstrated considerable success for a variety of image modalities. Our main contributions are listed below,

- **Spatial Attention-based Super Resolution cGAN**: We propose a spatial attention based super resolution cGAN architecture for capsule endoscopy.
- **High fidelity loss function**: We introduce EndoL2H loss which is a weighted hybrid loss function specifically optimized for endoscopic images. It combines strengths of perceptual, content, texture and pixel-based loss descriptions and improves image quality in terms of pixel values, content and texture. This maintains image quality even under high scaling factors.
- **Qualitative and quantitative study**: We conduct a detailed quantitative analysis to assess the effectiveness of our proposed approach and compare to alternative approaches. We also conduct qualitative visual Turing tests performed by 16 gastroenterologists for assessing the clinical applicability of our method.

II. RELATED WORK

Super-resolution algorithms can be classified based on various criteria such as number of images used, transformation domain (spatial or frequency domain) and color space etc. With recent advances in GPUs and data-set availability, learning-based super-resolution techniques have increasing been used. Glasner et al. make use of patch redundancies across scales within an image to achieve super-resolution. Huang et al. extend self dictionaries by further allowing for small transformations and shape variations. Gu et al. proposed an approach which uses a convolutional sparse coding that works on the entire image instead of focusing only on overlapping patches resulting in the super-resolution image being improved. Tai et al. combine an edge-directed SR algorithm relying on a gradient profile prior to the advantage of learning-based detail synthesis to regenerate higher-frequency image details. Gaussian Process Regression, Trees or Random Forests may also aid the regression performance. In the context of super-resolution, deep learning methods learn the nonlinear mapping between low-resolution and high-resolution images in an end-to-end fashion utilizing neural networks such as Dong et al. making use of super-resolution Convolutional Neural Network (SRCNN) to cover high-frequency representation in a given low-resolution like sparse-coding-based strategies with the additional advantage of joint optimization.

Generating a SR image by using a mapping learned between the set of original images and a down-sampled LR versions of that set is widely applied for medical images. For MR images, as an example, Rueda et al. uses HR and LR dictionaries that are learned from MRI. However, the fact that network performs well in the training and not in the testings, indicates an overfit. Mahapatra et al. uses progressive generative adversarial networks (P-GANs) so that more accurate detection of anatomical landmarks and pathology segmentation in medical images can be made. Numerous SR techniques have also been utilized to increase quality of images acquired by low resolution endoscopic cameras. In Hafner et al. propose a SR algorithm that is based on the projection onto convex sets (POCS) approach. Their aim is to reveal details such as mucosal structures that may not be seen on limited HD endoscope magnification. The construction is based on registration, fusion and image restoration. In the work of Köhler et al. uses a SR technique for range images acquired by a ToF (Time of Flight) sensor. In the proposed work, an HR image is generated by multiple LR images with known subpixel displacements. This is done by estimating movements of the endoscope held by a surgeon by using optical flow on RGB data from the endoscope camera. In Rupp et al. proposes an SR method in order to improve the calibration accuracy of flexible endoscopes.

III. METHODS

A. Preliminaries

Image super-resolution aims at recovering the corresponding HR image from an LR image. We model the LR image as the output of the following degradation process:

$$I_x = D(I_y; \delta),$$

where $I_y$ is the corresponding HR image, $D$ represents a degradation mapping function, and $\delta$ denotes the parameters of the degradation process. The degradation process (i.e., $D$ and $\delta$) is unknown and LR images are provided as input for the inverse degradation function. The goal is to recover the corresponding HR image $I_y$ from the LR image $I_x$, so that $I_y$ is identical to the ground truth HR image $I_y$, following the process:

$$\hat{I}_y = F(I_x; \theta),$$

where $F$ is the super-resolution model and $\theta$ represents the parameters of $F$. The degradation process is unknown and can be affected by various factors such as defocusing, compression artefacts, anisotropic degradations, sensor noise and speckle noise, etc. Accordingly, in this work, we model the degradation as a combination of several distortion effects:

$$D(I_y; \delta) = (I_y \ast \kappa) \downarrow s + n_\varsigma; \kappa, s, \varsigma \in \delta,$$

where $I_y \ast \kappa$ represents the convolution between a blur kernel $\kappa$ and the ground truth HR image $I_y$, $\downarrow s$ is a downsampling operation with the scaling factor $s$ (e.g., bicubic interpolation with antialiasing), and $n_\varsigma$ is an additive white Gaussian noise with standard deviation $\varsigma$. The objective of the super-resolution process is as follows:

$$\theta = \arg \min_\theta (\mathcal{L}(\hat{I}_y, I_y) + \lambda \Phi(\theta)),$$
where $L(\hat{I}_y, I_y)$ represents the loss function between the generated HR image $\hat{I}_y$ and the ground truth image $I_y$. $\Phi(\theta)$ is the regularization term and $\lambda$ is the trade-off parameter.

B. Proposed Method

In order to generate SR image $I^{SR}$, the following steps are used. First, a Gaussian filter and a down-sampling with factor $r$ is applied to a high-resolution image $I^{HR}$ to obtain the low-resolution version $I^{LR}$. Here, an image $I^{LR}$ with $C$ color channels is described by a real-valued tensor of size $W \times H \times C$. Since $r$ is used as down-sampling factor, $I^{HR}$, $I^{SR}$ are described as $rW \times rH \times C$. Then, we train a generator network as a feed-forward CNN $G_{\theta_G}$ parametrized by $\theta_G$ where $\theta_G = \{W_{1:LR}; b_{1:LR}\}$ denotes the weights and biases of a $L$-layer deep network and it is obtained by optimizing a SR-specific loss function $L^{SR}$. Ultimately, a generating function $\hat{G}$ estimating the correct HR image for a given LR image. For training images $I^{LR}_n$, $n = 1, \ldots, N$ with corresponding $I^{HR}_n$, $n = 1, \ldots, N$, we solve:

$$\hat{\theta}_G = \arg \min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} L^{SR}(G_{\theta_G}(I^{LR}_n), I^{HR}_n) \quad (5)$$

The typical problem in the output SR images is the lack of texture detail. The minimization of the mean squared error (MSE) between the ground truth and the reconstructed HR capsule endoscopic images is the main target of SR algorithms. This fact holds while minimizing MSE. It should also maximize the peak signal-to-noise ratio (PSNR), which is a widely used measure to evaluate SR algorithms. Unfortunately, the ability of MSE in taking perceptually relevant differences is limited due to its objective based on strict mathematical and pixel-wise formulation of image similarity ignoring human perceptual aspects of the reconstructed image. In this study, we specifically propose a novel Endol2H loss $L^{SR}$ that is a weighted combination of several loss functions to model distinct desirable characteristics of the up-sampled endoscopic SR image.

C. Generator and Discriminator

We define a generator network $G_{\theta_G}$ that tries to transform a low-resolution (LR) image to its counterpart high-resolution (HR) image, and a discriminator network $D_{\theta_D}$ that tries to distinguish generated endoscopic super-resolution images from real patient endoscopic images. Optimization of $G$ and $D$ networks is performed in an alternating manner referring to the following adversarial min-max problem:

$$\min_{\theta_G} \max_{\theta_D} \left( E_{I^{HR} \sim \text{ptrue}(I^{HR})} \log D_{\theta_D}(I^{HR}) \right) +$$

$$E_{I^{LR} \sim \text{pgen}(I^{LR})} \log (1 - D_{\theta_D}(G_{\theta_G}(I^{LR}))) \quad (6)$$

Eq[6] dictates to train the generator $G$ with the prospect of faking the differentiable discriminator $D$, whereas, at the same time, it dictates the generator to learn how to super-resolve input images into highly realistic endoscopic SR images so that $D$ gets confused and classifies the super-resolved image as real. These constraints encourage perceptually superior solutions residing in the subspace manifold of real endoscopic images in contrast to SR based on pixel-wise error minimization, such as the MSE. In an encoder-decoder network, the input is passed through a series of layers that progressively downsample, until a latent feature representation is learnt, which is sequentially upsampled by the following decoder layers [29]. All the information is back-propagated through the bottleneck feature layer, which encodes high-level information about the super-resolution process. However, in many applications, the low level information shared between the input and output is needed across the layers, e.g., LR and HR images in an SR network share the location of prominent texture pixels. To circumvent this issue, we add skip connections to concatenate all channels between each layer $i$ and layer $n-i$, where $n$ is the total number of layers.

In order to accelerate training of deep CNNs and reduce the risk of overfit, we use batch normalization to reduce internal covariate shift of networks [30]. To be concrete, we perform the normalization for each mini-batch and train two extra transformation parameters for each channel to preserve the representation ability. The batch normalization calibrates the intermediate feature distribution and mitigates the vanishing gradient problem, which allows us to use much higher learning rates and be less careful about initialization.

Let $C_k$ denote a stack of convolutional layer with $k$ filters, batch normalization and ReLU layer. $CD_k$ denotes a stack of convolutional layer with $k$ filters, batch normalization, dropout with a rate of $50\%$ and ReLU layer. All convolutions are $4 \times 4$ spatial filters applied with stride 2. Convolutional layers in the encoder and the discriminator downsample the input tensor by a factor of 2, whereas they upsample by a factor of 2 in the decoder. The encoder-decoder architecture of the generator $G$ consists of:

- **encoder:**
  - C64-C128-C256-C512-C512-C512-C512-C512
- **decoder:**
  - CD512-CD512-CD512-C512-C256-C128-C64

We add a convolutional layer with $\tanh$ non-linearity after the last layer in the decoder to map the input to the number of output channels, i.e. 3. As an exception to the above notation, we omit batch normalization in the first C64 layer of the encoder. All ReLUs in the encoder are leaky with a slope value of $\alpha = 0.2$, whereas they are non-leaky in the decoder. The activation functions of the bottleneck layer are zeroed by the batch normalization, which effectively makes the innermost layer skipped. We fix this issue by removing the batch normalization from this layer.

We follow the general shape of the U-Net architecture [31], except with skip connections between each layer $i$ in the encoder and layer $n-i$ in the decoder, which concatenate the outputs of activation layers from layer $i$ to layer $n-i$. Skip connections change the number of channels in the decoder:

- **U-Net decoder:**
  - CD512-CD1024-CD1024-CD1024-C1024-C512-C256-C128

The architecture of the discriminator $D$ is shown in Fig. 2. Our network design is inspired and modified from [32]. We use LeakyReLU activation with slope parameter $\alpha = 0.2$.
Fig. 2: Overall system architecture of EndoL2H super-resolution framework. On the left, low resolution image is fed to generator network to produce estimated high resolution image, which is then passed to discriminator to classify as whether real or fake. Discriminator network takes as both LR image and the corresponding HR image (real or generated), and tries to distinguish whether the HR image is real or fake. On the right, layers and skip-connections of attenuation U-Net is depicted in detail and below that, the flow diagram of spatial attenuation blocks are shown.

and avoid max-pooling throughout the network, following the architectural guidelines summarized by Radford et al. [33]. Discriminator network is trained to solve the maximization problem in Equation 6, which contains eight convolutional layers with an increasing number of filter kernels of size $3 \times 3$. The number of features in each convolutional layer is increased by a factor of 2 from 64 to 512 as in the VGG network [34]. We use strided convolutions to reduce the image resolution each time the number of features is doubled. The resulting 512 feature maps are followed by two fully-connected layers and a final sigmoid activation function to obtain a probability for super-resolution image classification. In order to model high-frequency information in the input LR image, we analyze the structure in local image patches. Unlike the other conventional GANs, we design a discriminator architecture using PatchGAN that penalizes structure at the scale of patches [32]. The $D$ network classifies if each $N \times N$ patch in an image is real or fake, where $N$ is the patch size that can be significantly lower than the input image size. We run $D$ convolutionally across the image and average all responses to provide the ultimate output for the input. Such a patch based discriminator design effectively models the image as a Markov random field, which assumes the pixels separated by more than a patch size $N \times N$ are conditionally independent. This Markovian assumption was previously explored in [35], which is also the common assumption in texture [36] and style [37] modelling. Therefore, our PatchGAN also models texture and style.

The attention model is a non-local convolution process. For a given input $X$, we can define the non-local operation as follows,

$$Z = f(X, X^T)g(Y)$$

where $f$ represents the relationship of each pixel to another on the input tensor $X$. For spatial attention block, we can write:

$$Z = \text{softmax} \left( \theta(X) \phi(X^T) \right) g(Y)$$

The non-local operation in spatial attention calculates relative weights of all positions on the feature maps. The dot product of $\theta(X) \phi(X^T)$ gives a measure of the input covariance, which is the degree of tendency between two feature maps at different channels. A convolution operation or channel attention model

- **discriminator:** C64-C128-C256-C512

We add a convolutional layer after the last layer to map to a 1-dimensional output, followed by a Sigmoid function. As an exception, we remove batch normalization in the first C64 layer. We use leaky ReLUs with a slope parameter $\alpha = 0.2$.

D. Spatial Attention Blocks (SAB)

Spatial Attention Blocks (SAB) are one of the major contributions of our work into L2H resolution domain. Details of the SAB are shown in Fig 2(yellow block in Fig. 2 with word SAB). The basic idea of SAB is to learn cross-correlation between features. We add after the first convolution block to extract the attention maps as input to the next down-sampling layer, which is also propagated to the up-sampling layers with the skip-connections. Inside spatial attention blocks, there are three convolution layers that decompose the input data into three components: $\theta$, $\phi$ and $g$. Afterwards, two dot product operations are performed using two of the three components. There is a short connection between input to the output so that the attention model learns the residual mapping. The spatial attention block takes all the features extracted by the convolution layer to calculate the map.

The attention model is a non-local convolution process. For a given input $X$, we can define the non-local operation as follows,
extracts features from the weighted input only in a local region while the attention model covers the whole data, which is similar to idea of the Principal Component Analysis (PCA). As shown in Fig. 4, input tensor $X$ is decomposed into $\theta(X)$ and $\phi(X^T)$. Then we vectorize the feature maps along the channel dimension so that $i$-th vector represents the feature map at $i$-th channel. Their dot product calculates the autocorrelation of the input data and the softmax operation normalizes each of the vectors to a unit vector, which corresponds to a principal axis of the input data. The dot product of $g(X)$ and the normalized vectors projects the data to a new coordinate system. The output of the softmax is the global weight matrix that measures the importance of each feature map. Unlike PCA that uses the statistical correlation of a dataset to reduce the data dimensionality by extracting the eigenvectors, the spatial attention calculates the feature correlation across the channel domain to find the principal features across the entire spatial domain.

E. Learning Objectives for Super-Resolution

In the super-resolution field, loss functions are used to measure the difference between generated HR images and ground truth HR images, and guide the model optimization. In the early studies, the pixel-wise L2 loss is widely used in the field to enforce correctness at low frequencies but later it is discovered that it cannot measure the reconstruction quality very accurately. In addition, L2 and L1 losses tend to produce blurry results on image reconstruction problems [39]. Therefore, we adopt a variety of loss functions (e.g., content loss [40], adversarial loss [41], etc.) to better measure the reconstruction error. We combine the GAN objective with a more traditional loss, such as L2 distance [29], and the content loss that encourages less blurring. The content loss enforces the low-frequency correctness, whereas the GAN discriminator models the high-frequency structure. The objective of the discriminator remains unchanged but the generator is optimized not only to fool the discriminator but also to reduce the error between the ground truth and the generated outputs. The notations in this section follow Sec. [III] except that we ignore the subscript $y$ of the target HR image $I_y$ and generated HR image $\hat{I}_y$ for brevity.

Pixel Loss. Pixel loss measures pixel-wise difference between two images and mainly includes L1 loss (i.e., mean absolute error):

$$L_{\text{pixel}_1}(\hat{I}, I) = \frac{1}{hwc} \sum_{i,j,k} |\hat{I}_{i,j,k} - I_{i,j,k}|,$$

where $h$, $w$ and $c$ are the height, width and number of channels of the evaluated images, respectively. For numerical stability, we use a variant of the pixel L1 loss, namely Charbonnier loss [38], [43], given by:

$$L_{\text{pixel}_{\text{Cha}}}(\hat{I}, I) = \frac{1}{hwc} \sum_{i,j,k} \sqrt{(\hat{I}_{i,j,k} - I_{i,j,k})^2 + \epsilon^2},$$

where $\epsilon$ is a small constant (e.g., $1e-3$).

The pixel loss constrains the generated HR image $\hat{I}$ to be close enough to the ground truth HR image $I$ on the pixel values. Comparing with L1 loss, the L2 loss penalizes larger errors but is more tolerant to small errors. In practice, the L1 loss shows improved performance and convergence over L2 loss [44]. Since the definition of PSNR (Sec. [IV-B1]) is highly correlated with pixel-wise difference and minimizing pixel loss directly maximize PSNR, the pixel loss is a widely used loss function in this field. However, since the pixel loss actually is not capable of modelling the image quality (e.g., perceptual quality [45], textures [46]), it often lacks high-frequency details and produces perceptually unsatisfying results with overly smooth textures [47].

Content Loss. To evaluate the perceptual quality of an image, the content loss is introduced into super-resolution domain [40], [48], which measures the semantic differences between images using a pre-trained image classification network. Essentially the content loss transfers the learned knowledge of hierarchical image features from the classification network to the SR network. In contrast to the pixel loss, the content loss encourages the perceptually similarity between the output image and the target image instead of forcing them to exactly match the corresponding pixel values. Thus, it produces visually more perceptible results and is also widely used in this field [40], where the VGG [34] and ResNet [49] are the most commonly used pre-trained CNNs.

We define the Content loss based on the methods introduced in Gatys et al. [50], Bruna et al. [51] and Johnson et al. [40]. The defined loss function aims to make comparisons at higher feature levels rather than at pixel level and it is based on the ReLU activation layers of the pre-trained VGG network as described in Simonyan and Zisserman [53]. Let $\phi_{i,j}$ represent the feature map in the VGG network that is obtained by the $j$-th convolution layer before the $i$-th max-pooling layer. We define the VGG loss as the Euclidean distance between the feature representations $\phi_{i,j}$ of a reconstructed image $G_{\theta_G}(I^{LR})$ and the reference image $I^{HR}$.

$$L_{\text{content}} = \frac{1}{w_{i,j}h_{i,j}} \sum_{x=1}^{w_{i,j}} \sum_{y=1}^{h_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR})_{x,y})^2, \quad (9)$$

where $w_{i,j}$ and $h_{i,j}$ are the dimensionality of the respective feature maps within the VGG network.

Texture Loss. The texture loss (a.k.a. style reconstruction loss) is used in the super-resolution domain to capture the same texture of the target image. Following [52], we regard the texture of an image as the correlations between different feature channels and define it as the Gram matrix $G^{(l)} \in \mathbb{R}^{c \times c}$, where $G^{(l)}_{ij}$ is the inner product between the vectorized feature maps $i$ and $j$ on layer $l$:

$$G^{(l)}_{ij}(I) = \text{vec}(\phi^{(l)}_{i}(I)) \cdot \text{vec}(\phi^{(l)}_{j}(I)), \quad (10)$$

where $\text{vec}(\cdot)$ denotes a vectorization operation, and $\phi^{(l)}_{i}(I)$ represents the $i$-th channel of the feature maps on layer $l$ of image $I$. Based on these definitions, we describe the texture loss as:

$$L_{\text{texture}}(\hat{I}, I; \phi, l) = \frac{1}{c_l} \sqrt{\sum_{i,j} (G^{(l)}_{ij}(\hat{I}) - G^{(l)}_{ij}(I))^2}. \quad (11)$$
By employing texture loss, the SR model creates more realistic textures and produce visually more satisfactory results \cite{46}. Despite this, determining the size of the patch to match the target texture is still empirical. Too small patches lead to artifacts in textured region, whereas too large patches induce artifacts throughout the entire image because texture statistics are averaged over regions of varying textures.

**Adversarial Loss.** In recent years, the GANs \cite{53} have been more and more popular and used in various vision-related tasks. GANs consist of a generator performing generation (e.g., SR image), and a discriminator which takes the generated output and real instances as input and discriminates whether each input comes from the target distribution. The training alternates between two steps:

- freeze the generator and train the discriminator to better discriminate,
- freeze the discriminator and train the generator to generate realistic outputs and fool the discriminator.

Through adversarial iterations, the trained generator produces outputs consistent with the true target distribution, while the discriminator cannot distinguish the generated data from the real data anymore.

We treat the generator as the SR model, and additionally define a discriminator to judge whether the generated SR image is indistinguishable from the true SR images. Unlike the cross entropy-based GAN loss proposed in SRGAN \cite{41}, we use adversarial loss based on least square error for more stable training process and higher quality results \cite{7, 54}, given by:

\[
L_{gan, \text{ls}}(\hat{I}; D) = (D(\hat{I}) - 1)^2, \tag{12}
\]

\[
L_{adv}(\hat{I}, I_\text{s}; D) = (D(\hat{I}))^2 + (D(I_\text{s}) - 1)^2. \tag{13}
\]

**EndoL2H Loss.** Several loss types of generators have been investigated to guide the EndoL2H network convergence. The adversarial loss, L1-norm pixel loss, content loss and texture loss are the main components of the loss function as described in the previous sections. The overall loss function (EndoL2H) of our network is:

\[
L_{L2H} = \alpha(n) L_{adv} + (1 - \alpha(n))(1 - \beta_1)(1 - \gamma_1) \cdot L_{cha} + \gamma_1 \cdot L_{content} + \beta_1 \cdot L_{texture}. \tag{14}
\]

The $L_{adv}$ is the vanilla GAN adversarial loss described in Eq. 12. $L_{cha}$ represents the L1-norm between the target high-resolution and the generated high-resolution images described in \cite{8}. $L_{texture}$ represents the texture loss to capture the textures in the SR image, described in \cite{11}.

The L1 and L2 lead to rather vaguer results, they tend to give satisfactory results in lower frequency features leaving the detection of high frequency features to discriminator. Consequently, feature detection in wider ranges of frequencies is guaranteed. Without L1 loss, conditional GANs are used to generate fake images that discriminator only fails to distinguish. However, addition of L1 loss forces conditional GANs also to generate resulting images that are close to reference image. With the combination of L1 loss and usual conditional GAN loss, weaknesses of each loss functions is circumvented and they work in a complementary manner. Similarly, the target of the content loss is to describe the visual attribute. A version of the VGG-19 using pre-trained weights on ImageNet as a universal feature extractor is preferred in this paper, as described in \cite{9}. The variable $\alpha$ dynamically modulates the influence of the adversarial loss into the overall loss. Similarly, $\beta_1$ and $\gamma_1$ control the importance of the texture and content losses, respectively.

**IV. EXPERIMENTAL SETUP AND EVALUATION METRICS**

**A. Dataset**

We used a portion of Kvasir dataset for training and testing. The dataset is gathered using endoscopic equipment at Vestre Viken Health Trust (VV) in Norway and composed of 8000 images taken from several organs including bowel and stomach (v2) \cite{55} and the portion we used consists of 1500 images of it. We divided the dataset into three subsets consisting of A images (1100) for training, B images for validation 100 and C images 300 for test. To avoid overfitting, the dataset is shuffled 10 times to obtain different training and test subsets, keeping the validation set constant. We obtained LR images by downsampling HR ground truth images (RGB, C = 3) using bicubic kernel with downsampling factor $r = 8$, resulting in 125 $\times$ 125 resolution input images. Random jitter was applied by resizing the input images to 150 $\times$ 150 and then randomly cropping back to the original size. This augments data to faithfully represent existing state-of-the-art endoscopic cameras which has resolutions varying from 150 $\times$ 150 to 400 $\times$ 400 pixels due to energy and area limitations.

**B. Image Quality Assessment**

1) **Peak Signal-to-Noise Ratio:** Peak signal-to-noise ratio (PSNR) is a full-reference image quality measure of pixel-wise similarity. It is defined by maximum possible pixel value $L$ and the mean-squared-error $MSE$ between image and the formula is given as

\[
\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (I(i) - \hat{I}(i))^2, \tag{15}
\]

\[
\text{PSNR} = 10 \cdot \log_{10}\left(\frac{L^2}{\text{MSE}}\right). \tag{16}
\]

where $I$ is the reference image and $\hat{I}$ is the reconstructed image and both consists of N pixels. In practice, $L$ is fixed and equals to 256 for 8-bit encoded images ($= 2^8$), and the PSNR is determined by mean-squared-error of pixel values of two images while ignoring structural and perceptual similarity. Higher PSNR values represent greater pixel-wise similarity. Even though it often fails to assess image quality of SR models, it is still widely used image quality metric and it should be taken into account when comparing performances with other models.

Nevertheless; as it can be seen from the Table \cite{4} EndoL2H over-performs SRGAN and bicubic interpolation in terms of PSNR with higher mean. In terms of pixel-wise values; Figure \cite{6} shows examples of three intensity profiles of red, green and blue channels for a sample image tuple composed of a ground truth and corresponding LR and 3 different HR image...
Fig. 3: Image tuples are shown in the first row represent ground truth and HR images obtained by bicubic interpolation, SRGAN and EndoL2H. The second row shows the zoomed versions of the areas depicted inside the green squares above. Notice that EndoL2H was able to estimate finer details more accurately than two other methods.

2) Structural Similarity: Structural similarity index (SSIM) is a well-known full-reference quality metric that takes into account structural composition of pixels. By making use of luminance, contrast and structure values of images; SSIM aims to measure perceived quality by human visual system. It gets values between 0 and 1 and it is more sensitive to high frequency content -such as an edge- than PSNR. Although it is considered to be a better assessment method in finding perceptual quality of SR images than PSNR, it may still disfavor GAN generated HR images.

3) Learned Perceptual Image Patch Similarity (LPIPS): As discussed in content loss (Sec. III-B), although PSNR and SSIM are two widely used quality metrics in image reconstruction, they often fail to catch perceptually linked differences. An image having high PSNR or SSIM score can be blurry than an image generated by a generative network and with less score. To evaluate the performance of our method and to compare with others in terms of perceptual features quantitatively, we used an open source framework ‘Learned Perceptual Image Patch Similarity’ (LPIPS) provided by [56]. Similar to our content loss, LPIPS utilizes extracted feature maps from convolutional layers of well-known deep neural networks pre-trained on classification task which then trained on a large dataset consisting of image patches with human opinion scores for perceptual distance evaluation. It than infers a perceptual distance between a reference image and a transformed version of it analyzing them in small patches. We performed our calculations using version 0.1 of LPIPS repository with pre-trained AlexNet model and default input arguments. Mean and standard deviation of perceptual distances for HR images generated by EndoL2H, SRGAN and bicubic interpolation are given in Table I. As it can be seen from the Table I, EndoL2H produced HR images have lower mean perceptual distances, indicating that EndoL2H generated images are more similar to ground truth counterparts in terms of perceptual features.

4) Mean opinion score (MOS) evaluations: Mean opinion score (MOS) is a survey based subjective image quality metric in which humans participants are asked to choose a score regarding a set of specified features (e.g., sharpness). Although MOS testing can asses perceptual quality of images better than objective methods such as PSNR and SSIM, it introduces human error factor and it is susceptible to bias and also it can be time consuming and expensive. Nevertheless, some SR models might have low PSNR and SSIM scores whereas the perceptual quality of HR images could be still high, in which case MOS testing should be employed. We performed a MOS test to subjectively quantify the sharpness, suitability for diagnosis and detail level of the reconstructed super-resolution images of different approaches. Specifically, we asked 16 gastroenterologists to assign an integral score to 15 output images from 1 (bad quality) to 5 (excellent quality) for each of 3 comparison metrics and reconstructed images, respectively.

V. EVALUATIONS AND RESULTS

A. Optimization and inference

We trained and tested neural networks on NVIDIA® Tesla® V100 instance of Google GPU cluster taking 8h for each training and 81h in total. EndoL2H was trained with $10^5$ update iterations at a learning rate of $10^{-4}$ and another $10^5$ iterations at a lower rate of $10^{-5}$. We follow the approach
Fig. 4: PSNR, SSIM and Perceptual distance (LPIPS) comparisons of EndoL2H (green) and SRGAN (orange). For PSNR and SSIM scores, higher is better. For LPIPS distance metric, lower is better.

Fig. 5: SSIM heatmaps for high resolution image obtained by a sample LR image using bicubic interpolation, SRGAN and EndoL2H, respectively. Each point represents the local SSIM value for $11 \times 11$ Gaussian window. Red color denotes lower SSIM values (low structural similarity with the original image) and blue color represents high SSIM values (high structural similarity with the ground truth).

**TABLE I: Mean and Standard Deviations for PSNR, SSIM and LPIPS**

|        | ENDOL2H     | ENDOL2H-Pix  | ENDOL2H-Content | ENDOL2H-Texture | SRGAN       | Bicubic     |
|--------|-------------|--------------|-----------------|-----------------|-------------|-------------|
| PSNR   | $34.14 \pm 2.82$ | $32.44 \pm 2.14$ | $31.37 \pm 1.69$ | $30.89 \pm 1.23$ | $33.68 \pm 2.32$ | $28.13 \pm 2.12$ |
| SSIM   | $0.80 \pm 0.062$ | $0.88 \pm 0.075$ | $0.87 \pm 0.082$ | $0.88 \pm 0.076$ | $0.84 \pm 0.064$ | $0.68 \pm 0.091$ |
| LPIPS  | $0.18 \pm 0.056$ | $0.19 \pm 0.042$ | $0.19 \pm 0.065$ | $0.20 \pm 0.053$ | $0.21 \pm 0.080$ | $0.28 \pm 0.072$ |

**TABLE II: MOS Results according to Parameters.**

| MOS Sharpness Results | MOS Suitability for Diagnosis Results | MOS Detail Level Results |
|-----------------------|--------------------------------------|-------------------------|
| Image                 | EndoL2H     | SRGAN      | Bicubic Interpolation | EndoL2H     | SRGAN      | Bicubic Interpolation | EndoL2H     | SRGAN      | Bicubic Interpolation |
| Mean                  | 4.43        | 4.11        | 2.12               | 4.60        | 4.47        | 2.34               | 4.60        | 4.20        | 2.11               |
| Std.                  | 0.34        | 0.47        | 0.47               | 0.29        | 0.34        | 0.39               | 0.27        | 0.36        | 0.24               |
| Max.                  | 4.9         | 4.9         | 2.7                | 4.9         | 4.8         | 3.1                | 4.9         | 4.8         | 2.5                |
| Min.                  | 3.8         | 3.3         | 1.3                | 3.8         | 3.8         | 1.8                | 3.9         | 3.6         | 1.8                |

from Goodfellow et. al. [53] to optimize our EndoL2H network by alternating between one gradient descent step on $D$, then one step on $G$. Following the standard GAN approach, we train to maximize $\log D(x, G(x, z))$ rather than training $G$ to minimize $\log(1 - D(x, G(x, z)))$. Dividing the objective by 2 while optimizing $D$, we slow down the rate at which $D$ learns relative to $G$. We use the Adam optimizer [58], with momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. At inference time, we run the generator net in exactly the same manner as during the training phase. Unlike usual procedure, we apply dropout at test time, and we apply batch normalization using the statistics of the test batch, rather than aggregated statistics of the training batch. Batch normalization approach, when the batch size is set to 1, has been termed instance normalization and has been demonstrated to be effective at image generation tasks. All networks were trained from scratch. Weights were initialized from a Gaussian distribution with mean 0 and standard deviation 0.02.

**B. Investigation of EndoL2H loss**

We investigated the effect of loss choices explained in [III-E] for the EndoL2H network. Specifically we investigate $\mathcal{L}_{L2H} = \alpha \mathcal{L}_{adv} + (1 - \alpha) \mathcal{L}_X$ for the following losses $\mathcal{L}_X$:

- EndoL2H-Pix: $\mathcal{L}_{Cha}$, to investigate the adversarial network with the standard pixel loss.
- EndoL2H-Content: $\mathcal{L}_{content}$ a loss defined on feature maps representing lower-level features [59].
- EndoL2H-Texture: $\mathcal{L}_{texture}$ a loss defined on the correlations between feature channels with more potential to focus on the texture of the images.

Quantitative results are listed in Table II. Even combined with the adversarial loss, EndoL2H-Pix provides reconstructions...
with the highest PSNR values that are, however, perceptually rather smooth and less convincing than results achieved with a loss component more sensitive to visual perception. This is caused by competition between the pixel-based loss and the adversarial loss. We further attribute minor reconstruction artifacts, which we observed in a minority of EndoL2H-Pix-based reconstructions, to those competing objectives. We could not determine a significantly best loss function for EndoL2H with respect to MOS score on the sharpness. However, EndoL2H significantly outperformed other variants on suitability for diagnosis in terms of MOS. Although EndoL2H-Texture yields better texture details than the other variants with higher MOS score on the detail level (see Appendix Table III), we observed a trend that using higher level VGG feature maps with EndoL2H-Content achieving higher scores when compared to the lower level features of VGG.

C. Performance of EndoL2H network

We quantitatively and qualitatively analysed reconstructed super-resolution images and compared the performance of the EndoL2H algorithm with the state-of-the-art (e.g., SRGAN) and standard (bicubic interpolation) approaches. Figure 5 demonstrates output super-resolution images for each method and the ground truth for one test sample. Quantitative assessment scores verify the effectiveness of EndoL2H against SRGAN by in terms of PSNR, SSIM, LPIPS for 150 images, respectively, and the statistics of those figures together with the scores of bicubic interpolation and EndoL2H loss types described in Table II. These results show that the EndoL2H outperforms SRGAN for most of the cases confirming EndoL2H raises the bar for a higher level on the benchmark dataset. MOS results for all reference methods on the Kvasir dataset are listed in Table II. It can be seen that all differences in MOS are highly significant on the Kvasir dataset (with a mean score of 4.43, 4.60 and 4.60 for each success criteria, respectively). Thereby, EndoL2H sets a new state-of-the-art for realistic endoscopic SR images outperforming all reference methods by a large margin. Appendix Table III shows that all differences in MOS are highly significant on the Kvasir dataset, except EndoL2H-Texture vs. EndoL2H-Content.

D. Conclusion

We have introduced a deep super-resolution network EndoL2H specifically designed and optimized for capsule endoscopic images. The proposed network combines content, texture and pixel loss functions with an adversarial loss by training a cGAN. Our investigation on network structure indicates that deeper networks to be more beneficial in super-resolution tasks. We believe that the feature maps of deeper layers mainly focus on the content while leaving the adversarial loss focusing on texture details. We further believe that the attention U-NET design has a substantial impact on the performance of deeper networks and decreases the EndoL2H loss, respectively. This; however, comes at the cost of longer training and testing times due to convergence challenges. We demonstrate the performance of EndoL2H using quantitative and qualitative tests for (8×) upsampling factor and compared it with the state-of-the-art reference methods. The test results prove that reconstructions of EndoL2H are perceptually more plausible in terms of diagnosis applications and have higher similarity scores with respect to corresponding original images than the reconstructions obtained with the compared methods.

References

[1] J. Park, Y. Hwang, J.-H. Yoon, M.-G. Park, J. Kim, Y. J. Lim, and H. J. Chun, “Recent development of computer vision technology to improve capsule endoscopy,” Clinical endoscopy, vol. 52, no. 4, p. 328, 2019.
[2] M. Pellissé, G. Fernández-Esparra, A. Cárdenas, O. Sendino, E. Ricart, E. Vaquero, A. Z. Gimeno-García, C. R. de Miguel, M. Zahalza, A. Ginés et al., “Impact of wide-angle, high-definition endoscopy in the diagnosis of colorectal neoplasia: a randomized controlled trial,” Gastroenterology, vol. 135, no. 4, pp. 1062–1068, 2008.
[3] 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.
[4] F. Mahmood, R. Chen, and N. J. Durr, “Unsupervised reverse domain adaptation for synthetic medical images via adversarial training,” IEEE transactions on medical imaging, vol. 37, no. 12, pp. 2572–2581, 2018.
[5] F. Mahmood and N. J. Durr, “Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy,” Medical image analysis, vol. 48, pp. 230–243, 2018.
[6] C. F. Sabottke and B. M. Spieler, “The effect of image resolution on deep learning in radiography,” Radiology: Artificial Intelligence, vol. 2, no. 1, p. e190015, 2020.
[7] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, “Deep endovo: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots,” Neurocomputing, vol. 275, pp. 1861–1870, 2018.
[8] D. Fischer, R. Schreiber, D. Levi, and R. Eliakim, “Capsule endoscopy: the localization system,” Gastrointestinal Endoscopy Clinics, vol. 14, no. 1, pp. 25–31, 2004.
[9] T. Nakamura and A. Terano, “Capsule endoscopy: past, present, and future,” Journal of gastroenterology, vol. 43, no. 2, pp. 93–99, 2008.
[10] N. Ben-Eliezer, M. Irani, and L. Friedman, “Super-resolved spatially encoded single-scan 2d mri,” Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 63, no. 6, pp. 1594–1600, 2010.
[11] D. Kouame and M. Ploquin, “Super-resolution in medical imaging: An illustrative approach through ultrasound,” in 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, 2009, pp. 249–252.
[12] P. Akhtar and F. Azhar, “A single image interpolation scheme for enhanced super resolution in bio-medical imaging,” in 2010 4th International Conference on Bioinformatics and Biomedical Engineering. IEEE, 2010, pp. 1–5.
[13] M. D. Robinson, S. J. Chiu, C. A. Toth, J. A. Izatt, J. Y. Lo, and S. Farsiu, “New applications of super-resolution in medical imaging,” in Super-Resolution Imaging. CRC Press, 2017, pp. 401–430.
[14] D. Huang and H. Liu, “A short survey of image super resolution algorithms,” Journal of Computer Science Technology Updates, vol. 2, pp. 19–29, 2015.
[15] F. Wang, H. S. S. Lai, L. Liu, P. Li, H. Yu, Z. Liu, Y. Wang, and W. J. Li, “Super-resolution endoscopy for real-time wide-field imaging.” Opt. Express, vol. 23, no. 13, pp. 16803–16811, Jun 2015. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-23-13-16803
[16] D. Glasner, S. Bagom, and M. Irani, “Super-resolution from a single image,” in 2009 IEEE 12th International Conference on Computer Vision, Sep. 2009, pp. 349–356.
[17] J.-B. Huang, A. Singh, and N. Ahuja, “Single image super-resolution from transformed self-exemplars,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5197–5206.
[18] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, “Convolutional sparse coding for image super-resolution,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1823–1831.
### APPENDIX A. MOS RESULTS

#### TABLE III: MOS Results: Summary

#### MOS SHARPNESS RESULTS

|       | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | #11 | #12 | #13 | #14 | #15 | Mean | Std. | Max. | Min. |
|-------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| EndoL2H | 4.3 | 4.8 | 4.4 | 4.5 | 4.8 | 4.2 | 4.1 | 4.9 | 4.1 | 4.2  | 3.8  | 4.2  | 4.9  | 4.6  | 4.7  | 4.43 | 0.34 | 4.9  | 3.6  |
| EndoL2H-Texture | 3.5 | 4.4 | 3.4 | 3.6 | 4.0 | 3.5 | 3.5 | 4.0 | 3.8 | 4.0  | 3.5  | 4.0  | 4.0  | 4.3  | 3.3  | 3.79 | 0.34 | 4.4  | 3.3  |
| SRGAN | 3.3 | 4.7 | 3.7 | 3.9 | 4.1 | 3.8 | 3.7 | 4.5 | 4.4 | 4.1  | 3.9  | 4.8  | 4.3  | 4.9  | 3.6  | 4.11 | 0.38 | 4.4  | 3.3  |
| Bicubic Interpolation | 2.3 | 1.7 | 2.4 | 2.3 | 2.7 | 2.6 | 1.7 | 2.4 | 2.3 | 2.7  | 1.3  | 2.3  | 2.1  | 1.3  | 1.3  | 2.12 | 0.47 | 2.7  | 1.3  |

#### MOS SUITABILITY FOR DIAGNOSIS RESULTS

|       | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | #11 | #12 | #13 | #14 | #15 | Mean | Std. | Max. | Min. |
|-------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| EndoL2H | 4.7 | 4.9 | 4.5 | 4.6 | 4.7 | 4.6 | 4.8 | 4.9 | 4.3 | 3.8  | 4.3  | 4.8  | 4.5  | 4.6  | 4.6  | 4.60 | 0.29 | 4.9  | 3.8  |
| EndoL2H-Texture | 4.5 | 4.6 | 4.4 | 4.3 | 4.7 | 4.6 | 4.4 | 4.6 | 4.5 | 4.4  | 4.0  | 4.5  | 4.6  | 4.6  | 4.3  | 4.47 | 0.18 | 4.7  | 4.0  |
| SRGAN | 3.8 | 4.7 | 4.4 | 4.2 | 4.4 | 4.1 | 3.9 | 3.8 | 4.4 | 4.5  | 4.1  | 4.8  | 4.6  | 4.5  | 3.8  | 4.27 | 0.34 | 4.8  | 3.8  |
| Bicubic Interpolation | 2.6 | 2.5 | 2.9 | 3.1 | 2.6 | 2.4 | 2.1 | 1.9 | 2.3 | 2.4  | 1.9  | 2.7  | 2.1  | 1.8  | 1.9  | 2.34 | 0.39 | 3.1  | 1.8  |

#### MOS DETAIL LEVEL RESULTS

|       | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | #11 | #12 | #13 | #14 | #15 | Mean | Std. | Max. | Min. |
|-------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| EndoL2H | 4.8 | 4.9 | 4.5 | 4.4 | 4.7 | 4.3 | 4.5 | 4.9 | 4.6 | 3.9  | 4.5  | 4.6  | 4.9  | 4.8  | 4.7  | 4.60 | 0.27 | 4.9  | 3.9  |
| EndoL2H-Texture | 4.6 | 4.8 | 4.2 | 4.3 | 4.4 | 4.3 | 4.3 | 4.6 | 4.4 | 4.5  | 4.6  | 4.7  | 4.7  | 4.8  | 4.55 | 0.20 | 4.8  | 4.2  |
| SRGAN | 4.4 | 4.8 | 4.0 | 4.3 | 4.2 | 4.0 | 4.2 | 4.0 | 4.3 | 4.2  | 4.4  | 4.6  | 4.4  | 4.6  | 4.6  | 4.30 | 0.24 | 4.8  | 4.0  |
| Bicubic Interpolation | 4.2 | 4.8 | 4.3 | 4.2 | 3.9 | 3.9 | 3.8 | 4.4 | 4.3 | 3.8  | 3.9  | 4.3  | 4.4  | 4.3  | 4.4  | 4.20 | 0.36 | 4.8  | 3.6  |

#### APPENDIX B. RGB CHANNEL COMPARISONS

![Fig. 6: Red, green and blue channels of a sample image is shown for a pair of RL and SR generated by EndoL2H. Each channel is depicted as an RGB image with only one non-zero respective color channel. Intensity profile comparisons of red, green and blue channels for same tuples are also provided.](image-url)
Fig. 7: Green channel of different sample of images are shown for a pair of RL and SR generated by EndoL2H. Intensity profile comparisons of green channel for various tuples are also provided.
APPENDIX D. TEST RESULTS

Fig. 8: Image tuples are shown in the first and third rows represent ground truth and HR images obtained by bicubic interpolation, SRGAN and EndoL2H. Second and fourth rows show zoomed versions of the areas depicted inside the green squares at above rows. Notice that EndoL2H was able to estimate finer details more accurately than two other methods.
Fig. 9: Image tuples are shown in the first and third rows represent ground truth and HR images obtained by bicubic interpolation, SRGAN and EndoL2H. Second and fourth rows show zoomed versions of the areas depicted inside the green squares at above rows. Notice that EndoL2H was able to estimate finer details more accurately than two other methods.
Fig. 10: SSIM heatmaps for high resolution image obtained by a sample LR image using bicubic interpolation, SRGAN and EndoL2H, respectively. Each point represents the local SSIM value for $11 \times 11$ Gaussian window. Red color denotes lower SSIM values (low structural similarity with the original image) and blue color represent high SSIM values (high structural similarity with the original image).