Improved Super Resolution of MR Images Using CNNs and Vision Transformers

Dwarikanath Mahapatra
Inception Institute of AI, UAE

Abstract. State-of-the-art magnetic resonance (MR) image super-resolution methods (ISR) using convolutional neural networks (CNNs) leverage limited contextual information due to the limited spatial coverage of CNNs. Vision transformers (ViT) learn better global context that is helpful in generating superior quality HR images. We combine local information of CNNs and global information from ViTs for image super resolution and output super resolved images that have superior quality than those produced by state of the art methods. We include extra constraints through multiple novel loss functions that preserve structure and texture information from the low resolution to high resolution images.

Keywords: MRI, super resolution, disentanglement, CNN, ViT

1 Introduction

Image super-resolution (ISR) takes low resolution (LR) image inputs and reconstructs its corresponding high resolution (HR) version thus enabling detailed examination of interesting regions. This is particularly relevant for medical image analysis where physics of the imaging systems limits the spatial resolution of radiological images (e.g., MRI, X-ray) since obtaining HR images requires longer scanning time, and leads to lower signal-to-noise ratio and smaller spatial coverage [16, 158, 156, 143, 123, 119, 102, 101, 97, 84, 74, 199, 147, 95, 71]. HR images provide more detailed information about local structures and textures resulting in higher accuracy in disease diagnosis and planning [18, 55, 70, 113, 97, 98, 114, 117, 67, 66, 69, 125, 65, 141, 139, 138, 150, 60]. Since originally acquired LR images pose challenges for accurate analysis it is important to have a reliable ISR method.

Recent works demonstrate the potential of convolutional neural networks (CNNs) in generating HR images by using SRCNN [25, 26, 21, 154, 85, 161, 173, 118, 86, 46, 180, 89, 10, 87, 180, 89, 10, 87, 50, 13, 187, 109, 112], residual learning in VDSR (Very Deep Super Resolution) [27, 166, 165, 179, 177, 172, 122, 176, 121, 198, 94, 93, 99, 119, 146, 94, 116, 170, 172, 115, 127, 169, 168, 96], and the information distillation network (IDN) [39, 192, 145, 148, 134, 124, 62, 64, 183, 63, 61, 140, 134, 136, 137, 135, 133, 132, 126, 123, 133, 143, 120] leverage hierarchical features in residual deep networks (RDN) while [8] combine 3D dense networks and adversarial learning for MRI super resolution. MR images have
inherent characteristics such as repeating structural patterns making them less complex than natural images. Secondly, they have a large proportion of background pixels. Since most approaches give the background and foreground equal importance it does not lead to good feature learning. Also, CNN methods capture mostly local context information and do not explore the global aspects. Zhang et al. in [191,153,152,155,157,24], [22,130,151], [79,51,78,49], [42,43,11], [90,77,88], [168,8,75,76] propose a squeeze and excitation network to capture the global characteristics thus leading to improved super resolution output. However, squeeze and excitation relies on CNN features to capture global context which is not optimal.

Vision transformers (ViT) [27] are an exciting new development that effectively capture long range contextual information from images. Given sufficient training data ViTs have been shown to outperform state of the art CNN based methods for classification and segmentation. In this work we propose to combine CNNs and ViT for performing super resolution of MR images. CNNs learn local details while ViT captures the global context much better than previously proposed methods. Inherent to the ViT is a self attention module that focuses on the important parts of the image and thus improves SR quality. Our method also uses feature disentanglement to improve super resolution.

2 Related Work

MR Image Super-resolution: ISR has been widely applied to MR images [171,100], and spectroscopy MRI [10,11,108,83], [52,33,128], [82,73,80,6,100], [101,105,5,104,4]. Initial methods achieved multiple frame image super resolution via alignment of multiple noisy LR images which proved to be very challenging [194]. Recent deep learning based ISR approaches show superior performance for MR image super resolution [19,138,143] but use large models that pose challenges in real world settings. Zhang et al. in [191] propose a squeeze and excitation attention network as part of a lightweight model for ISR, [29] achieve multi contrast MRI super resolution using multi stage networks. [37] use graph convolution networks for MRI super resolution, while in other related work recent methods have proposed hybrid-fusion networks for Multi-modal synthesis of MRI [196,106,107], [103,102,124], [111,31,1], [129,35,167], and [20] synthesize multi-contrast MRI using conditional GANs.

Attention Mechanism: Attention mechanisms enables adaptive resource allocation by focusing on important image regions [36] and are popular for many tasks like image recognition [2] and image captioning [188], as well as ISR [38,192]. They can be highly effective for MRI super resolution due to repeating patterns of relatively simpler structures and less informative background.

Vision Transformers: Dosovitskiy et al. [27] demonstrate state-of-the-art performance on image classification datasets using large-scale pre-training and fine-tuning, and [14,197] use ViTs for object detection. Hierarchical vision transformers with varying resolutions and spatial embeddings [57,184] have been used to reduce feature resolution, while [28] demonstrate success in high resolution
image synthesis. Recent work on transformer-based models for 2D image segmentation include the SETR model that uses a pre-trained transformer encoder with different CNN decoders [195] for multi-organ segmentation in [15], and a transformer-based axial attention mechanism for 2D medical image segmentation [181]. Hatamizadeh et al. propose UNETR [34] for 3D medical image segmentation using transformers as the main encoder of a segmentation network and directly connecting to the decoder via skip connections. For 3D medical image segmentation, [185] use a backbone CNN for feature extraction, a transformer to process the encoded representation and a CNN decoder for predicting segmentation outputs, while [185] use transformers in the bottleneck of a 3D encoder-decoder CNN for semantic brain tumor segmentation. However, none of the methods use ViT for medical image super resolution.

Motivation And Contribution: Context information is especially relevant for medical ISR since they provide additional cues to generate superior quality HR images. Our contributions are: 1) We combine CNNs and ViTs for image super resolution. Local contextual cues from CNNs and global information from ViTs result in superior quality super resolved images than those produced by state of the art methods. A pre-trained ViT is finetuned using self supervised learning. 2) Using multiple loss functions we incorporate extra constraints that preserve structural and semantic information in the generated super resolved image. 3) By comparing with results from [37] we also demonstrate our method’s better ability to learn global features compared to graph based super resolution methods.

3 Method

3.1 Overview

Given a low resolution (LR) image $x \in \mathbb{R}^{N \times N}$ our objective is to train a model that outputs a high resolution (HR) image $y \in \mathbb{R}^{M \times M}$, where $M > N$. Figure 1 shows the workflow of our proposed method. The LR image $x$ goes through a generator network consisting of a series of convolution blocks and an upsampleler that increases the image dimensions from $N \times N$ to $M \times M$. The discriminator module ensures $y$ satisfies the following constraints.

1. The HR and LR image should have similar semantic characteristics since a higher resolution version should not alter image semantics. For this purpose we disentangle the image into structure and texture features, and ensure their respective semantic information is consistent across both images.
2. The HR image should preserve global and local context of the original LR image. To achieve it we use features extracted using a pre-trained ViT to effectively capture the relations in LR image and ensure this relationship is preserved in the HR image.
Fig. 1. (a) Workflow of our proposed method. LR images goes through a generator to get HR image, and multiple loss functions ensure that semantic information of the LR image is preserved in the HR image. (b) Architecture of feature disentanglement network using swapped autoencoders.

3.2 Vision Transformers

Vision transformers play an important role in our super resolution framework by serving as a robust and accurate feature extractor that integrates long range context and structural information. We use the ViT from UNETR [34] pre-trained for MR image segmentation and fine tune it for our task. We briefly describe the architecture below (for full details please refer to Appendix B) and also explain our modifications. UNETR uses the contracting-expanding pattern consisting of a stack of transformers as the encoder which is connected to the decoder using skip connections. A 1D sequence from the 3D input volume $x \in \mathbb{R}^{H \times W \times D \times C}$ with image dimension $(H, W, D)$ and $C$ input channels is created by dividing $x$ into $N = (H \times W \times D) / P^3$ flattened non-overlapping patches of size $P \times P \times P$ and denote this set as $x_v$. A linear layer projects the patches onto a $K$ dimensional embedding space. To preserve spatial information a 1D learnable positional embedding $E_{pos} \in \mathbb{R}^{N \times K}$ is added to the projected patch embedding $E \in \mathbb{R}^{P^3 \times C \times K}$ as $z_0 = [x_1 E; x_2 E; \cdots; x_N E] + E_{pos}$. Then multiple transformer blocks [27] are used that have multi-head self-attention (MSA) and multilayer perceptron (MLP) sublayers according to

$$
\begin{align*}
    z'_i &= MSA(\text{Norm}(z_{i-1})) + z_{i-1}, i = 1 \cdots L \\
    z_i &= MLP(\text{Norm}(z'_i)) + z'_i, i = 1 \cdots L
\end{align*}
$$

(1)

where $\text{Norm}()$ denotes layer normalization [3]. MLP has two linear layers with GELU activation functions, $i$ denotes intermediate block and $L$ denotes transformer layers. $SA$ maps a query ($q$) and the corresponding key ($k$) and value ($v$) representations in a sequence $z \in \mathbb{R}^{N \times K}$. Attention weights ($A$) measure similarity between elements in $z$ and their key-value pairs according to $A = \text{Softmax} \left( \frac{ak^T}{\sqrt{K_h}} \right)$, where $K_h = K/n$ is a scaling factor. Thus, $SA(z) = Av$,
where $v$ denotes input sequence values, and MSA output is:

$$MSA = [SA_1(z); SA_2(z); \cdots ; SA_n(z)] W_{msa},$$

where $W_{msa} \in \mathbb{R}^{n.K_h \times K}$ represents the multi-headed trainable parameter weights.

**Self Supervised Learning:** A pre-trained transformer network such as UNETR has the advantage of being trained on medical images. We take the UNETR network and finetune it in a self supervised manner using images from the different datasets that we use for super resolution. Self supervised finetuning of ViT has attracted a fair bit of attention of late with different approaches using contrastive learning [17] and masked auto-encoding [16, 27]. We investigate both approaches and identify [27] as more stable for our task. We remove the pre-trained prediction head and attach a zero-initialized $D \times K$ feedforward layer, where $K$ is the number of downstream classes, and $D$ is the dimension of the flattened patches. We define a pre-text task to identify the primary organ in the images, which is akin to a classification problem involving $K$ classes.

### 3.3 Feature Disentanglement

In order to separate the images into structure and texture components we train an autoencoder (AE) shown in Figure 1 (b). In a classic AE the encoder $E$ and generator $G$ form a mapping between image $x$ and latent code $z$ using an image reconstruction loss

$$L_{rec}(E, G) = \mathbb{E}_{x \sim X} [||x - G(E(x))||_1]$$

To ensure that the generated image is realistic we have discriminator $D$ that calculates the adversarial loss for generator $G$ and encoder $E$ as:

$$L_{adv}(E, G, D) = \mathbb{E}_{x \sim X} [-\log(D(G(E(x))))]$$

As shown in Figure 1 (b) we divide the latent code into two components - a texture component $z_t$ and a structural component $z_s$. Then amongst similar images $X^1, X^2$ from the same dataset in a minibatch we swap the two components and enforce the constraint that the resulting images be realistic, using the ‘swapped-GAN’ loss [162]

$$L_{swap}(E, G, D) = \mathbb{E}_{x^1, x^2 \sim X, x^1 \neq x^2} [-\log(D(G(z_s^{1}, z_t^{2}))))]$$

Here $z_s^1, z_t^2$ are the first and second components of $E(x^1)$ and $E(x^2)$. The intuition is to combine the structure component of one image with the texture component of another image. The two images are not identical although they belong to the same dataset. As shown in Figure 1 (b) the shapes of $z_s$ and $z_t$ are asymmetric. $z_s$ is designed to be a tensor with spatial dimensions so it can learn the structural properties associated with spatial configurations, and $z_t$ is vector that encodes the texture information. At each training iteration we randomly sample two images $x^1$ and $x^2$, and enforce $L_{rec}, L_{adv}$ for $x^1$, while
applying $L_{\text{swap}}$ to the combination of $x^1$ and $x^2$. The final loss function for feature disentanglement is given in Eqn. 6 and more details are given in Appendix A:

$$L_{\text{Disent}} = L_{\text{Rec}} + 0.7L_{\text{Adv}} + 0.7L_{\text{swap}}$$ (6)

We first train this disentanglement autoencoder that can extract the two separate features for a given input image (high or low resolution). The structure and texture features of the HR and LR images are used to train the super resolution network.

Since the HR and LR images are different versions of the same image swapping the structure code $z_s^{LR}$ (or texture $z_t^{LR}$) of the LR image with that of the HR image $z_s^{HR}$ (or $z_t^{HR}$) should still generate an image that is close to the original. Patches of size $n \times n$ are extracted around the center of the LR image and corresponding patches of size $mn \times mn$ are extracted from the center of the HR image, $m$ being the upscaling factor. This ensures that the two patches show the same region of interest. Swapping $z_i^{LR}$ with $z_i^{HR}$ and combining with $z_s^{LR}$ and $z_s^{HR}$ combine to give a fairly similar representation of the higher resolution image.

**Training The Super Resolution Network:** We use two pre-trained networks - the ViT and the feature disentanglement network. Given the LR image $x$ and the intermediate generated HR image $y$, we obtain their respective disentangled latent feature representations as $z_s^x, z_t^x$ and $z_s^y, z_t^y$. Thereafter we calculate the semantic similarity between them using the cosine similarity loss as

$$L_{\text{str}} = 1 - \langle z_s^x, z_s^y \rangle$$

$$L_{\text{tex}} = 1 - \langle z_t^x, z_t^y \rangle.$$ (7)

where $\langle, \rangle$ denotes cosine similarity. Additionally we also obtain the ViT based feature vectors of the HR ($f_{\text{VIT}}^{HR}$) and LR ($f_{\text{VIT}}^{LR}$) images from the ViT described previously and calculate their corresponding cosine similarity loss as

$$L_{\text{VIT}} = 1 - \langle f_{\text{VIT}}^{LR}, f_{\text{VIT}}^{HR} \rangle.$$ (8)

Once the above loss terms are obtained we train the whole super resolution network in an end to end manner using the following loss function. Thus the final loss function is

$$L_{\text{SR}}(X, Y) = L_{\text{adv}} + \lambda_1 L_{\text{VIT}}(X, Y) + \lambda_2 L_{\text{str}}(X, Y) + \lambda_3 L_{\text{tex}}(X, Y).$$ (9)

**4 Experiments And Results**

**Dataset Description:** We use two datasets for our experiments: 1) **fastMRI** [190] - following [189], we filter out 227 and 24 pairs of proton density (PD) and fat suppressed proton density weighted images (FS-PDWI) volumes for training and validation. 2) The **IXI dataset**: Three types of MR images are included in the datasets (i.e., PD, T1, and T2). Each of them has 500, 70, and 6 MR

[1] http://brain-development.org/ixi-dataset/
volumes for training, testing, and validation respectively. Subvolumes of size 240 × 240 × 96 are used and due to using 2D images we get 500 × 96 = 48,000 training samples.

### Table 1. Quantitative Results for IXI Dataset.

Higher values of PSNR and SSIM, and lower value of NMSE indicate better results.

| Method         | PSNR | SSIM | NMSE |
|----------------|------|------|------|
| Bicubic        | 30.4 | 0.9531 | 0.042 |
| 2D             | 31.7 | 0.982 | 0.035 |
| 2D             | 38.96 | 0.9836 | 0.022 |
| 2D             | 40.31 | 0.987 | 0.021 |
| 2D             | 41.28 | 0.9885 | 0.02 |
| 2D             | 41.66 | 0.9902 | 0.019 |
| 2D             | 42.9 | 0.9936 | 0.018 |
| Proposed       | 44.3 | 0.9972 | 0.016 |

#### Ablation Studies

| Method         | PSNR | SSIM | NMSE |
|----------------|------|------|------|
| $L_{tex} + L_{VIT}$ | 41.1 | 0.9826 | 0.021 |
| $L_{str} + L_{VIT}$ | 43.1 | 0.9902 | 0.018 |
| $L_{VIT}$        | 36.9 | 0.9745 | 0.027 |
| $L_{tex} + L_{str}$ | 37.6 | 0.9815 | 0.026 |

#### 4.1 Implementation Details

**ViT Parameters:** For self supervised finetuning we use a batch size of 6 and cross entropy loss, the AdamW optimizer [58] with initial learning rate of 0.0001 for 20,000 iterations. For the specified batch size, the average training time was 10 hours for 20,000 iterations. **AE Network:** The encoder consists of 4 convolution blocks followed by max pooling after each step. The decoder is also symmetrically designed. 3 × 3 convolution filters are used and 64, 32, 16 filters are used in each conv layer. The input to the AE is 256 × 256 and dimension of $z_{tex}$ is 256, while $z_{str}$ is 64 × 64.

**Super Resolution Network:** We train our model using Adam [48] with $\beta_1 = 0.9, \beta_2 = 0.999$, a batch size of 256 and a weight decay of 0.1, for 100 epochs. We implement all models in PyTorch and train them using one NVIDIA Tesla V100 GPU with 32GB of memory. $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = 0.9$ (from Eqn.9).

#### 4.2 Quantitative Results

For a given upscaling factor we first downsample the original image by that factor and recover the original size using different super resolution methods, and compare the performance using different metrics such as peak signal to noise ratio (PSNR), Structural Similarity Index Metric (SSIM), and Normalized Mean Square Error (NMSE). Tables [1]2 show the average values of different methods for the IXI and fastMRI datasets at upscaling factors of 2× and 4×. Our method
shows the best performance for both datasets and beats the next best method by a significant margin. While there is an expected noticeable performance drop for higher scaling factors, our method still outperforms other methods significantly. Our proposed method's advantage is the combination of CNN and ViT features that improve the image quality significantly. Although image quality degrades at higher magnification factor, our method performs better than others due to its ability to leverage local and global information.

**Ablation Studies:** Tables 1, 2 also show ablation study outcomes where different loss terms are excluded during training. Excluding the ViT features results in reduced performance. However, it is still better than most other methods because of using feature disentanglement that leads to better super resolution based on texture and structure features. On the other hand, excluding only one or more of structure and texture features leads to poor performance despite including ViT features. Thus, we conclude that both global and local information is important for accurate super resolution.

|                | IXI-T2 Images | Fast MRI      |
|----------------|---------------|---------------|
|                | 2x            | 4x            | 2x            | 4x            |
| PSNR/SSIM/NMSE |               |               |               |
| jpeg           | 30.2/0.891/0.034 | 28.1/0.878/0.034 | 26.66/0.812/0.036 | 18.367/0.708/0.082 |
| [56]           | 37.3/0.978/0.027 | 29.4/0.965/0.032 | 28.17/0.997/0.051 | 21.81/0.976/0.067 |
| [193]          | 28.5/0.943/0.029 | 29.33/0.951/0.030 | 28.870/0.670/0.048 | 23.255/0.907/0.062 |
| 32.6/0.941/0.031 |               | 29.484/0.988/0.049 | 28.219/0.974/0.059 |
| 32.6/0.941/0.029 |               | 31.705/0.979/0.045 | 29.819/0.961/0.054 |
| 32.6/0.941/0.027 |               | 34.2/0.955/0.047 |               |
| Proposed       | 41.7/0.995/0.017 |               | 34.6/0.997/0.024 |               |
|                | 36.7/0.997/0.024 |               | 34.6/0.997/0.024 |               |

**Table 2.** Quantitative Results for IXI and fastMRI dataset super resolution output. Higher values of PSNR and SSIM, and lower value of NMSE indicate better results.

### 4.3 Qualitative results:

In Figure 2, we show visualization results where the recovered images and their corresponding difference image with the original image is shown. Our method shows a very accurate reconstruction with minimal regions in the error map, while the recovered images from other methods are blurred and of poor quality. These results demonstrate the effectiveness of our approach.

### 5 Conclusion

We proposed a novel method for MR image super resolution by combining CNNs and Vision transformers. ViTs provide more global context features while CNNs
provide discriminative local information. We achieve feature disentanglement using swapped auto encoders to obtain texture and structure features. We enforce constraints that the original and super resolved images should have similar semantic information by minimizing the cosine loss of the respective structure and texture features, as well as minimizing the difference between the respective ViT features. Experimental results show our method outperforms state of the art techniques on benchmark public datasets, and ablation studies demonstrate the importance of our proposed loss terms.

A Feature Disentanglement

Similar to a classic autoencoder, the encoder $E$ produces a latent code $z \sim Z$ for image $x \sim X$. The $G$ reconstructs the original image from $z$ using an image reconstruction loss that is defined as:

$$
\mathcal{L}_{Rec}(E, G) = \mathbb{E}_{x \sim X} \left[ \| x - G(E(x)) \| \right]
$$

Additionally, the generated image should be realistic as determined by the Discriminator $D$ and is enforced using the adversarial loss defined as:

$$
\mathcal{L}_{Adv}(E, G, D) = \mathbb{E}_{x \sim X} \left[ -\log(D(G(E(x)))) \right]
$$

Furthermore, as part of our objective to achieve feature disentanglement we decompose the latent code $z$ into two components $[z_{str}, z_{tex}]$ corresponding to the structure and texture components. We enforce that swapping these components of the latent code with those from other images still produces realistic images.
This is achieved by using a modified version of the adversarial loss, which we term as the swapped GAN loss, and is defined as:

$$L_{\text{swap}}(E, G, D) = \mathbb{E}_{x^1, x^2 \sim X, x^1 \neq x^2} \left[ -\log(D(G(z^1_{\text{tex}}, z^2_{\text{str}}))) \right]$$  \hspace{1cm} (12)

Here, $z^1_{\text{tex}}, z^2_{\text{str}}$ are the first and second components of images $X^1, X^2$'s latent representations, and $X^1, X^2$ from the same dataset in a minibatch. The component $z_{\text{str}}$ is a tensor with spatial dimensions, while $z_{\text{tex}}$ is a vector that encode structure and texture information.

$L_{\text{Rec}}$ and $L_{\text{Adv}}$, are applied to image $X^1$ while $L_{\text{swap}}$ is applied to the latent components from $X^1, X^2$. The final loss function for feature disentanglement is defined as

$$L_{\text{Disent}} = L_{\text{Rec}} + 0.7 L_{\text{Adv}} + 0.7 L_{\text{swap}}$$  \hspace{1cm} (13)

### B UNETR Architecture

We use the ViT from UNETR [34] pre-trained for MR image segmentation and describe its architecture below. UNETR uses the contracting-expanding pattern consisting of a stack of transformers as the encoder which is connected to the decoder using skip connections. A 1D sequence of 3D input volume $x \in \mathbb{R}^{H \times W \times D \times C}$ with image dimension $(H, W, D)$ and $C$ input channels is created by dividing it into flattened uniform non-overlapping patches $x_{v} \in \mathbb{R}^{N \times (P^3 \cdot C)}$ where $P \times P \times P$ denotes the resolution of each patch and $N = (H \times W \times D) / P^3$ is the length of the sequence.

A linear layer projects the patches onto a $K$ dimensional embedding space which remains constant throughout the transformer layers. To preserve spatial information a 1D learnable positional embedding $E_{\text{pos}} \in \mathbb{R}^{N \times K}$ is added to the projected patch embedding $E \in \mathbb{R}^{P^3 \cdot C \times K}$ according to

$$z_0 = [x^1_{v}E; x^2_{v}E; \cdots x^N_{v}E] + E_{\text{pos}}$$  \hspace{1cm} (14)

Then multiple transformer blocks [27] are used that have multi-head self-attention (MSA) and multilayer perceptron (MLP) sublayers according to

$$z'_i = \text{MSA}(\text{Norm}(z_{i-1})) + z_{i-1}, i = 1 \cdots L$$  \hspace{1cm} (15)$$

$$z_i = \text{MLP}(\text{Norm}(z'_i)) + z'_i, i = 1 \cdots L$$  \hspace{1cm} (16)

where $\text{Norm}()$ denotes layer normalization [3]. MLP has two linear layers with GELU activation functions, $i$ denotes intermediate block and $L$ denotes transformer layers. A MSA sublayer comprises of $n$ parallel self-attention (SA) heads. Specifically, the SA block, is a parameterized function that maps a query ($q$) and the corresponding key ($k$) and value ($v$) representations in a sequence $z \in \mathbb{R}^{N \times K}$. Attention weights ($A$) measure similarity between elements in $z$ and their key-value pairs according to

$$A = \text{Softmax} \left( \frac{qk^T}{\sqrt{K_h}} \right).$$  \hspace{1cm} (17)
\( K_h = K/n \) is a scaling factor for maintaining the number of parameters to a constant value with different values of the key \( k \). Using the computed attention weights, the output of SA for values \( v \) in the sequence \( z \) is computed as
\[
SA(z) = Av,
\]
\( v \) denotes input sequence values, and MSA output is:
\[
MSA = [SA_1(z); SA_2(z); \cdots; SA_n(z)] W_{msa},
\]
where \( W_{msa} \in \mathbb{R}^{n.K_h \times K} \) represents the multi-headed trainable parameter weights.

At the encoder bottleneck (i.e. output of transformer’s last layer), a deconvolutional layer is applied to the transformed feature map to increase its resolution by a factor of 2. The resized feature map is concatenated with the feature map of the previous transformer output and fed into consecutive \( 3 \times 3 \) convolutional layers, whose output is upsampled using a deconvolutional layer. This process is repeated for all the other subsequent layers up to the original input resolution where the final output is fed into a \( 1 \times 1 \) convolutional layer with a softmax activation function to generate voxel-wise semantic predictions.

### B.1 Loss Function

The loss function is a combination of soft dice loss and cross-entropy loss, and it can be computed in a voxel-wise manner according to
\[
\mathcal{L}(G, Y) = 1 - \frac{2}{I} \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{G_{i,j} Y_{i,j}}{G_{i,j} + Y_{i,j}^2} - \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} G_{i,j} \log Y_{i,j}
\]
where \( I \) is the number of voxels; \( J \) is the number of classes; \( Y_{i,j} \) and \( G_{i,j} \) denote the probability output and one-hot encoded ground truth for class \( j \) at voxel \( i \), respectively. For a detailed explanation of all terms we urge the reader to refer to [34].

The UNETR was implemented by the authors in PyTorch and MONAI and trained using a NVIDIA DGX-1 server. All models were trained with the batch size of 6, using the AdamW optimizer [58] with initial learning rate of 0.0001 for 20,000 iterations. For the specified batch size, the average training time was 10 hours for 20,000 iterations. The transformer-based encoder follows the ViT-B16 [27] architecture with \( L=12 \) layers, an embedding size of \( K=768 \). The patch resolution was \( 16 \times 16 \times 16 \). For inference a sliding window was used with an overlap portion of 0.5 between the neighboring patches. The authors did not use any pre-trained weights for the transformer backbone (e.g. ViT on ImageNet) since it did not demonstrate any performance improvements for the medical images.

### C Additional Visual Results

In this section we show additional visual results (Figures 3, 4, 5) from the IXI and Fast MRI dataset at different super resolution factors for the different ablation
settings. The figures show the original image and the reconstructed image along with the difference image. They clearly illustrate the important contribution of each of the loss terms, and the adverse impact on super resolution if we exclude different terms.

### D Computation Time

The original UNETR model has 92.58 Million parameters, and our finetuned model has similar number of parameters at 93.4 Million. The training time on a NVIDIA Tesla V100 GPU was 10 hours for 20,000 iterations for the finetuning stage. The feature disentanglement network took 18 hours to train for 100 epochs. For the actual super resolution step, it took us 14 hours to train for 80 epochs. Note that feature disentanglement and ViT finetuning were pre-trained and while training the super resolution network we only extracted features from them.

The original UNETR model’s inference time was 12.08s. Feature extraction from the finetuned UNETR model took 1.3s, while the disentangled feature extraction took 0.05 seconds per image. For the actual super resolution at inference stage it took 1.2 seconds for 2x upsampling for a 512 × 512 image.

### E Architecture of Super Resolution Network

Figure 6 shows the detailed architecture of the super resolution network’s generator and discriminator components. In the generator (Figure 6 (a)) the input low resolution image $I_{LR}$ is passed through a convolution block followed by ReLU activation. The output is passed through a residual block with skip connections. Each block has convolutional layers with 3 × 3 filters and 64 feature maps, followed by batch normalization and ReLU activation. This output is subsequently passed through multiple residual blocks. Their output is passed through a series of upsampling stages, where each stage doubles the input image size. The output is passed through a convolution stage to get the super resolved image $I_{SR}$. Depending upon the desired scaling, the number of upsampling stages can be changed. The discriminator outputs the $L_{adv}$ in Eqn 9 and is defined as:

$$L_{adv,SR}(E,G,D) = \mathbb{E}_{lr \sim LR}[-\log(D(G(E(lr)))))$$

where $LR$ is the set of low resolution images and $G(E(lr))$ is the super resolved high resolution image. The other two loss terms, $L_{V,iT}, L_{tex}, L_{str}$, have been defined before.

### F Loss Plots

In figure 7 we show the loss plots for training, validation and test data splits on the IXI brain image dataset. We see that the training error decreases gradually, which is also observable for the validation and test errors, although their magnitudes are higher than the training error. The plots show that there is minimal chance of overfitting of the models and the results are not biased.
G Comparison With \cite{30}

The results of \cite{30} come up as worse than bicubic interpolation on the IXI dataset. This is surprising considering that they use a task transformer network. In our re-implementation we report better results than those reported on the paper \cite{30} since we devote significant bit of time in finetuning the parameters. Our experiments show that by removing the task transformer component the performance degrades but is still better than the numbers in \cite{30}. While it is difficult to ascertain the reason behind their low performance, a possible reason could be the architecture of the task transformer network. This requires further investigation and is beyond the scope of our current work.

References

1. Antony, B., Sedai, S., Mahapatra, D., Garnavi, R.: Real-time passive monitoring and assessment of pediatric eye health. In: US Patent App. 16/178,757 (2020)
2. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention (2015)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016)
4. Bastide, P., Kiral-Kornek, I., Mahapatra, D., Saha, S., Vishwanath, A., Cavallar, S.V.: Machine learned optimizing of health activity for participants during meeting times. In: US Patent App. 15/426,634 (2018)
5. Bastide, P., Kiral-Kornek, I., Mahapatra, D., Saha, S., Vishwanath, A., Cavallar, S.V.: Visual health maintenance and improvement. In: US Patent 9,993,385 (2018)
6. Bastide, P., Kiral-Kornek, I., Mahapatra, D., Saha, S., Vishwanath, A., Cavallar, S.V.: Crowdsourcing health improvements routes. In: US Patent App. 15/611,519 (2019)
7. Bozorgtabar, B., Mahapatra, D., von Teng, H., Pollinger, A., Ebner, L., Thiran, J.P., Reyes, M.: Informative sample generation using class aware generative adversarial networks for classification of chest x-rays. Computer Vision and Image Understanding \textbf{184}, 57–65 (2019)
8. Bozorgtabar, B., Mahapatra, D., von Teng, H., Pollinger, A., Ebner, L., Thiran, J.P., Reyes, M.: Informative sample generation using class aware generative adversarial networks for classification of chest x-rays. In: arXiv preprint arXiv:1904.10781 (2019)
9. Bozorgtabar, B., Mahapatra, D., Thiran, J.P.: Exprada: Adversarial domain adaptation for facial expression analysis. In Press Pattern Recognition \textbf{100}, 15–28 (2020)
10. Bozorgtabar, B., Mahapatra, D., Thiran, J.P., Shao, L.: SALAD: Self-supervised aggregation learning for anomaly detection on x-rays. In: In Proc. MICCAI, pp. 468–478 (2020)
11. Bozorgtabar, B., Mahapatra, D., Vray, G., Thiran, J.P.: Anomaly detection on x-rays using self-supervised aggregation learning. In: arXiv preprint arXiv:2010.09856 (2020)
12. Bozorgtabar, B., Mahapatra, D., Zlobec, I., Rau, T., Thiran, J.: Computational pathology. Frontiers in Medicine \textbf{7} (2020)
13. Bozorgtabar, B., Rad, M.S., Mahapatra, D., Thiran, J.P.: Syndemo: Synergistic deep feature alignment for joint learning of depth and ego-motion. In: In Proc. IEEE ICCV (2019)
14. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 213–229. Springer International Publishing, Cham (2020)

15. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation (2021)

16. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1691–1703. PMLR (13–18 Jul 2020), https://proceedings.mlr.press/v119/chen20s.html

17. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers (2021)

18. Chen, Y., Shi, F., Christodoulopou, A.G., Xie, Y., Zhou, Z., Li, D.: Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network. In: MICCAI pp. 91–99 (2018)

19. Chen, Y., Xie, Y., Zhou, Z., Shi, F., Christodoulopou, A.G., Li, D.: Brain MRI super resolution using 3d deep densely connected neural networks. CoRR abs/1801.02728 (2018), http://arxiv.org/abs/1801.02728

20. Dar, S.U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., Çukur, T.: Image synthesis in multi-contrast mri with conditional generative adversarial networks. IEEE Transactions on Medical Imaging 38(10), 2375–2388 (2019). [https://doi.org/10.1109/TMI.2019.2901750]

21. Das, S.D., Dutta, S., Shah, N.A., Mahapatra, D., Ge, Z.: Anomaly detection in retinal images using multi-scale deep feature sparse coding. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2022). [https://doi.org/10.1109/ISBI52829.2022.9761713]

22. Das, S.D., Dutta, S., Shah, N.A., Mahapatra, D., Ge, Z.: Anomaly detection in retinal images using multi-scale deep feature sparse coding. In: arXiv preprint arXiv:2202.09988 (2022)

23. Devika, K., Mahapatra, D., Subramanian, R., Oruganti, V.R.M.: Outlier-based autism detection using longitudinal structural mri. IEEE Access 10, 27794–27808 (2022). [https://doi.org/10.1109/ACCESS.2022.3157613]

24. Devika, K., Mahapatra, D., Subramanian, R., Oruganti, V.R.M.: Outlier-based autism detection using longitudinal structural mri. In: arXiv preprint arXiv:2202.09988 (2022)

25. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV. pp. 184–199 (2014)

26. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 295–307 (2016)

27. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR abs/2011.11929 (2020). https://arxiv.org/abs/2011.11929

28. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis (2020)

29. Feng, C.M., Fu, H., Yuan, S., Xu, Y.: Multi-contrast mri super-resolution via a multi-stage integration network. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy,
uncertainty estimation. IEEE Transactions on Medical Imaging pp. 1–1 (2022). https://doi.org/10.1109/TMI.2022.3141425
45. Ju, L., Wang, X., Zhao, X., Lu, H., Mahapatra, D., Bonnington, P., Ge, Z.: Synergistic adversarial label learning for grading retinal diseases via knowledge distillation and multi-task learning. IEEE JBHI 100, 1–14 (2020)
46. Ju, L., Wang, X., Zhao, X., Lu, H., Mahapatra, D., Ge, Z.: Relational subsets knowledge distillation for long-tailed retinal diseases recognition. In: In MICCAI 2021. pp. 1–11 (2021)
47. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral) (June 2016)
48. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: arXiv preprint arXiv:1412.6980, (2014)
49. Kuanar, S., Athitsos, V., Mahapatra, D., Rajan, A.: Multi-scale deep learning architecture for nucleus detection in renal cell carcinoma microscopy image. In: arXiv preprint arXiv:2104.13557 (2021)
50. Kuanar, S., Athitsos, V., Mahapatra, D., Rao, K., Akhtar, Z., Dasgupta, D.: Low dose abdominal ct image reconstruction: An unsupervised learning based approach. In: In Proc. IEEE ICIP. pp. 1351–1355 (2019)
51. Kuanar, S., Mahapatra, D., Athitsos, V., Rao, K.: Gated fusion network for sao filter and inter frame prediction in versatile video coding. In: arXiv preprint arXiv:2105.12229 (2021)
52. Kuanar, S., Rao, K., Mahapatra, D., Bilas, M.: Night time haze and glow removal using deep dilated convolutional network. In: arXiv preprint arXiv:1902.00855 (2019)
53. Kuanar, S., Mahapatra, D., Bilas, M., Rao, K.: Multi-path dilated convolution network for haze and glow removal in night time images. The Visual Computer 38(3), 1121–1134 (2022)
54. Kuang, H., Guthier, B., Saini, M., Mahapatra, D., Saddik, A.E.: A real-time smart assistant for video surveillance through handheld devices. In: In Proc: ACM Intl. Conf. Multimedia. pp. 917–920 (2014)
55. Li, Z., Mahapatra, D., J.Tielbeek, Stoker, J., van Vliet, L., Vos, F.: Image registration based on autocorrelation of local structure. IEEE Trans. Med. Imaging 35(1), 63–75 (2016)
56. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution (2017)
57. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021)
58. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019)
59. Lyu, Q., Shan, H., Steber, C., Helis, C., Whitlow, C., Chan, M., Wang, G.: Multi-contrast super-resolution mri through a progressive network. IEEE Transactions on Medical Imaging 39(9), 2738–2749 (2020). https://doi.org/10.1109/TMI.2020.2974858
60. Mahapatra, D.: Elastic registration of cardiac perfusion images using saliency information. Sequence and Genome Analysis – Methods and Applications pp. 351–364 (2011)
61. Mahapatra, D.: Neonatal brain mri skull stripping using graph cuts and shape priors. In: In Proc: MICCAI workshop on Image Analysis of Human Brain Development (IAHBD) (2011)
62. Mahapatra, D.: Cardiac lv and rv segmentation using mutual context information. In: Proc. MICCAI-MLMI. pp. 201–209 (2012)
63. Mahapatra, D.: Groupwise registration of dynamic cardiac perfusion images using temporal information and segmentation information. In: In Proc. SPIE Medical Imaging (2012)

64. Mahapatra, D.: Landmark detection in cardiac MRI using learned local image statistics. In: Proc. MICCAI-Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges (STACOM). pp. 115–124 (2012)

65. Mahapatra, D.: Skull stripping of neonatal brain MRI: Using prior shape information with graph cuts. J. Digit. Imaging 25(6), 802–814 (2012)

66. Mahapatra, D.: Cardiac image segmentation from cine cardiac MRI using graph cuts and shape priors. J. Digit. Imaging 26(4), 721–730 (2013)

67. Mahapatra, D.: Cardiac MRI segmentation using mutual context information from left and right ventricle. J. Digit. Imaging 26(5), 802–814 (2013)

68. Mahapatra, D.: Graph cut based automatic prostate segmentation using learned semantic information. In: Proc. IEEE ISBI. pp. 1304–1307 (2013)

69. Mahapatra, D.: Joint segmentation and groupwise registration of cardiac perfusion images using temporal information. J. Digit. Imaging 26(2), 173–182 (2013)

70. Mahapatra, D.: An automated approach to cardiac RV segmentation from MRI using learned semantic information and graph cuts. J. Digit. Imaging. 27(6), 794–804 (2014)

71. Mahapatra, D.: Combining multiple expert annotations using semi-supervised learning and graph cuts for medical image segmentation. Computer Vision and Image Understanding 151(1), 114–123 (2016)

72. Mahapatra, D.: Retinal image quality classification using neurobiological models of the human visual system. In: In Proc. MICCAI-OMIA. pp. 1–8 (2016)

73. Mahapatra, D.: Consensus based medical image segmentation using semi-supervised learning and graph cuts. In: arXiv preprint arXiv:1612.02166 (2017)

74. Mahapatra, D.: Semi-supervised learning and graph cuts for consensus based medical image segmentation. Pattern Recognition 63(1), 700–709 (2017)

75. Mahapatra, D.: AMD severity prediction and explainability using image registration and deep embedded clustering. In: arXiv preprint arXiv:1907.03075 (2019)

76. Mahapatra, D.: Generative adversarial networks and domain adaptation for training data independent image registration. In: arXiv preprint arXiv:1910.08593 (2019)

77. Mahapatra, D.: Registration of histopathology images using structural information from fine grained feature maps. In: arXiv preprint arXiv:2007.02078 (2020)

78. Mahapatra, D.: Interpretability-driven sample selection using self-supervised learning for disease classification and segmentation. In: arXiv preprint arXiv:2104.06087 (2021)

79. Mahapatra, D.: Learning of inter-label geometric relationships using self-supervised learning: Application to Gleason grade segmentation. In: arXiv preprint arXiv:2110.00404 (2021)

80. Mahapatra, D., Agarwal, K., Khosrowabadi, R., Prasad, D.: Recent advances in statistical data and signal analysis: Application to real world diagnostics from medical and biological signals. In: Computational and mathematical methods in medicine (2016)

81. Mahapatra, D., Antony, B., Sedai, S., Garnavi, R.: Deformable medical image registration using generative adversarial networks. In: In Proc. IEEE ISBI. pp. 1449–1453 (2018)

82. Mahapatra, D., Bozorgtabar, B.: Retinal vasculature segmentation using local saliency maps and generative adversarial networks for image super resolution. In: arXiv preprint arXiv:1710.04783 (2017)
83. Mahapatra, D., Bozorgtabar, B.: Progressive generative adversarial networks for medical image super resolution. In: arXiv preprint arXiv:1902.02144 (2019)
84. Mahapatra, D., Bozorgtabar, B., Garnavi, R.: Image super-resolution using progressive generative adversarial networks for medical image analysis. Computerized Medical Imaging and Graphics 71, 30–39 (2019)
85. Mahapatra, D., Bozorgtabar, B., Ge, Z.: Medical image classification using generalized zero shot learning. In: In IEEE CVAMD 2021. pp. 3344–3353 (2021)
86. Mahapatra, D., Bozorgtabar, B., Kuanar, S., Ge, Z.: Self-supervised multimodal generalized zero shot learning for gleason grading. In: In MICCAI-DART 2021. pp. 1–11 (2021)
87. Mahapatra, D., Bozorgtabar, B., Shao, L.: Pathological retinal region segmentation from OCT images using geometric relation based augmentation. In: In Proc. IEEE CVPR. pp. 9611–9620 (2020)
88. Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Shao, L.: Pathological retinal region segmentation from OCT images using geometric relation based augmentation. In: arXiv preprint arXiv:2003.14119 (2020)
89. Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Shao, L.: Structure preserving stain normalization of histopathology images using self supervised semantic guidance. In: In Proc. MICCAI. pp. 309–319 (2020)
90. Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Shao, L.: Structure preserving stain normalization of histopathology images using self supervised semantic guidance. In: arXiv preprint arXiv:2008.02101 (2020)
91. Mahapatra, D., Bozorgtabar, S., Hewavitahranage, S., Garnavi, R.: Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis. In: In Proc. MICCAI. pp. 382–390 (2017)
92. Mahapatra, D., Bozorgtabar, S., Thiran, J.P., Reyes, M.: Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In: In Proc. MICCAI (2). pp. 580–588 (2018)
93. Mahapatra, D., Buhmann, J.: Obtaining consensus annotations for retinal image segmentation using random forest and graph cuts. In: In Proc. OMIA. pp. 41–48 (2015)
94. Mahapatra, D., Buhmann, J.: Visual saliency based active learning for prostate MRI segmentation. In: In Proc. MLMI. pp. 9–16 (2015)
95. Mahapatra, D., Buhmann, J.: Visual saliency-based active learning for prostate magnetic resonance imaging segmentation. SPIE Journal of Medical Imaging 3(1), 014003 (2016)
96. Mahapatra, D., Buhmann, J.: Automatic cardiac RV segmentation using semantic information with graph cuts. In: Proc. IEEE ISBI. pp. 1094–1097 (2013)
97. Mahapatra, D., Buhmann, J.: Analyzing training information from random forests for improved image segmentation. IEEE Trans. Imag. Proc. 23(4), 1504–1512 (2014)
98. Mahapatra, D., Buhmann, J.: Prostate MRI segmentation using learned semantic knowledge and graph cuts. IEEE Trans. Biomed. Engg. 61(3), 756–764 (2014)
99. Mahapatra, D., Buhmann, J.: A field of experts model for optic cup and disc segmentation from retinal fundus images. In: In Proc. IEEE ISBI. pp. 218–221 (2015)
100. Mahapatra, D., Garnavi, R., Roy, P., Tennakoon, R.: System and method to teach and evaluate image grading performance using prior learned expert knowledge base. In: US Patent App. 15/459,457 (2018)
101. Mahapatra, D., Garnavi, R., Roy, P., Tennakoon, R.: System and method to teach and evaluate image grading performance using prior learned expert knowledge base. In: US Patent App. 15/814,590 (2018)

102. Mahapatra, D., Garnavi, R., Sedai, S., Roy, P.: Joint segmentation and characteristics estimation in medical images. In: US Patent App. 15/234,426 (2017)

103. Mahapatra, D., Garnavi, R., Sedai, S., Roy, P.: Retinal image quality assessment, error identification and automatic quality correction. In: US Patent 9,779,492 (2017)

104. Mahapatra, D., Garnavi, R., Sedai, S., Tennakoon, R.: Classification of severity of pathological condition using hybrid image representation. In: US Patent App. 15/426,634 (2018)

105. Mahapatra, D., Garnavi, R., Sedai, S., Tennakoon, R.: Generating an enriched knowledge base from annotated images. In: US Patent App. 15/429,735 (2018)

106. Mahapatra, D., Garnavi, R., Sedai, S., Tennakoon, R., Chakravorty, R.: Early prediction of age related macular degeneration by image reconstruction. In: US Patent App. 15/854,984 (2018)

107. Mahapatra, D., Garnavi, R., Sedai, S., Tennakoon, R., Chakravorty, R.: Early prediction of age related macular degeneration by image reconstruction. In: US Patent 9,943,225 (2018)

108. Mahapatra, D., Ge, Z.: Combining transfer learning and segmentation information with gans for training data independent image registration. In: arXiv preprint arXiv:1903.10139 (2019)

109. Mahapatra, D., Ge, Z.: Training data independent image registration with gans using transfer learning and segmentation information. In: In Proc. IEEE ISBI. pp. 709–713 (2019)

110. Mahapatra, D., Ge, Z.: Training data independent image registration using generative adversarial networks and domain adaptation. Pattern Recognition 100, 1–14 (2020)

111. Mahapatra, D., Ge, Z., Sedai, S.: Joint registration and segmentation of images using deep learning. In: US Patent App. 16/001,566 (2019)

112. Mahapatra, D., Ge, Z., Sedai, S., Chakravorty, R.: Joint registration and segmentation of xray images using generative adversarial networks. In: In Proc: MICCAI-MLMI. pp. 73–80 (2018)

113. Mahapatra, D., Gilani, S., Saini, M.: Coherency based spatio-temporal saliency detection for video object segmentation. IEEE Journal of Selected Topics in Signal Processing. 8(3), 454–462 (2014)

114. Mahapatra, D., J.Tielbeek, Makanyanga, J., Stoker, J., Taylor, S., Vos, F., Buhmann, J.: Automatic detection and segmentation of crohn’s disease tissues from abdominal mri. IEEE Trans. Med. Imaging 32(12), 1232–1248 (2013)

115. Mahapatra, D., J.Tielbeek, Makanyanga, J., Stoker, J., Taylor, S., Vos, F., Buhmann, J.: Active learning based segmentation of crohn’s disease using principles of visual saliency. In: Proc. IEEE ISBI. pp. 226–229 (2014)

116. Mahapatra, D., J.Tielbeek, Makanyanga, J., Stoker, J., Taylor, S., Vos, F., Buhmann, J.: Combining multiple expert annotations using semi-supervised learning and graph cuts for crohn’s disease segmentation. In: In Proc: MICCAI-ABD (2014)

117. Mahapatra, D., J.Tielbeek, Vos, F., Buhmann, J.: A supervised learning approach for crohn’s disease detection using higher order image statistics and a novel shape asymmetry measure. J. Digit. Imaging 26(5), 920–931 (2013)
118. Mahapatra, D., Kuanar, S., Bozorgtabar, B., Ge, Z.: Self-supervised learning of inter-label geometric relationships for gleason grade segmentation. In: In MICCAI-DART 2021. pp. 57–67 (2021)
119. Mahapatra, D., Li, Z., Vos, F., Buhmann, J.: Joint segmentation and groupwise registration of cardiac dce mri using sparse data representations. In: In Proc. IEEE ISBI. pp. 1312–1315 (2015)
120. Mahapatra, D., Routray, A., Mishra, C.: An active snake model for classification of extreme emotions. In: IEEE International Conference on Industrial Technology (ICIT). pp. 2195–2199 (2006)
121. Mahapatra, D., Roy, P., Sedai, S., Garnavi, R.: A cnn based neurobiology inspired approach for retinal image quality assessment. In: In Proc. EMBC. pp. 1304–1307 (2016)
122. Mahapatra, D., Roy, P., Sedai, S., Garnavi, R.: Retinal image quality classification using saliency maps and cnns. In: In Proc. MICCAI-MLMI. pp. 172–179 (2016)
123. Mahapatra, D., Roy, S., Sun, Y.: Retrieval of mr kidney images by incorporating shape information in histogram of low level features. In: In 13th International Conference on Biomedical Engineering. pp. 661–664 (2009)
124. Mahapatra, D., Saha, S., Vishwanath, A., Bastide, P.: Generating hyperspectral image database by machine learning and mapping of color images to hyperspectral domain. In: US Patent App. 15/949,528 (2019)
125. Mahapatra, D., Saini, M.: A particle filter framework for object tracking using visual-saliency information. Intelligent Multimedia Surveillance pp. 133–147 (2013)
126. Mahapatra, D., Saini, M., Sun, Y.: Illumination invariant tracking in office environments using neurobiology-saliency based particle filter. In: IEEE ICME. pp. 953–956 (2008)
127. Mahapatra, D., Schüßler, P., Tielbeek, J., Vos, F., Buhmann, J.: Semi-supervised and active learning for automatic segmentation of crohn’s disease. In: Proc. MICCAI, Part 2. pp. 214–221 (2013)
128. Mahapatra, D., Sedai, S., Garnavi, R.: Elastic registration of medical images with gans. In: arXiv preprint arXiv:1805.02369 (2018)
129. Mahapatra, D., Sedai, S., Halupka, K.: Uncertainty region based image enhancement. In: US Patent App. 10,832,074 (2020)
130. Mahapatra, D., Singh, A.: Ct image synthesis using weakly supervised segmentation and geometric inter-label relations for covid image analysis. In: arXiv preprint arXiv:2106.10230 (2021)
131. Mahapatra, D., Sun, Y.: Nonrigid registration of dynamic renal MR images using a saliency based MRF model. In: Proc. MICCAI. pp. 771–779 (2008)
132. Mahapatra, D., Sun, Y.: Registration of dynamic renal mr images using neurobiological model of saliency. In: Proc. ISBI. pp. 1119–1122 (2008)
133. Mahapatra, D., Sun, Y.: Using saliency features for graphcut segmentation of perfusion kidney images. In: In 13th International Conference on Biomedical Engineering (2008)
134. Mahapatra, D., Sun, Y.: Joint registration and segmentation of dynamic cardiac perfusion images using mrfs. In: Proc. MICCAI. pp. 493–501 (2010)
135. Mahapatra, D., Sun, Y.: Mrf based joint registration and segmentation of dynamic renal mr images. In: Second International Conference on Digital Image Processing, vol. 7546, pp. 285–290 (2010)
136. Mahapatra, D., Sun., Y.: An mr framework for joint registration and segmentation of natural and perfusion images. In: Proc. IEEE ICIP. pp. 1709–1712 (2010)
137. Mahapatra, D., Sun, Y.: Retrieval of perfusion images using cosegmentation and shape context information. In: Proc. APSIPA Annual Summit and Conference (ASC). vol. 35 (2010)
138. Mahapatra, D., Sun, Y.: Rigid registration of renal perfusion images using a neurobiology based visual saliency model. EURASIP Journal on Image and Video Processing. pp. 1–16 (2010)
139. Mahapatra, D., Sun, Y.: Mrf based intensity invariant elastic registration of cardiac perfusion images using saliency information. IEEE Trans. Biomed. Engg. 58(4), 991–1000 (2011)
140. Mahapatra, D., Sun, Y.: Orientation histograms as shape priors for left ventricle segmentation using graph cuts. In: In Proc: MICCAI. pp. 420–427 (2011)
141. Mahapatra, D., Sun, Y.: Integrating segmentation information for improved mr-based elastic image registration. IEEE Trans. Imag. Proc. 21(1), 170–183 (2012)
142. Mahapatra, D., Tielbeek, J., Buhmann, J., Vos, F.: A supervised learning based approach to detect crohn’s disease in abdominal mr volumes. In: Proc. MICCAI workshop Computational and Clinical Applications in Abdominal Imaging(MICCAI-ABD). pp. 97–106 (2012)
143. Mahapatra, D., Tielbeek, J., Vos, F., Buhmann, J.: Crohn’s disease tissue segmentation from abdominal mri using semantic information and graph cuts. In: Proc. IEEE ISBI. pp. 358–361 (2013)
144. Mahapatra, D., Tielbeek, J., Vos, F., Buhmann, J.: Localizing and segmenting crohn’s disease affected regions in abdominal mri using novel context features. In: Proc. SPIE Medical Imaging (2013)
145. Mahapatra, D., Tielbeek, J., Vos, F., Buhmann, J.: Weakly supervised semantic segmentation of crohn’s disease tissues from abdominal mri. In: Proc. IEEE ISBI. pp. 832–835 (2013)
146. Mahapatra, D., Vos, F., Buhmann, J.: Crohn’s disease segmentation from mri using learned image priors. In: In Proc. IEEE ISBI. pp. 625–628 (2015)
147. Mahapatra, D., Vos, F., Buhmann, J.: Active learning based segmentation of crohns disease from abdominal mri. Computer Methods and Programs in Biomedicine 128(1), 75–85 (2016)
148. Mahapatra, D., Winkler, S., Yen, S.: Motion saliency outweighs other low-level features while watching videos. In: SPIE HVEI. pp. 1–10 (2008)
149. Mahapatra, D.: Registration and segmentation methodology for perfusion mr images: Application to cardiac and renal images. - pp. – (2011)
150. Mahapatra, D.: Registration and segmentation methodology for perfusion mr images: Application to cardiac and renal images. - pp. – (2011)
151. Mahapatra, D.: Multimodal generalized zero shot learning for gleason grading using self-supervised learning. In: arXiv preprint arXiv:2111.07646 (2021)
152. Mahapatra, D.: Generalized zero shot learning for medical image classification. In: arXiv preprint arXiv:2204.01728 (2022)
153. Mahapatra, D.: Unsupervised domain adaptation using feature disentanglement and gcns for medical image classification. In: arXiv preprint arXiv:2206.13123 (2022)
154. Mahapatra, D., Ge, Z.: MR image super resolution by combining feature disentanglement CNNs and vision transformers. In: Medical Imaging with Deep Learning (2022)
155. Mahapatra, D., Ge, Z.: Mr image super resolution by combining feature disentanglement cnns and vision transformers. In: - (2022)
156. Mahapatra, D., Ge, Z., Reyes, M.: Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps. IEEE Transactions on Medical Imaging pp. 1–1 (2022). https://doi.org/10.1109/TMI.2022.3163232

157. Mahapatra, D., Korevaar, S., Tennakoon, R.: Gen based unsupervised domain adaptation with feature disentanglement for medical image classification. In: - (2022)

158. Mahapatra, D., Poellinger, A., Reyes, M.: Interpretability-guided inductive bias for deep learning based medical image classification and segmentation. Medical Image Analysis p. 102551 (2022)

159. Mahapatra, D., Poellinger, A., Shao, L., Reyes, M.: Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. IEEE TMI pp. 1–15 (2021)

160. Manjón, J.V., Coupé, P., Buades, A., Fonov, V., Collins, D.L.: Non-local mri upsampling. Medical Image Analysis 14(6), 1465–1476 (2010)

161. Pandey, A., Paliwal, B., Dhall, A., Subramanian, R., Mahapatra, D.: This explains that: Congruent image–report generation for explainable medical image analysis with cyclic generative adversarial networks. In: In MICCAI-iMIMIC 2021. pp. 1–11 (2021)

162. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A.A., Zhang, R.: Swapping autoencoder for deep image manipulation. In: Advances in Neural Information Processing Systems (2020)

163. Pham, C.H., Ducournau, A., Fablet, R., Rousseau, F.: Brain mri super-resolution using deep 3d convolutional networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). pp. 197–200 (2017). https://doi.org/10.1109/ISBI.2017.7950500

164. Plenge, E., et al: Super-resolution methods in mri: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? Magnetic resonance in medicine 68(6), 1983–1993 (2012)

165. Roy, P., Chakravorty, R., Sedai, S., Mahapatra, D., Garnavi, R.: Automatic eye type detection in retinal fundus image using fusion of transfer learning and anatomical features. In: In Proc. DIICTA. pp. 1–7 (2016)

166. Roy, P., Tennakoon, R., Cao, K., Sedai, S., Mahapatra, D., Maetschke, S., Garnavi, R.: A novel hybrid approach for severity assessment of diabetic retinopathy in colour fundus images. In: In Proc. IEEE ISBI. pp. 1078–1082 (2017)

167. Roy, P., Mahapatra, D., Garnavi, R., Tennakoon, R.: System and method to teach and evaluate image grading performance using prior learned expert knowledge base. In: US Patent App. 10.984.674 (2021)

168. Saini, M., Guthier, B., Kuang, H., Mahapatra, D., Saddik, A.: szoom: A framework for automatic zoom into high resolution surveillance videos. In: arXiv preprint arXiv:1909.10164 (2019)

169. Schüffler, P., Mahapatra, D., Tielbeck, J., Vos, F., Makanyanga, J., Pends, D., Nio, C., Stoker, J., Taylor, S., Buhmann, J.: A model development pipeline for crohns disease severity assessment from magnetic resonance images. In: In Proc: MICCAI-ABD (2013)

170. Schüffler, P., Mahapatra, D., Tielbeck, J., Vos, F., Makanyanga, J., Pends, D., Nio, C., Stoker, J., Taylor, S., Buhmann, J.: Semi automatic crohns disease severity assessment on mri imaging. In: In Proc: MICCAI-ABD (2014)

171. Scherrer, B., Gholipour, A., Warfield, S.K.: Super-resolution reconstruction to increase the spatial resolution of diffusion weighted images from orthogonal anisotropic acquisitions. Medical Image Analysis 16(7), 1465–1476 (2012)
172. Schüßler, P.J., Mahapatra, D., Vos, F.M., Buhmann, J.M.: Computer aided crohn’s disease severity assessment in mri. In: VIGOR++ Workshop 2014-Showcase of Research Outcomes and Future Outlook. pp. – (2014)

173. Sedai, S., Mahapatra, D., Antony, B., Garnavi, R.: Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using bayesian deep learning. In: In Proc. MICCAI-OMIA. pp. 219–227 (2018)

174. Sedai, S., Mahapatra, D., Ge, Z., Chakravorty, R., Garnavi, R.: Deep multiscale convolutional feature learning for weakly supervised localization of chest pathologies in x-ray images. In: In Proc. MICCAI-MLMI. pp. 267–275 (2018)

175. Sedai, S., Mahapatra, D., Hewavitharanage, S., Maetschke, S., Garnavi, R.: Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder. In: In Proc. MICCAI. pp. 75–82 (2017)

176. Sedai, S., Roy, P., Mahapatra, D., Garnavi, R.: Segmentation of optic disc and optic cup in retinal fundus images using shape regression. In: In Proc. EMBC. pp. 3260–3264 (2016)

177. Sedai, S., Roy, P., Mahapatra, D., Garnavi, R.: Segmentation of optic disc and optic cup in retinal images using coupled shape regression. In: In Proc. MICCAI-OMIA. pp. 1–8 (2016)

178. Srivastava, S., Yaqub, M., Nandakumar, K., Ge, Z., Mahapatra, D.: Continual domain incremental learning for chest x-ray classification in low-resource clinical settings. In: In MICCAI-FAIR 2021. pp. 1–11 (2021)

179. Tennakoon, R., Mahapatra, D., Roy, P., Sedai, S., Garnavi, R.: Image quality classification for dr screening using convolutional neural networks. In: In Proc. MICCAI-OMIA. pp. 113–120 (2016)

180. Tong, J., Mahapatra, D., Bonnington, P., Drummond, T., Ge, Z.: Registration of histopathology images using self supervised fine grained feature maps. In: In Proc. MICCAI-DART Workshop. pp. 41–51 (2020)

181. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation (2021)

182. Verma, R., Kumar, N., Patil, A., Kurian, N.C., Rane, S., Graham, S., Vu, Q.D., Zwager, M., Raza, S.E.A., Rajpoot, N., Wu, X., Chen, H., Huang, Y., Wang, L., Jung, H., Brown, G.T., Liu, Y., Liu, S., Jahromi, S.A.F., Khani, A.A., Montahehi, E., Baghshah, M.S., Behrouzi, H., Semkin, P., Rassadin, A., Dutandt, P., Lodaya, R., Baid, U., Baheti, B., Talbar, S., Mahbod, A., Ecker, R., Ellinger, I., Luo, Z., Dong, B., Xu, Z., Yao, Y., Lv, S., Feng, M., Xu, K., Zunair, H., Hamza, A.B., Smilea, S., Yin, T.K., Fang, Q.R., Srivastava, S., Mahapatra, D., Travasla, L., Zhang, H., Narayanan, P., Law, J., Yuan, Y., Tejomay, A., Mitkari, A., Koka, D., Ramachandra, V., Kini, L., Sethi, A.: Monusac2020: A multi-organ nuclei segmentation and classification challenge. IEEE Transactions on Medical Imaging 40(12), 3413–3423 (2021). https://doi.org/10.1109/TMI.2021.3085712

183. Vos, F.M., Tielbeek, J., Naziroglu, R., Li, Z., Schüffler, P., Mahapatra, D., Wiebel, A., Lavini, C., Buhmann, J., Hege, H., Stoker, J., van Vliet, L.: Computational modeling for assessment of IBD: to be or not to be? In: Proc. IEEE EMBC. pp. 3974–3977 (2012)

184. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 568–578 (October 2021)

185. Wang, W., Chen, C., Ding, M., Li, J., Yu, H., Zha, S.: Transbts: Multimodal brain tumor segmentation using transformer (2021)
186. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation (2021)
187. Xing, Y., Ge, Z., Zeng, R., Mahapatra, D., Seah, J., Law, M., Drummond, T.: Adversarial pulmonary pathology translation for pairwise chest x-ray data augmentation. In: In Proc. MICCAI. pp. 757–765 (2019)
188. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention (2016)
189. Xuan, K., Sun, S., Xue, Z., Wang, Q., Liao, S.: Learning mri k-space subsampling pattern using progressive weight pruning. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoczeau, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 178–187. Springer International Publishing, Cham (2020)
190. Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M.J., Defazio, A., Stern, R., Johnson, P., Bruno, M., Parente, M., Geras, K.J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzal, M., Romero, A., Rabbat, M., Vincent, P., Yakubova, N., Pinkerton, J., Wang, D., Owens, E., Zitnick, C.L., Recht, M.P., Sodickson, D.K., Lui, Y.W.: fastmri: An open dataset and benchmarks for accelerated mri (2019)
191. Zhang, Y., Li, K., Li, K., Fu, Y.: Mr image super-resolution with squeeze and excitation reasoning attention network. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13420–13429 (2021). https://doi.org/10.1109/CVPR46437.2021.01322
192. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018)
193. Zhao, C., Carass, A., Dewey, B., Woo, J., Oh, J., Calabresi, P., Reich, D., Sati, P., Pham, D., Prince, J.: A deep learning based anti-aliasing self super-resolution algorithm for mri. In: Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G., Frangi, A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018 - 21st International Conference, 2018, Proceedings. pp. 100–108. Springer Verlag (2018)
194. Zhao, X., Zhang, Y., Zhang, T., Zou, X.: Channel splitting network for single mri image super-resolution. IEEE Transactions on Image Processing 28(11), 5649–5662 (2019)
195. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers (2021)
196. Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L.: Hi-net: Hybrid-fusion network for multi-modal mri image synthesis. IEEE Transactions on Medical Imaging 39(9), 2772–2781 (2020). https://doi.org/10.1109/TMI.2020.2975344
197. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection (2021)
198. Zilly, J., Buhmann, J., Mahapatra, D.: Boosting convolutional filters with entropy sampling for optic cup and disc image segmentation from fundus images. In: In Proc. MLMI. pp. 136–143 (2015)
199. Zilly, J., Buhmann, J., Mahapatra, D.: Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. In Press Computerized Medical Imaging and Graphics 55(1), 28–41 (2017)
For IXI Image Dataset at 2x superresolution

Fig. 3. Visualization of superresolution. For each figure the top row is the original image followed by the difference image in the bottom row. Column 1 - original image; Reconstructed Image using: Column 2 - Our Proposed method; Column 3 - $L_{str} + L_{VIT}$; Column 4 - $L_{tex} + L_{VIT}$; Column 5 - $L_{tex} + L_{str}$; Column 6 - $L_{VIT}$. 
Fig. 4. Visualization of superresolution. For each figure the top row is the original image followed by the difference image in the bottom row. Column 1- original image; Reconstructed Image using: Column 2- Our Proposed method; Column 3 - $L_{str} + L_{V/IT}$; Column 4 - $L_{tex} + L_{V/IT}$; Column 5 - $L_{tex} + L_{str}$; Column 6 - $L_{V/IT}$. 

For IXI Image Dataset at 4x superresolution
Fig. 5. Visualization of superresolution results. For each figure the top row is the original image followed by the difference image in the bottom row. Column 1- original image; Reconstructed Image using: Column 2- Our Proposed method; Column 3 - $\mathcal{L}_{str} + \mathcal{L}_{ViT}$; Column 4 - $\mathcal{L}_{tex} + \mathcal{L}_{ViT}$; Column 5 - $\mathcal{L}_{tex} + \mathcal{L}_{str}$; Column 6 - $\mathcal{L}_{ViT}$. 
Fig. 6. (a) Generator Network; (b) Discriminator network. $n64s1$ denotes 64 feature maps (n) and stride (s) 1 for each convolutional layer.
Fig. 7. Loss plots for (a) UNETR Fine tuning using the IXI dataset; (b) Image super-resolution training for IXI dataset.