Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 28 (2025), No. 7, pp. 1191–1222
DOI : 10.47974/JSMS-1409

Advanced analysis of medical image modality using generative and vision transformer model

Hemanta Kumar Bhuyan *
Bhuvan Unhelkar †
Muma College of Business
University of South Florida
8350 N. Tamiami Trail Sarasota
Florida 34243
U.S.A.

S. Siva Shankar §
Department of Computer Science and Engineering
KG Reddy College of Engineering and Technology
Hyderabad 501504
Telangana
India

Prasun Chakrabarti ‡
Department of Computer Science and Engineering
Sir Padampat Singhania University
Udaipur 313601
Rajasthan
lndia

Abstract

Various medical image fusion challenges have been addressed by using convolutional neural network (CNN) based generative adversarial techniques. While CNNs are designed

* *ORCID-ID:* **0000-0002-9712-7280**
† *ORCID-ID:* **0000-0003-1118-3837**
§ *ORCID-ID:* **0000-0002-7616-6088**
‡ *ORCID-ID:* **0000-0001-8062-4144**

* *E-mail:* **hmb.bhuyan@gmail.com (Corresponding Author)**
† *E-mail:* **bunhelkar@usf.edu**
§ *E-mail:* **drsivashankars@gmail.com**
‡ *E-mail:* **drprasun.cse@gmail.com**

for small-scale processing, their inductive bias makes it difficult to learn contextual features. Our solution to this problem is modality-based image analysis, which utilizes vision transformers for their circumstantial compassion. The generator for the vision transformer uses a generative approach to a new cumulative local transformer (CLT) block, which associates with local convolutional and transformer modules as its fundamental bottleneck. The encoder and decoder image techniques are used to analyse the information bottleneck model, and different modality images, such as T1, T2, PD, and FLAIR, are designed to test the image evaluation matrix. Instead of creating separate fusion models for each source-target modality, we have developed a standard procedure. Multi-contrast magnetic resonance imaging (McMRI) relies on MRI scans to ascertain the missing fusion orders. We explored different existing approaches for comparative experimental analysis based on evaluation measures such as PSNR and SSIM. Our models demonstrate that LVT is significantly more successful than competing CNN- and transformer-based methods, making it the ideal choice.

Subject Classification: *92C55 Biomedical imaging and signal processing.*

Keywords: *Healthcare image, Local features, Transformer, CNN, Adversarial, Generative, MRI.*

1. Introduction

Disease image testing plays a vital role in the contemporary healthcare system by facilitating in-vivo clinical investigation of medical datasets. Moreover, comprehending tissue structure enhances the precision and accuracy of medical diagnostics. However, the routine use of multi-modal imaging is hindered by several issues, including patient reluctance and prolonged scan durations [1,2]. In medical images, contextual interactions can be found in healthy and diseased tissues. For example, distant voxels can be interdependent due to the distribution of cerebrospinal fluid in the ventricles or bone in the skull, spanning across spatially continuous or divided brain regions. Disease-specific patterns in the distribution of diseased tissues have been observed despite the absence of regular anatomical priors in normal tissues [3]. Conversely, rare malignant lesions cluster together in the same area, such as gliomas around the cerebrum and cerebellum or meningiomas near the skull [4]. Therefore, the location and morphology of lesions affecting healthy tissue are essential determinants in the spread of pathology. Vision transformers (VT) have great potential in this area, but their computational cost and limited localization make them difficult to apply for pixel-level testing [5]. Nonetheless, recent research [6-10] suggests that hybrid designs or computation-efficient attention operators could provide a solution by switching from healthcare tasks.

Although various testing methods have been developed and utilized for transformer and GAN models, they have all primarily focused on analysing global contextual feature image data to diagnose diseases or abnormalities in human health. Ren et al. [13] developed efficient framework based on Deep CNN model using the Feature Based Medical Image Retrieval (FBMIR) system to analyse the detection for Pneumonia from clinical images. Further, Dhar et. al., [14] have used CNN model for abnormality detection in chest

disease. However, this method has shown insufficient in addressing relevant diseases as observed in image testing data from the human body. Thus, enhancing image quality by capturing images in specific areas is crucial when gathering data from various image-capture technologies. It is necessary to conduct local contextual image analysis to address disease-related image analysis, as different medical problems require distinct types of image testing. Based on the above challenges, a deep learning framework is proposed for healthcare image fusion analysis, consisting of several component architectures. Our suggested system heavily relies on the vision transformer model combined with the CNN technique at various points. Our framework includes structures such as an encoder-decoder architecture, an information bottleneck, a cumulative local transformer (CLT) block, an input/output patching block, an upscaling/downscaling model, and so on, all based on the local vision transformer model (LVT). Other parts are examined outside the structure mentioned above according to methodological norms.

This paper presents several significant contributions, including the following:

- A transformer-based generator that utilizes an adversarial model and input/output patching processing to process imaging modalities.
- Cumulative local transformer (CLT) patches using patching processing and input/output feature mapping, to ensure that the local content of the image's framework is preserved.
- Reduction in computing load by using an information bottleneck with learnable positional encoding to distribute weights for CLT.
- Introducing a unified fusion paradigm that can be applied across various contexts with different source-target modalities.
- Validating our proposed framework and its multiple image modalities with mapping to enable comparison and analysis of numerous image data sets and their accompanying analytical methods.

The following sections of this paper provide detailed information on the research conducted. Section 2 focuses on related work on analysing image models and methodologies. Section 3 outlines the problem statements related to the research topic. Section 4 describes the framework and its various structural components, along with mathematical method analysis. Section 5 analyses the experimental setting, including the dataset, computational environments, and practical components. Section 6 explains the experimental performance, various outputs, and comparative result analysis. Section 7 is dedicated to discussing the paper as a whole, and in section 8, we conclude our framework with an experiment and future work.

2. Related Work

Different approaches under deep learning model have made significant strides in healthcare image systems, particularly in medical image fusion for

situations where modality data is absent. Previous research has proposed patch-level processing local networks [15, 16]. Still, the CNN model has proven more effective due to its ability to benefit from local networks. However, CNN models may not always be attuned to the larger image of the data [17, 18]. With the increasing availability of large image libraries, various types of image assessment are now utilizing the CNN model. In-depth examinations of CNN-based synthesis can be found in references [19–23]. Though CNN training models frequently employ loss functions to grasp undesirable loss data structures [20], significant advancements have been made possible thanks to these references.

A GAN model was utilized in [24] to enhance its ability to capture structural information, incorporating learning-based modalities that adhere to specific criteria. However, when retrieving high-resolution geographical data, the adversarial model cannot improve upon the existing baseline [20]. Nevertheless, GAN-based techniques have demonstrated remarkable performance in various fusion tasks through current research [20,25]. For instance, multi-contrast MRI fusion [26-30], unpaired cross-modality [31], and MR to CT [32-34] are just a few examples where the GAN model has been validated. Recent studies [35-37] have also included maps generated by the CNN model, utilizing spatial or channel attention processes to train the network to focus on potential trouble spots.

While ResViT is designed to translate imaging data between modalities, other methods focus on recreating images using data from a single modality. The hybrid model of Transformer GAN incorporates an adversarial approach that maintains the integrity of the information, while avoiding persistent coupling. Additionally, it utilizes a convolutional encoder-decoder model with a transformer that does not require local connectivity. In contrast, ResViT employs cross-attention transformers to map latent variables to images through an unconditional model, whereas SLATER is conditional on these factors. Recently, several articles have introduced transformer-based techniques for synthesizing medical images.

When differentiating between VTGAN and GANBERT, both utilize transformers alongside convolutional generators. On the other hand, ResViT's transformer-integrated generator intentionally employs a wide-ranging environment. Meanwhile, PTNet's design is devoid of convolutions and excludes a discriminator. In contrast, ResViT utilizes both a GAN and transformer model to produce authentic synthetic images that are both context-sensitive and highly localizable. Among the many tasks that ResViT undertakes are one-to-one and many-to-one translations and the consideration of T1 and T2-weighted images, which are employed for tissue measurements. [38]. The residual paths facilitate the seamless merging of convolutional and contextual representations, resulting in highly efficient processes. To achieve this, we have developed a local vision transformer model that employs CLT blocks as a generator and employs a transformer with skip connections to release image data. The local connections work by convolving and activating accumulative

relative features. Our innovative approach marks the first use of this robust component for testing of various image tasks.

3. Problem Statements

Medical image data has been examined using various transformer methods and GAN models. However, most of these tests focus on global contextual feature images and use different analyses to detect diseases or abnormalities. While these methods may work for image testing data from the human body, they are insufficient for accurately diagnosing diseases. To address this issue, capturing focused, on-the-ground images is crucial for recovering the excellence of images captured by image capture technology. Analysing images connected to disease is challenging because global contextual feature analysis often isn't enough to diagnose a specific illness. Therefore, we have developed Local Vision Transformers (LVT), a complex generative adversarial approach to medical image fusion that utilizes vision transformers' ostensible compassion. Local context image analysis is required for various diagnostic procedures for various disorders. Based on these difficulties, we have offered a deep learning framework for healthcare image fusion analysis to develop a model with various component architectures. The vision transformer model, which is integrated to many parts of the proposed system based on CNN model and has different potential uses in image analysis outside medical imaging.

4. Framework for Local Vision Transformer

We have carefully considered various components to design the Local Vision Transformer (LVT) model. We explain these components using different mathematical approaches, including the construction and reconstruction of images, transfer of image information from one component to another as required, and more.

4.1 *Transformer-based GAN Model*

The GAN model used is a transformer-based hybrid transfer and GAN model that employs convolution and deconvolution techniques, illustrated in Figure 1. The model comprises of two primary components: (a) a Generator and (b) a Discriminator. The generator, as per the model, utilizes the transformer-based approach to produce high-quality image samples from source images with a specific format: $s \in R^{W \times H \times C}$ using a convolution network model as per transformer format. CNN makes the discriminator with specific activation methods. The real image is designed through $x \in R^{W \times H \times C}$ with width (W), height (H), and channel (C). Thus, two categories of images are considered as (a) source image as $I_s \in R^{W \times H \times C}$ and quality-separated image $\widetilde{I_{se}} \in R^{W \times H \times C}$. The discriminator always discriminates the difference between original or inputs and generated sample images i.e., $I_s \; and \; \widetilde{I_{se}}$ images. Using adversarial learning, the generator creates several images for $\widetilde{I_{se}}$. Now, we considered two tools of the GAN model as follows [39]

4.1.1 Design of Generator

We have considered the use of convolution and deconvolution techniques for generators that utilize specific layers and kernel sizes. In our approach, we perform downsampling to extract features through convolution and upsampling to fuse features through deconvolution.

$$I_n^{Conv} = Conv_n(x) \tag{1}$$

Where, $I_n^{Conc} \in R^{W_n \times H_n \times C_n}$ contains down sampling nth feature extraction with $W_n = \frac{W}{2^t}$ and $H_n = \frac{H}{2^t}$. Similarly, deconvolution up sampling fusion feature I_n is defined by

$$I_{n-1}^{Deconv} = Deconv_n(I_n) \tag{2}$$

Here, $I_n \in R^{W_n \times H_n \times 2C_n}$ is the nth layer fusion feature image is created by concatenation of convolution and deconvolution features as:

$$I_n = Concat(I_n^{Conv}, I_n^{Deconv}) \tag{3}$$

We also considered I_{n-1}^{Deconv} for extracting nth fusion features. For n=1, $I_{n-1}^{Deconv} = I_0^{Deconv}$ creates separated images.

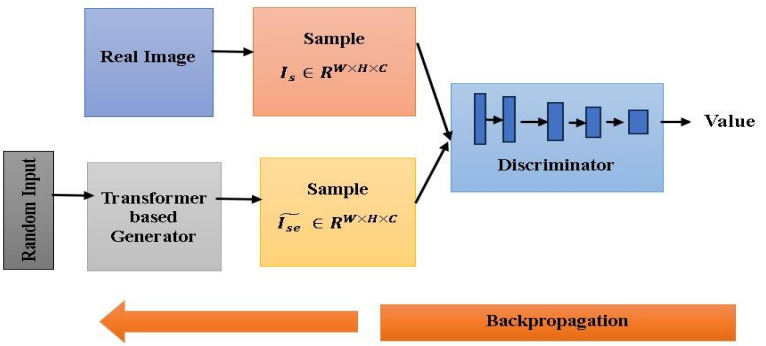


Figure 1
Transformer-based GAN model

4.1.2 Discriminator

In order to distinguish between original and generated images, we utilized a discriminator that employed an activation function (tanh) that was used to create multiple images with normalized feature values (-1, 1) across n-layers. For developing separate images, we considered feature values > 0. Otherwise, it is false.

$$value = \tanh(Conv_n\tilde{I}), \tag{4}$$

Here, we have considered the convolution network's different parameters, which are the same as the generator's. Further, tanh(value) is defined as

$$\tanh(value) = \frac{e^{v} - e^{-value}}{e^{value} - e^{value}}, \tag{5}$$

The equ. (5) is developed the meeting speed as compared to the sigmoid activation function. The role of the generator and discriminator is essential for the GAN model approach. In the next section, we have explained the proposed LVT model with different component analyses.

4.2 *Local Vision Transformer (LVT) model*

Our team investigated using a Local Vision Transformer (LVT) in an adversarial model to test healthcare images that combine multiple source-target modality (STM) formations into a solo framework for improved functionality. The LVT combines CNN operators and transformer blocks to integrate global features based on operational resolution, as depicted in Figure 2. The generator and discriminator both use image translators, with the discriminator subnetwork utilizing convolutional operators. The generator's bottleneck includes cumulative local transformer (CLT) blocks with cascading modules, and a transformer module that collects concealed relative features. Skip connections are strategically placed on modules to establish multiple information channels in the block, and various paths are considered to broadcast information among modules based on the input, including the use of both transformer and CNN techniques for input features, evaluating relative features with the transformer module, and computing fusion local-relative features. Ultimately, this approach results in a combined illustration that synergistically integrates various levels of input features with relative information, using local transformers with local CNN approaches in CLT blocks.

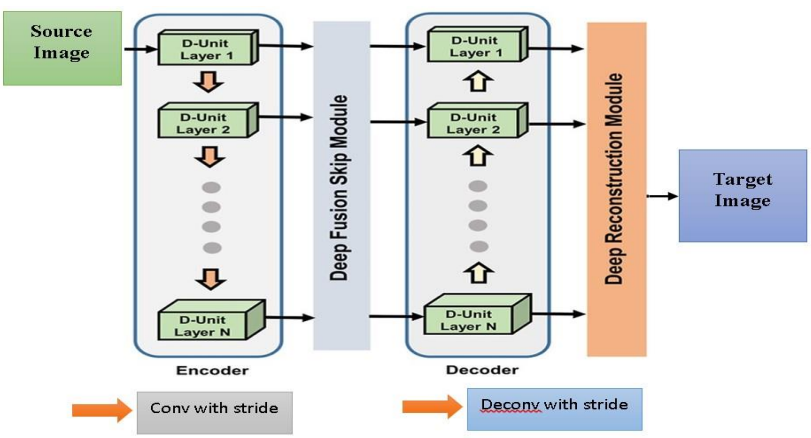


Figure 2
Encoder and decoder model

In Figure 2, the Encoder receives the source image using a convolution layer with a confident stride and transfers it to the decoder through a deep fusion skip module. After receiving encoded images, the decoder uses a deconvolution layer with a confident stride and creates the target image through a deep reconstruction module.

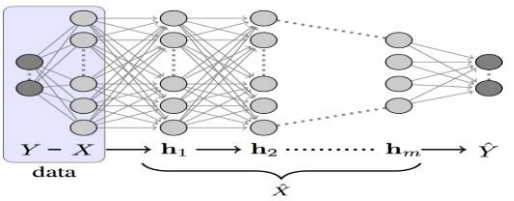


Figure 3

Generic Information Bottleneck. The structure of the Information Bottleneck using a DNN model that maintains the source layer x and produces target layer Y with different hidden layers $h_1,\ldots,h_m$ and the final prediction $\hat{Y}$

We can gain insight into deep learning techniques by utilizing information theory to create an information bottleneck (IB). This IB is also determined through statistical measurements, such as the joint distribution of X and Y, which are included in the training data and various sampled observations. Generally, it has focused on the hidden layer weights (h_i) as our requirements and input variable X, which are both high-dimensional random variables. In the classification settings, the expected value and the ground truth target Y are random variables with lower dimensions. Each layer can be thought of as one state of a Markov Chain if the hidden layers of a DNN are labeled as h_1, h_2,..., and h_m as shown in the above image.: $h_i \rightarrow h_{i+1}$. Although different authors have designed this information bottleneck as their requirements, we don't focus on this component in other ways. We have explained this component with different methodologies in this section per our requirements. Based on Figures 2 and 3, we processed the input image data and performed accordingly as per the Local Vision Transformer model in Figure 4.

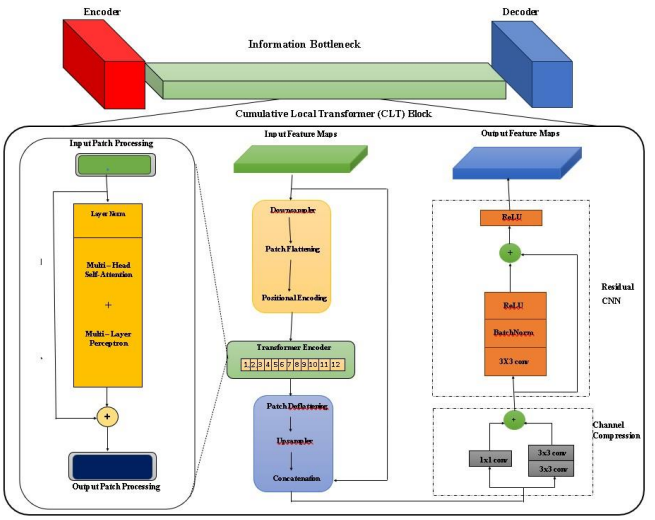


Figure 4

Local Vision Transformer model

The bottlenecked medical image representations for spatial and feature dimensions occur in the centre LVT segment with CLT blocks. The Encoder processes the middle part by using spatially down-sampled feature maps. However, CC modules in CLT blocks have been developed according to the transformer module and the prior CLT block. The series maps for feature dimension are down-sampled by CC modules to extract task-relevant convolutional and relative features. CNN approaches are commonly used for high resolution, which increases local feature compassion as convolutional layers are computationally inexpensive [11]. Usually, vision transformers analyse feature maps at a lower resolution due to computationally demanding self-attention layers. This design ensures that the CNN feature map and transformer module resolutions are compatible. We utilized several modules from [40] to analyse our model and experiments accordingly. The description of different components as per our proposed model is as follows: we took into account the information bottleneck between the encoder and decode is detailed below.

4.2.1 Encoder

The initial segment of the deep network functions as an encoder network comprising multiple convolutional layers. These layers capture localized features from the source images that constitute the first part of LVT. As LVT can function as a cohesive fusion model, the Encoder receives source and target modalities from the imaging protocol as inputs. For source modalities, an identity mapping is utilized, while any unavailable destination modalities are concealed in accordance with the proposed model. Thus, the multi-channel input X^G is formulated as equ. (1).

$$X_i^G = a_i \, . m_i + FGI_k \tag{6}$$

Here 'i' is considered for the input index for the Encoder and m_i is considered as i^{th} modality of image and a_i is the availability of the i^{th} modality.

$$a_i = \begin{cases} 1 \text{ if } m_i \text{ is a source modality} \\ 0 \text{ if } m_i \text{ is a target modality} \end{cases} \tag{7}$$

And FGI_k is a fine-grained image that is made by using both brightness (B_{rt}) and contrast (C_{rt}) of the image with various parameter values defined in next section. This FGI is considered a requirement of image quality.

The multi-modal protocol is used during training to consider various source-target configurations. The availability circumstances in each test subject are used to infer the precise source-target configuration. The Encoder uses convolutional filters to map the different channel input X^G using a feature map $f_{n_e} \in R^{N_C,H,W}$ where N_C is the number channel as the channel count, H and W are the height and width of feature maps. The details of the multi-modal contrast parameter are explained as follows.

The innovative conditional image fusion model seamlessly merges various STM formations into a unified model to enhance image performance during training and testing. (a) LVT incorporates a range of input images during

training using diverse modality protocols, according to STM. For multi-fusion tasks, LVT showcases different formations of STM as the appropriate states. (b) The correct STM formation is determined for all test images based on the inference images using the fitting forms.

4.2.1.1 Multi-contrast Image

We have considered parameter-based multi-contrast images with the following formulas. Let the original image I_O need to change its image quality using multi-contrast parameters. Fine-grained image (FGI) can be made using both brightness (B_{rt}) and contrast (C_{rt}) of the image with various parameter values, Let α_i and β_i are used to perform the different format of F_{gi}. Thus, F_{gi} is formulated as

$$\bigcup_{k=1}^{p} FGI_k = \sum_{i=1}^{n} \sum_{j=1}^{m} (\alpha_i \times B_{rt} + \beta_j \times C_{rt}) \quad (8)$$

Here, α_i and β_j are use used to improve both the brightness and contrast of the image. The 'k' is used p number of image quality. For example, if i = {1,2,3,4,5} and j = {1,2,3,4,5} are used for image quality, then 5 x 5 =25 images can be formed for FGI. From that group of FGI, we can select which one is better than the others for our experiments. Accordingly, quality of images can be used for various required experimental evaluations.

4.2.2 Deep neural network-based Information Bottleneck

There is ongoing debate regarding the information bottleneck model and its relationship with actual evidence. To determine the relevant information volumes during image data processing, various evidence and methods are examined thoroughly. One approach is to use Deep Neural Networks (DNN) and the theory of Information Bottleneck (IB), where any hidden network layer is considered between starting and ending layers. The information bottleneck maintains a trade-off through mutual information (MI) measurement, quantifying the knowledge between starting and ending layers. Research suggests that a DNN undergoes two distinct training phases: a fitting phase that involves increases and a compression phase that involves decreases. However,

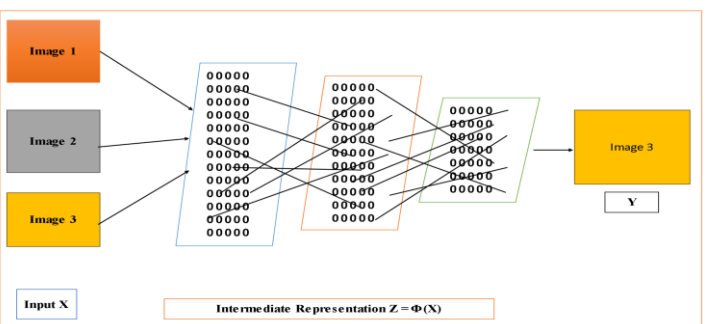


Figure 5
Information bottleneck with deep neural network parameters X, Y, Z.

some argue that the compression phenomena in DNNs are not universal and depend on the specific activation function. For instance, Saxe et al. [41] claim that ReLu activation functions prevent compression from occurring in images. This paper has delved into the intricacies of the information bottleneck (IB) as [42], utilizing a specialized notation that allows for implicit minimization of extraneous information and maximization of contextual knowledge, as illustrated in Figure 5. To further explore its potential, we have employed various formulas to analyse the benefits of IB in deep learning and identify any related errors, which appear to be closely tied to the degree of information bottleneck.

We have used specific notation to extract information with the representation of $Z = \emptyset(X) \in \mathbb{Z}$ as per input $X \in \chi$. The IB is considered to balance the information based on lessening the MI between X and Z as I(X; Z) while maximizing the MI between Y and Z as I(Y; Z). There is a connection between I(X; Z) and the machine learning model for various understandings of the IB model. Thus, the training data is designed with probability as per [43]. The training data $s = \{(x_i, y_i)\}_{i=1}^{n}$ is drawn with probability 1-δ from the same probability distribution with a random variable (X, Y). The error gradient Δ(s) is defined through statistical expectation as follows:

$$\Delta s = E_{X,Y}[\ell(F^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f^s(x_i), y_i), \quad (9)$$

This error has a bounded form as

$$\Delta(s) \leq \sqrt{\frac{2^{I(X;Z_l^s)+log\frac{2}{\delta}}}{2n}},$$

Here, the training model is f^s, the result of the hidden layer of the Encoder. $\emptyset_l^s$ is generated by $\emptyset_l^s(X)$. This is achieved after the first layers. The data step moves towards the Encoder using IB with a particular bounded form. Thus, Hafez-Kolahi et al. [44] considered the determinate form as per the input layer and created an input compact bound for binary classification. The classification is designed with Y = {0,1} and ℓ is the (0 - 1) loss. The training dataset s is defined with probability at least 1 – δ is

$$\Delta(s) = \tilde{\mathcal{O}}\left(\sqrt{\frac{2^{H(X)}}{n}}\right) \quad (10)$$

The equ. (10) is satisfied for $2^{\frac{1}{\epsilon}}$ where,

$$\epsilon = \Omega\left(\sqrt{(2^{\frac{6H(X)}{\epsilon}} + \log\left(\frac{1}{\delta}\right) + 2)/n}\right)$$

The generalization error is defined by high probability is as follows.

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{I(X;Z_l^s|Y)+1}{n}}\right) \text{ as } n \to \infty \quad (11)$$

The equ. (11) is improved the numerator bound form with improving $I(X; Z_l^s)$ to smaller quantity $I(X; Z_l^s|Y)$.

We aim to examine the establishment of complexity bounds in relation to the interplay between information bottleneck (IB) and generalization errors.

4.2.2.1 Independent training data in Encoder

We considered the mathematical model for IB with conditional MI for optimizing control of expected loss. We try to minimize the conditional MI $I(X;Z_l^s|Y)$ and training loss as per the stable Encoder $\emptyset_l^s$ and independent training data s. It is very useful for data independent of s for transfer and unsupervised learning models. We have considered two theorems for understanding IB with an encoder. The reader can follow the reference [45].

Theorem 1: *For different layers $\ell \in \{1,2,\dots D\}$. If the stable Encoder $\emptyset_l^s$ and independent training data s for BI, then, for any positive δ with probability 1- δ on training data s, the generalization error is formulated as:*

$$\Delta(s) \leq G_3^l \sqrt{\frac{I(X;Z_l^s|Y)\ln(2)+G_2^l}{n}} + \frac{G_1^l(0)}{\sqrt{n}} \tag{12}$$

Here, $G_1^l(0) = \tilde{\mathcal{O}}(1)$, $G_2^l(0) = \tilde{\mathcal{O}}(1)$, $G_3^l(0) = \tilde{\mathcal{O}}(1)$, as n $\rightarrow \infty$.

The proof of this theorem is available in [45]. The reader can follow this proof for more details. In theorem 1, we found two significant bounds by comparing previous bounds.

(i) In the first case, we lessen the dependency using an exponential growth rate $2^{I(X:Z_l^s)}$ as per the linear growth rate $I(X;Z_l^s)$.

(ii) In the second case, we are improving $I(X;Z_l^s)$ to smaller quantity $I(X;Z_l^s|Y)$ as per the expected MI between X and Z_l^s with conditioned on Y.

We minimize the $I(X;Z_l^s|Y) \geq 0$ with maximizing the predicted information as $I(Y;Z_l^s) \geq 0$. This is done through BI.

4.2.2.2 Learning Encoder-based training data

The basic concept of IB is to try to minimize $I(X;Z_l^s) = I(X;\emptyset_l^s(X))$ as per the variable of the Encoder $\emptyset_l^s$ with discriminative objectives. This for IB rule, we can set the learning encoder. $\emptyset_l^s$ with data set s. The output is obtained from IB regularize $I(X;Z_l^s|Y)$ with MI of encoder and training data set $I(\emptyset_l^s;S)$.

Theorem 2: *For all layers $D \subseteq \{1,2\dots.,D+1\}$. Then, for any δ >0, the generalization bound holds with probability 1 – δ over the training set s.*

$$\Delta(s) \leq \min_{l\in D} Q_l \tag{13}$$

Where, for $l \leq D$,

$$Q_l = G_3^l \sqrt{\frac{\left(I(X;Z_l^s|Y)+I(\emptyset_l^S;S)\right)\ln(2)+\tilde{G}_2^l}{n}} + \frac{G_1^l(\zeta)}{\sqrt{n}}; \tag{14}$$

And for $l = D+1$,

$$Q_l = R(f^s)\sqrt{\frac{I(\emptyset_l^S;S)\ln(2)+G_2^l}{2n}} \quad (15)$$

Here, $S \sim P^{\otimes n}$, $G_1^l(\zeta) = \tilde{O}\left(\sqrt{I(\emptyset_l^S;S)+1}\right)$, $\tilde{G}_1^l = \tilde{O}(1)$, $\tilde{G}_2^l = \tilde{O}(1)$, $\tilde{G}_3^l = \tilde{O}(1)$, as n → ∞.

The proof of this theorem and the formula of $G_1^l(\zeta)$, $\tilde{G}_1^l, \tilde{G}_2^l, \tilde{G}_3^l$ are available in [45].

The learning encoder sets the complexity bound of IB $\emptyset_l^s$ with data set s. The random variable Z_l^S is considered with l layers with depending training data set s and D is considered for all layers with the last hidden layers Z_D^S and Z_{D+1}^S as output layer, theorem 2 is useful as per trained Encoder with independent data set s. Since the Encoder is supposed to keep less dependency over S (target data) in transferring the learning model, it inclines to lessen $I(\emptyset_l^S;S)$.

4.3 *Decoder*

The generator of the proposed model's third and final part is a decoder built using altered convolutional layers. LVT decoder maintains a multi-contrast approach inside various modal protocols since it can act as a unified model, irrespective of the individual combination. The feature maps f_A that the bottleneck has reduced are sent into the decoder, which then generates multi-modality images Y_i^G in distinct channels with synthesized modality.

Here, two concatenation methods are defined for synthetic and acquired images. PatchGAN calculates the probabilities of two outcomes as per getting input image.

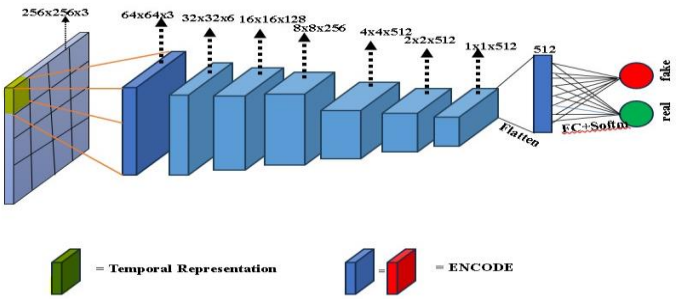


Figure 6
The construction PatchGAN model

Figure 6 shows the pGAN structure to identify whether an image is real or fake using discriminator architecture. Although the part is not our goal, we avoid this part for details.

4.4 *Multi-modal contrast parameter*

Running various sequences with different weighted tissue measurement images has been considered, using contrast parameters as follows: There are

three primary images - T1-weighted (highest T1 grey contrast), T2-weighted (bright T2 contrast), and weighting based on proton density (PD) (hydrogen proton density). Multi-modal imaging techniques that utilize a combination of various imaging modalities have shown promising results for diagnosis, therapy, and outcome prediction. MRI programs such as T1 dark and T2 bright, among others, enable the multi-parameter display of anatomical structures. However, the effectiveness of MRI detection is limited due to the prevalence of false-positive diagnoses, as MR images are naturally produced in black and white. In this paper, we introduce an intriguing dual-modal program based on contrast agent augmentation. Our focus will be on the interaction between T1 and T2 relaxation mechanisms, shedding light on the molecular foundation of MRI through contrast agents.

In clinical diagnosis, the atomic nuclear magnetization signal is the most commonly used anatomical tool. It's reconstructed into 2D and 3D images using non-invasive, non-ionizing, and radiation-free MRI technology. Weighted imaging techniques based on various factors such as longitudinal (T1) or transverse (T2) relaxation times, molecular diffusion, and proton density can be used instead of the standard MRI imaging techniques. T1 varies among tissue types in nature and is commonly used in brain anatomic imaging. T2, on the other hand, describes the length of time required and is also highly influenced by tissue type. T2-weighted MRI is typically accomplished by echo time (TE) weighted MRI pulse sequences, which display brighter for a longer T2 for the FLA1R sequence. T1-weighted MRI is typically produced by inversion-recovery MRI pulse sequences, which show darker for longer T1. The use of Proton density-weighted (PDw) MRI is quite common in medical imaging. This technique was designed to minimize the impact of T1 and T2 effects by utilizing optimal parameters, resulting in images that are primarily dependent on the density of protons within the imaging volume. Furthermore, there are a range of alternative imaging sequences that come in parameter packages. Recently, new MRI sequences have been developed that allow for quantitatively parametrical imaging that is clinically practical. MRI scans can provide quantitative mapping of both T1 and T2 simultaneously, making them a valuable tool for practitioners. However, imaging protocols are often influenced by the practitioner's empirical judgment. On the other hand, T2-weighted imaging is typically used to analyse water content, which appears as a bright signal.

4.5 *Modality Basis MRI*

As a result of the Zeeman splitting phenomenon, nuclei are divided into high and low energy levels, exhibiting a Boltzmann equilibrium state. To simplify matters, we have designated T1 and T2 as the relaxation rate constants in the xy and z directions, respectively. Various rotations or directions of dimension were explored to obtain T1 and T2 measurements, as illustrated in Figure 7. In Figure 8: (a) Dephasing of M_{xy} correlates with the phenomenon of T2 relaxation. (b) The recovery of the M_z in z-direction with nuclear spin is referred to as the T1 relaxation phenomenon. This fundamental process results in

a substantial T2 dephasing effect that attenuates T1 in the measurement. (c) The T1 and T2 relaxation increase the water protons which results from the effective T1 and T2 relaxation enhancement. (d) The multiple-parameter MRI demonstration is featured in the logic which can be defined on T1 and T2 as OFF-ON condition as 'o' for OFF and '1' for ON with different states. The reader can follow [47] for details of this model.

A standard technique for MRI investigations involves sequentially acquiring T1 and T2 images using a T1-T2 dual-modal approach. Pathological abnormalities or inflammations often coincide with changes in water content. This technique is believed to significantly improve diagnostic accuracy for diseased tissues. In modern medicine, accurate diagnosis can have a significant impact on treatment and outcomes. Contrast agents can provide mutual-confirmatory information from both pre- and post-contrast T1 and T2 images.

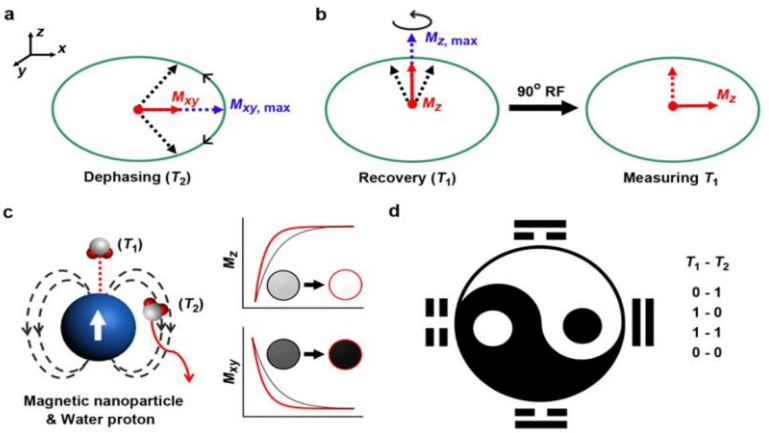


Figure 7
Formation of T1 and T2 as per direction

Despite its potential, the T1-T2 dual-modal MRI technique has shown significant variances in responsive T1 or T2 MRI investigations, which raises doubts about its broad applicability. However, this technique holds promise in shedding insight into the complex design of responsive MRI systems, making it an area of interest for future research. Developing T1-T2 dual-modal responsive MRI applications could lead to new insights into the process of magnetic material-induced MRI contrast enhancement.

4.6 *Parametric Maps*

We used four parameters, namely T1, T2, PD, and FLAIR, to evaluate quantitative performance based on learning-based approaches. To transform the input T1 and T2 weighted images, we utilized an endwise drawing function. These parameters - T1 (spin-lattice relaxation time or darkness or hypo intensity), T2 (spin-spin relaxation time or brightness or hyperintense), PD

(proton density), and FLAIR (fluid-attenuated inversion recovery) - were used for both one-to-one mapping and many-to-one mapping. We considered the ranges of these parameters based on white matter (WM0), gray matter (GM), and cerebrospinal fluid (CSF) to generate synthetic brain volumes. To avoid any corresponding background, we used weighted parameters for T1 and T2 as T1W and T2W images with their appropriate intensity. We took into account performance validation for the proposed model, utilizing parametric maps and different weighted image modalities in accordance with those used for the CNN model. The quality of synthesis was evaluated using direct mapping with synthesized weighted images. To assess quantitative parameters such as PNSR and SSIM, we employed various parameter mapping techniques with multiple GAN and Transformer approaches, which are discussed in detail in the experimental sections.

5. Experiments

5.1. *Datasets*

We have considered different healthcare image datasets such as the Brain MRI datasets (BRATS [48]) and MRI-CT datasets [49]. The proposed LVT model was tested on these two datasets with various evaluation parameters and compared with existing models.

(1) Information extraction from Images (IXI) Dataset: Our study analysed brain MR images from a group of healthy participants. To explore various multi-contrast weighted parameters, we utilized data sets that included T1, T2, and PD-weighted parameters. Prior to conducting any evaluations or experiments, we established a mapping mechanism among the T1, T2, and PD-weighted images as our requirement. Additionally, our data set factored in the echo time (TE) and spatial resolution [50].

(2) Brain Tumor Segmentation (BRATS) Dataset: Brain MR scans of participants were examined using T1, T2, and T1, T2-weighted images0, and FLAIR images. Twenty-five data were considered for training, ten for validation, and twenty for testing. Here, the BRATS dataset includes images taken with various scanners and clinical regimens from different institutions.

(3) MRI-CT Dataset: We examined the male pelvis in images of our patients for our experiments. Nine participants were reserved for training, two for validation, and four for testing. We examined ninety axial cross-sections from each patient. The following acquisition criteria were used. Group 1, T2-weighted images, Time to Echo (TE) = 97 ms, TR = 6000–6600 ms, spatial resolution = 0.875 x 0.875 x 2.5 mm^3. Group 2, with TE = 91-102 ms, Repetition Time (TR) of 12000–16000 ms, and spatial resolution of 0.875–1.1 x 0.875–1.1 x 2.5 mm3. Group 1 of the CT scans has a spatial resolution of 0.98 mm3 and a kernel of B30f. Group 2: Kernel = FC17, spatial resolution = 0.10.12mm3. Images from this collection were gathered using multiple methods.

5.2 *Differing Approaches*

We tested the LVT model against several cutting-edge image generation techniques. The standard methods, including attention-augmented convolutional models, are explained as follows.

5.2.1 Convolutional models

(a) **pGAN-based:** In our approach, we utilized the patchGAN (pGAN) model with a ResNet backbone [20] to assess the authenticity of a single input image through probability determination. Unlike a scalar output, pGAN generates an NxN output vector to calculate outcome probabilities for individual image sections. Its generator employs the same Encoder and decoder as LVT, with a series of residual CNN blocks located in the bottleneck. Two key components of pGAN were instrumental in image identification, including the bottleneck procedure between the Encoder and decoder aided by CNN blocks.

(b) **pix2pix**: This model was considered for the GAN model with a U-Net backbone [58]. Here, encoder-decoder techniques were used by skip connections in pix2pix.

(c) **MM-GAN:** The convolutional GAN-based fusion model was considered [29]. The discriminator and generator networks in MM-GAN were CNN-based generators. A single network was trained using MM-GAN in a variety of source-target modalities. The unification approach used by ResViT and MM-GAN were identical.

(d) **$pGAN_{uni}$:** To combine several synthesis tasks, an integrated pGAN model was trained. The unification process was the same as with ResViT.

5.2.2 Enhanced Attention based Convolutional Models

(a) **Attention U-Net (A-UNet):** Here, the discriminator was the same as ResViT's, and the original A-UNet model was used to generate a conditional GAN model [51].

(b) **SAGAN, or Self-Attention GAN:** This model is considered to develop the image quality using soft max and attention map as shown below [52]. In this work, SAGAN is presented to yield acceptable hyperspectral samples. To mitigate the effects of over-fitting, it seeks to rise the quantity and quality of training dataset. The SAGAN model is designed in Figure 8.

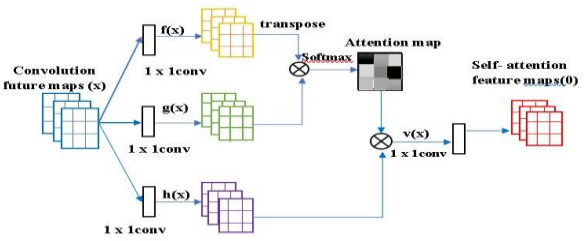


Figure 8
SAGAN (Self-Attention Generative Adversarial Networks) Model

Deep learning's rise as a model is a significant turning point in generative modeling i.e., GAN which creates learning model as deep learning approaches.

5.2.3 Transformer Models

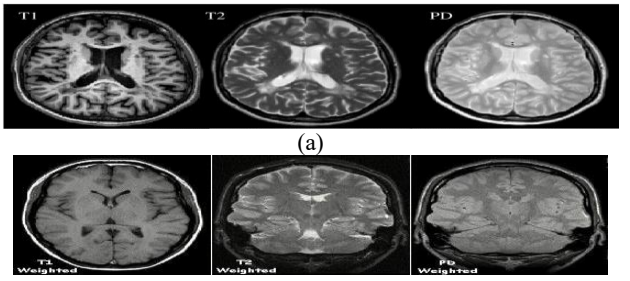
(a) **TransUNet:** We used the initial TransUNet model to generate a GAN model with a discriminator that was the same as ResViT [5]. Furthermore, we added a convolutional layer for synthesis in place of the segmentation head. In applications for healthcare image segmentation, a Transformers-based U-Net architecture achieves up-to-date performance.

(b) **PTNet**: Recently, a transformer architecture without convolution was considered [38]. Pyramid Transformer Net (PTNet), a brand-new framework for MRI synthesis was introduced.

(c) **TransUNet$_{uni}$:** Multiple synthesis tasks were consolidated using the TransUNet architecture. The unification process was the same as with ResViT.

5.3 *T1, T2, and FLAIR-based MRI image*

We have examined basic parametric images, as depicted in Figure 9(a). To distinguish between various types of images, we have utilized T1, T2, and PD-based MRI images, which are represented by dark, bright, and (light + bright) colors, respectively. Additionally, Figure 9(b) includes the corresponding weighted images.



(a)

(b)

Figure 9

(a) T1, T2 and PD based MRI images, (b) T1 weighted, T2 weighted and PD weighted images

We considered three weighted MRI images based on TR (Repetition Time) and TE (Time to Echo) times, where T1-weighted used a short period, T2-weighted used an extended period, and T1-weighted used a very long period.

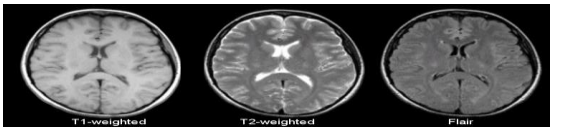


Figure 10

Comparison of T1 vs. T2 vs. Flair (Brain)

When arranging mutual MRI scans, Figure 10 is often used to produce both T1 and T2-weighted images. These images are made through short TE and TR times and highlight the tissue's T1 properties. Similarly, it happened for T2-weighted images. To differentiate between the two images, CSF can be used since it appears on T1 and T2-weighted imaging. The Flair is comparable to T2-weighted images but employs very long TE and TR times.

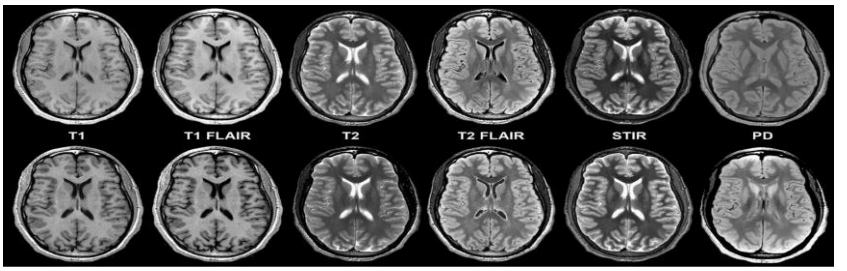


Figure 11
Different parametric images

Moreover, we can distinguish various parametric images as depicted in Figure 11. The figure comprises two rows of images, each featuring contrast-based images.

6. Experimental performance analysis

6.1 *Quantitative performance*

In these experiments, we utilized two quantitative approaches to compare different methods: the peak signal-to-noise ratio (PSNR) [12] and the structural similarity index measure (SSIM) [12]. Both metrics were used to estimate the separative performance quantitatively. PSNR determines the ratio between the optimal signal and noise or the separated image from the source image. It focuses on the difference between pixels and does not consider structural similarity. On the other hand, SSIM combines contrast and texture of images to define structural similarity. To achieve better image construction, we considered higher values of PSNR and SSIM. The PSNR is defined through a log-scale formula as follows.

$$PSNR = 10log_{10}\frac{MAX_I^2}{MSE} \tag{16}$$

Here, MAX_I is the optimal value of the Image, and the mean-variance between source and generated images determines MSE. Similarly, the SSIM is defined as a similarity structure between the source image (I_s) and the separated or generated image ($\widetilde{I_{se}}$) as follows.

$$SSIM\left(I_s, \widetilde{I_{se}}\right) = \frac{(2\mu_{I_s}\mu_{\widetilde{I_{se}}}+k_1)(2\sigma_{I_s\widetilde{I_{se}}}+k_2)}{(\mu_{I_s}^2+\mu_{\widetilde{I_{se}}}^2+k_1)(\sigma_{I_s}^2+\sigma_{\widetilde{I_{se}}}^2+k_2)} \tag{17}$$

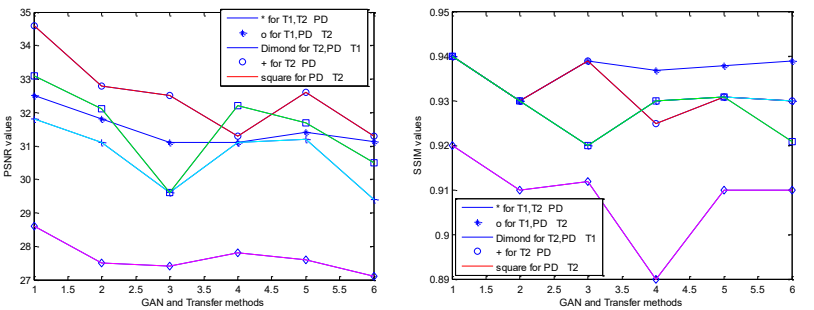
Here, we considered $\mu_{I_s},\mu_{\widetilde{I_{se}}}$ determined the mean of (I_s) and ($\widetilde{I_{se}}$) ; $\sigma_{I_s}^2$, $\sigma_{\widetilde{I_{se}}}^2$ determined the variance of $\mu_{I_s},\mu_{\widetilde{I_{se}}}$

$\sigma_{I_s\widetilde{I_{se}}}$ is defined as the covariance of (I_s) and ($\widetilde{I_{se}}$); k_1, k_2 maintained stability of constant values through a certain range of pixel values as $k_1 = (E_1L)^2$ and $k_2 = (E_2L)^2$ where L determined the various range of pixel values and E_1, E_2 are empirical values. Similarly, the SSIM is determined within the values between 0 and 1. If the SSIM value is close to 1, the source image seems too similar to the generated or separated images.

6.2 *Multi-Contrast MRI (McMRI) fusion testing*

6.2.1 Image Fusion Model

After careful consideration, we have found that the LVT method is highly effective in detecting image fusion models for McMRI. In order to compare enhanced CNN models (such as A-UNet and SAGAN), convolutional models (pGAN and pix2pix), and modern transformer architectures (TransUNet and PTNet), we first analysed healthy brain images from the IXI dataset. Figure 16 was used to compute PSNR and SSIM measurements for utilizing various mappings such as (T1, T2) →PD, (T1, PD)→T1, (T2, PD)→T1, T2→PD, and PD→T2. In both challenges, LVT outperformed other methods.



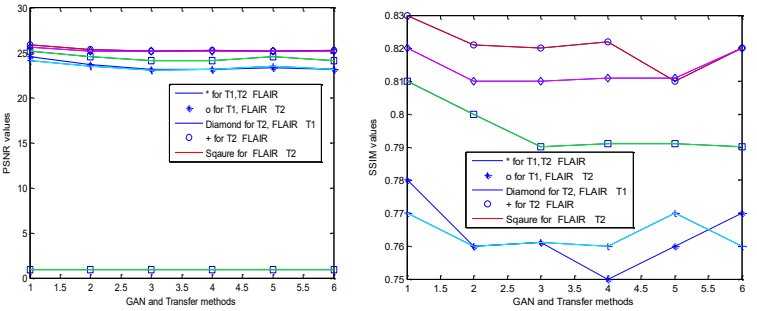
(a) PSNR (b) SSIM

Figure 12

Evaluation of fusion models on different mapping tasks for PSNR and SSIM.

According to our findings in Figure 12, we have discovered that using the parameter mapping of (T1, PD) → T2 yields better performance for PSNR evaluation compared to other mappings, as shown in Figure 12(a). On the other hand, for SSIM evaluation, the parameter mapping of (T1, T2) → PD performs better, as shown in Figure 12(b). We conducted experiments using various mappings and different methods, including LVT, pGAN, pix2pix, A-Unet, SAGAN, and TransUNet. After analysing the results, we found that the LVT method outperforms the others. To determine the best method, we considered the performance of both PSNR and SSIM, and chose the one with the highest score. As the LVT method was the first one we tested, it had the highest scores in both metrics, PSNR and SSIM. In our evaluation of various methods, we have explored different mappings between T1, T2, and FLAIR, as illustrated in Figure

13. We found that LVT demonstrates strong performance for most tasks, with the exception of T2 → FLAIR, where A-UNet slightly outperforms SSIM. Compared to transformer models, LVT exhibits superior results, with an average improvement of 1.56dB in PSNR and 1.63% in SSIM.



(a) PSNR (b) SSIM

Figure 13

Evaluation of fusion models on different mapping among T1, T2, and FLAIR tasks for PSNR and SSIM.

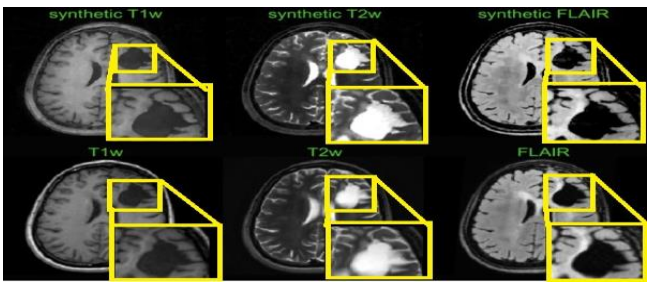


Figure 14

LVT improves the synthesis performance through a clinical reference for tumor region.

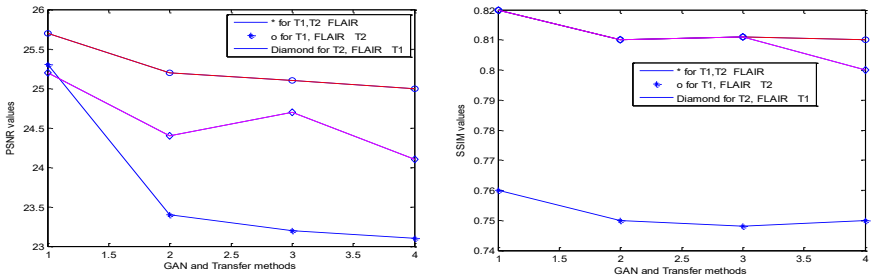
We have explored various synthetic weighted imaging options, including Synthetic T1W, Synthetic T2W, and Synthetic FLAIR tissue protocol mapping, in order to accurately identify tumor regions. Specifically, we examined three different weighted images and compared their performance to clinical reference images in Figure 14. The first row showcases synthesized images while the second row displays the clinical reference images. Our findings revealed that the synthesized images outperformed the clinical reference images in detecting tumors from MRI images.

6.2.2 Integrated Fusion Models

To increase performance, the LVT model is trained and evaluated to complete a single fusion task; nevertheless, a different model must be created for each task. We then showed how LVT may be used to learn Integrated Fusion

Models for multi-contrast MRI. Comparisons are made between an integrated LVT (LVT_{uni}) and unified convolutional ($pGAN_{uni}$, MMGAN) and transformer models ($TransUNet_{uni}$). LVT_{uni} continues to perform at the top level in different tasks in both IXI and BRATS (p 0.05). LVT_{uni} performs better than $pGAN_{uni}$, MM-GAN, and $TransUNet_{uni}$ in IXI by 1.12dB PSNR, 1.80% SSIM, 2.37dB PSNR, and 2.69dB PSNR, respectively (p0.05). LVT performs better in BRATS than pGANuni, MM-GAN, and TransUNetuni (p 0.05) by 0.74 dB PSNR. In Fig. 15, representative target images are shown. LVT creates target images that are sharper and have fewer artifacts than baselines. These findings indicate that models for various source-target configurations can be successfully consolidated using a single LVT model.

As expected, attention-augmented models (AAM) surpass pure convolutional models compared to standard models. However, LVT, which directly integrates contextual relationships, surpasses all standard models. In Figure 15 (a and b), sample target images are displayed for different tasks. LVT produces more precise goal images in terms of artifact levels and clearer tissue descriptions compared to standard models. Notably, LVT correctly represents brain lesions in patients, unlike competing techniques such as TransUNet, which produce incorrect images.



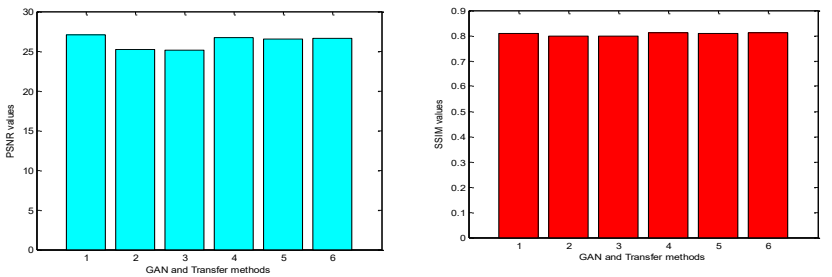
(a) PSNR (b) SSIM

Figure 15
Performance of United Fusion Models (T1, T2 → FLAIR, T1, FLAIR → T2, AND T2, FLAIR → T1)

We considered united fusion models through the LVT model. Thus, we considered the proposed united models LVT_{uni} which is compared with other united methods such as $pGAN_{uni}$, MM-GAN, and $TrabsUNet_{uni}$. We have taken three mappings as shown in figure 15. We also obtained that LVT_{uni} has the highest performance than other methods. The superior pathology representation in LVT highlights the significance of CLT blocks in preserving local accuracy and relative steadiness in medical image fusion. TransUNet's bottleneck simply utilizes a transformer to propagate narrow convolutional features using an encoder-decoder skip links model, which can be unsuccessful at reducing better frequency artifacts [54].

6.2.3 Fusion Across Modalities

The LVT model proves to be a powerful tool for cross-modal synthesis, taking into account pelvic datasets. Figure 16 showcases the impressive performance of representative target images. Notably, LVT excels in the vicinity of bone structures in CT images. Research indicates that MRI-CT synthesis affords contextual representations a larger proportional weight. In this pursuit, LVT offers reliable performance and precise tissue portrayal through its residual transformer blocks.



(a) PSNR (b) SSIM

Figure 16
Evaluation of Modality based Fusion Task (T2-Weighted MIR → CT) in the Pelvic MRI-CT Dataset

Figure 16 reveals that the LVT method performed admirably when compared to other methods in both evaluation metrics such as PSNR and SSIM. In Figure 16, 1st Bar is LVT method performance, remaining are other methods as per proposed method.

6.3 *Analysis of denoised image*

(a) In evaluating denoised image form, we utilized NIFTI images and transformed input images into T1 and T2 images. To test the U-net model, we considered two evaluation matrices, namely PNSR and SSIM, based on an input shape of (128,128,64). For U-net testing, we implemented encoder and decoder techniques in the following manner. For Encoder, two convolutions are considered as follows: (a) for conv1, Conv2D (64,3) is designed with Relu activation amd pool1 is considered with Maxpooling2D(pool_size(2,2)). (b) Similarly, for Conv2, Conv2D (128,3) is designed with Relu activation amd pool2 is considered with Maxpooling2D (pool_size (2,2)).

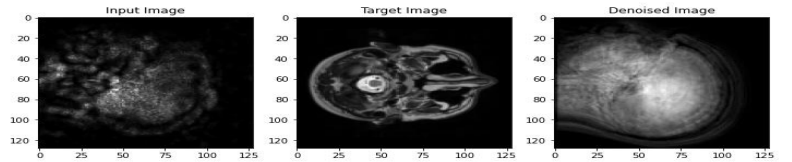
(b) Upsampler is designed with Conv2D (64,2) and relu activation for decoder. It also use upsampling2D size (2,2) with Conv2D. Merge3 is designed with concatenation of {Conv1 and Upsampler}. Conv3 is considered with upsampler and merge3 and repeated with itself.

Our process involves generating input and target data using numpy's random function, with input_data and target_data being of shape (1, 128, 128, 64). We then resize both images to a common 128x128 shape, with the denoised images being resized to match the target images. Our testing involved a single epoch with 8 batches, and the resulting loss values are presented in Table 1.

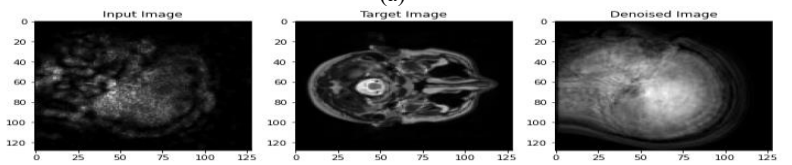
Table 1
Experimented loss evaluation

Epoch	Batch	Loss
1/1	1/8	92653.7656
1/1	2/8	81936.5547
1/1	3/8	84883.0312
1/1	4/8	96804.6094
1/1	5/8	69065.0000
1/1	6/8	103969.5781
1/1	7/8	7989.2588
1/1	8/8	54195.7266

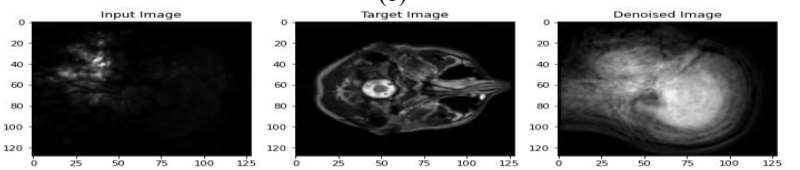
The evaluation of PSNR and SSIM are as (a) psnr = calculate_psnr (target_ image, denoised_images) (b) ssim = calculate_ssim (target_image, denoised_ images). Thus, we got Validation PSNR: -46.26 and Validation SSIM: 0.0990. The experimental images are shown in Figure 17 as input images and produce target and denoise images.



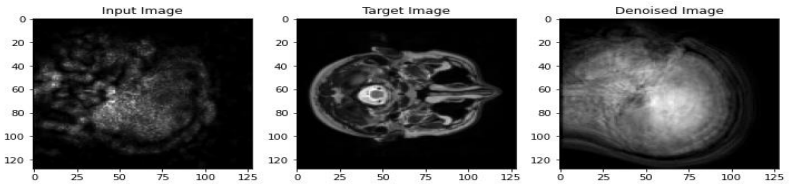
(a)



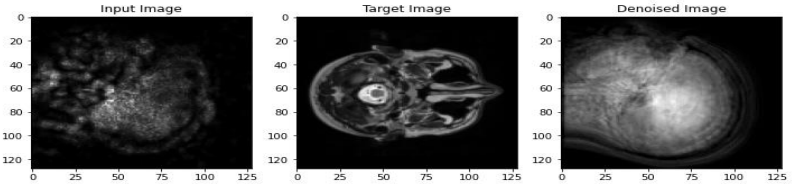
(b)



(c)



(d)



(e)

Figure 17
All Input image, Target image and Denoise image using Unet model

We utilized NumPy arrays to evaluate all images using the A-Unet model. T1 was selected for the input image, while T2 was chosen for the target image. To ensure consistency, all images were maintained at a common size of 128x128 for Input, Target, and Denoise images. Additionally, all images were set to a figsize of (12,4) and cmap = 'gray'. Figure 17 (a,b,c,d,e) required varying amounts of testing time per step for the aforementioned images, ranging from 71ms to 63ms. Notably, all target and denoised images were obtained from their corresponding input images.

7. Discussion

Different approaches are considered to evaluate the model based on the advantages of LVT's essential part of the method. Firstly, it considered synthesis tasks in the test set. LVT has consistently demonstrated optimal or near-optimal performance across these tasks, while maintaining a significantly lower FID than the variation loss. Compared to traditional metrics like PSNR or SSIM, FID is widely considered a more accurate indicator of the perceived benefits of adversarial learning [53], as it takes into account higher frequencies.

Secondly, we compared LVT with ablated variations that omitted the weight-tying process across transformer modules. LVT outperformed the variants in all typical tasks ($p < 0.05$), which yielded a similar SSIM in the T1, T2 → PD. Next, we conducted an analysis of the down- and up-sampling blocks and skip connections in the proposed architecture to gauge their effectiveness. Our findings indicate that LVT outperforms all other versions in terms of performance (with a significance level of $p < 0.05$). We also compared LVT to variations that involved removing all down- and up-sampling modules from CLT blocks and substituting unlearned max-pooling/bilinear interpolation modules, as well as variants created by accelerating Encoder down- and decoder up-sampling rates. All of the variations described in these experiments perform

worse than LVT (with a significance level of $p < 0.05$). Lastly, we evaluated the strong characteristics acquired from transformers in distilled representations inside CLT blocks.

Finally, we would like to highlight the significant role played by LVT self-attention processes in enhancing synthesis performance. Our experimentation indicates that both LVT and pGAN exhibit similar performance, with LVT being more effective in minimizing fusion errors in areas of greater attentional focus. The image synthesis is further supported by the use of synthetic images and error maps. According to research, traditional GANs using convolutional operators are unable to capture long-range images [56]. To address this, we included weighted parameters among transformer modules to simplify the model. Our findings indicate that LVT continues to outperform competing approaches in terms of performance. Convolutional Neural Networks (CNNs) is handling abnormal anatomical differences between subjects [20,46]. Convolutional filtering remains the primary method for feature representation extraction.

Recent research has proposed transformer-based models for medical image fusion, with only a handful of studies published [10,38]. Our work stands out in several ways. Firstly, we utilize a transformers model in an LVT generator to process latent relative representations from source images, which differs from [10]. Secondly, to maintain realism, we employ an adversarial loss instead of the mean-squared error loss used in [38], as the latter can over-smooth target images [20]. Thirdly, we proposed the contextual sensitivity of transformers with CNNs' localization skills, unlike [38], which used a transformer design without convolution. Lastly, the LVT helps preserve low-level characteristics due to the hourglass's midpoints' significantly reduced spatial resolution. Encoder-decoder skip links are not present in this model for several reasons. For instance, the LVT maintains high resolution at the Encoder's output (e.g., 64x64 maps), reflecting relatively lower-level data in its bottleneck.

The encoder and transformer modules downscaled 256 x 256 images by 16 times cumulatively, making LVT usable in various ways and at various image resolutions. To construct the Encoder, non-integer downsampling rates can be used, or the resolution can be rounded up to the nearest power of two after zero-padding [55]. Incorporating multi-scale modules into the decoder can also aid LVT in preserving fine image features [57].

8. Conclusions

We have developed a fusion technique for multi-modal imaging that uses constrained deep adversarial networks. By taking into account the information bottleneck of the LVT model, we were able to create a multi-layered, cumulative local transformer model that includes encoder and decoder components. We also considered input feature mapping models that preserve up/down sampling approaches for the transformer encoder during input and output patching. To plan the information bottleneck, we utilized a self-attention model. Additionally, we incorporated multi-modal protocols and the LVT model discriminator on the

pGAN model. To further evaluate our technique, we used a variety of convolutional, attention convolutional, and transformer models. Our LVT technique outperformed state-of-the-art techniques in fusion quality on McMRI and making it a strong candidate for medical image fusion. With many potential clinical uses, such as identifying small items from satellite photos while traveling or at a specific location, LVT has great promise.

References

[1] B. B. Thukral, "Problems and preferences in pediatric imaging," *Indian Journal of Radiology and Imaging*, vol. 25, no. 4, pp. 359–364 (Oct. 2015).

[2] K. Krupa and M. Bekieśińska-Figatowska, "Artifacts in magnetic resonance imaging," *Polish Journal of Radiology,* vol. 80, pp. 93–106 (Feb. 2015).

[3] A. Adam, A. Dixon, J. Gillard, C. Schaefer-Prokop, R. Grainger, and D. Allison, Grainger & Allison's Diagnostic Radiology, Amsterdam, The Netherlands: Elsevier (2014).

[4] D. Ellison, Neuropathology: A Reference Text of CNS Pathology, Amsterdam, The Netherlands: Elsevier (2012).

[5] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Zhou, and L. Xing, "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306 (2021).

[6] Y. Luo, H. Gong, J. Liu, Y. Liu, and H. Huang, "3D transformer-GAN for high-quality PET reconstruction," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021, Cham, Switzerland: Springer, pp. 276–285 (2021).

[7] H. K. Bhuyan and V. K. Ravi, "An integrated framework with deep learning for segmentation and classification of cancer disease," *International Journal on Artificial Intelligence Tools*, vol. 32, no. 02, Article no. 2340002 (2023).

[8] Y. Korkmaz, S. U. Dar, M. Yurt, M. Özbey, and T. Çukur, "Unsupervised MRI reconstruction via zero-shot learned adversarial transformers," *IEEE Transactions on Medical Imaging,* early access, Jan. 27 (2022), doi: 10.1109/TMI.2022.3147426.

[9] H. K. Bhuyan, A. Vijayaraj, and V. K. Ravi, "Development of secrete images in image transferring system," *Multimedia Tools and Applications*, vol. 82, no. 5, pp. 7529–7552 (2023).

[10] R. S. M. Patibandla and N. Veeranjaneyulu, "A SimRank based ensemble method for resolving challenges of partition clustering methods," *Journal of Scientific & Industrial Research*, vol. 79, no. 4, pp. 323–327 (2022).

[11] A. Dureja and P. Pahwa, "Medical image retrieval for detecting pneumonia using binary classification with deep convolutional neural networks," unpublished.

[12] C. Bowles, L. Chen, R. Guerrero, P. Bentley, D. Rueckert, and A. Hammers, "Pseudo-healthy image synthesis for white matter lesion segmentation," in Simulation and Synthesis in Medical Imaging, Cham, Switzerland: Springer, pp. 87–96 (2016).

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (Jun. 2016).

[14] K. Dhar, P. Bhattacharya, N. Kumar, and A. Mitra, "Abnormality detection in chest diseases using a convolutional neural network," *Journal of Information & Optimization Sciences*, vol. 44, no. 1, pp. 97–111 (2023), doi: 10.47974/JIOS-1298.

[15] H. K. Bhuyan, A. Vijayaraj, and V. Ravi, "Diagnosis system for cancer disease using a single setting approach," *Multimedia Tools and Applications,* Springer US, pp. 1–27 (2023).

[16] A. Torrado-Carvajal, M. Sanches, D. Pascau, J. C. Desco, and J. M. Santos, "Fast patch-based pseudo-CT synthesis from T1-weighted MR images for PET/MR attenuation correction in brain studies," *Journal of Nuclear Medicine*, vol. 57, no. 1, pp. 136–143 (Jan. 2016).

[17] V. Sevetlidis, M. V. Giuffrida, and S. A. Tsaftaris, "Whole image synthesis using a deep encoder-decoder network," in Simulation and Synthesis in Medical Imaging, Cham, Switzerland: Springer, pp. 127–137 (2016).

[18] D. J. Eckman, S. G. Henderson, and S. Shashaani, "Diagnostic tools for evaluating and comparing simulation-optimization algorithms," *INFORMS Journal on Computing*, vol. 35, no. 2, pp. 350–367 (2023).

[19] H. K. Bhuyan, V. Ravi, B. Brahma, and N. K. Kamila, "Disease analysis using machine learning approaches in healthcare system," *Health and Technology*, vol. 12, no. 5, pp. 987–1005 (2022).

[20] S. U. H. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image synthesis in multi-contrast MRI with conditional generative adversarial networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2375–2388 (Oct. 2019).

[21] N. Cordier, H. Delingette, M. Le, and N. Ayache, "Extended modality propagation: Image synthesis of pathological cases," *IEEE Transactions on Medical Imaging*, vol. 35, no. 12, pp. 2598–2608 (Dec. 2016).

[22] H. K. Bhuyan, C. Chakraborty, Y. Shelke, and S. K. Pani, "COVID-19 diagnosis system by deep learning approaches," *Expert Systems*, vol. 39, no. 3, p. e12776 (2022).

[23] K. Bahrami, F. Shi, X. Zong, H. W. Shin, H. An, and D. Shen, "Reconstruction of 7T-like images from 3T MRI," *IEEE Transactions on Medical Imaging*, vol. 35, no. 9, pp. 2085–2097 (Sep. 2016).

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 27, pp. 2672–2680 (2014).

[25] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331 (Dec. 2018).

[26] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, "Hi-Net: Hybrid-fusion network for multi-modal MR image synthesis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2772–2781 (Sep. 2020).

[27] R. S. M. L. Patibandla, A. Yaswanth, and S. I. Hussani, "Water-body segmentation from remote sensing satellite images utilizing hierarchical and contour-based multi-scale features," in Mobile Radio Communications and 5G Networks, Lecture Notes in Networks and Systems, vol. 588, Singapore: Springer, pp. 245–253 (2023).

[28] H. K. Bhuyan, M. Saikiran, M. Tripathy, and V. Ravi, "Wide-ranging approach-based feature selection for classification," *Multimedia Tools and Applications*, vol. 82, no. 15, pp. 23277–23304 (Jun. 2023).

[29] A. Sharma and G. Hamarneh, "Missing MRI pulse sequence synthesis using multi-modal generative adversarial network," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1170–1183 (Apr. 2020).

[30] G. Wang, W. Li, M. A. Zuluaga, B. Dou, A. Vercauteren, and T. Fletcher, "Synthesize high-quality multi-contrast magnetic resonance imaging

from multi-echo acquisition using multi-task deep generative model," *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, pp. 3089–3099 (Oct. 2020).

[31] Y. Hiasa, Y. Otake, M. Takao, K. Matsuoka, and K. Sato, "Cross-modality image synthesis from unpaired data using CycleGAN: Effects of gradient consistency loss and training data size," in Simulation and Synthesis in Medical Imaging, Cham, Switzerland: Springer, pp. 31–41 (2018).

[32] C.-B. Jin, J.-H. Kim, D.-Y. Lee, and J.-M. Kim, "Deep CT to MR synthesis using paired and unpaired data," *Sensors,* vol. 19, no. 10, p. 2361 (May 2019).

[33] Y. Yang, K. Zhang, and Y. Fan, "sDTM: A supervised Bayesian deep topic model for text analytics," *Information Systems Research*, vol. 34, no. 1, pp. 137–156 (2022).

[34] H. K. Bhuyan and V. K. Ravi, "Analysis of sub-feature for classification in data mining," *IEEE Transactions on Engineering Management*, vol. 70, no. 8, pp. 2732–2746 (2023).

[35] H. Liang and Y. Xue, "Save face or save life: Physicians' dilemma in using clinical decision support systems," *Information Systems Research*, vol. 33, no. 2, pp. 737–758 (2023).

[36] H. K. Bhuyan, V. Ravi, and M. S. Yadav, "Multi-objective optimization-based privacy in data mining," *Cluster Computing*, vol. 25, no. 6, pp. 4275–4287 (2022).

[37] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, "SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network," *IEEE Transactions on Medical Imaging,* vol. 39, no. 7, pp. 2289–2301 (Jul. 2020).

[38] X. Zhang, Y. Wang, H. Wang, and C. Liu, "PTNet: A high-resolution infant MRI synthesizer based on transformer," arXiv preprint arXiv:2105.13993 (2021).

[39] Y. Su, D. Jia, Y. Shen, and L. Wang, "Single-channel blind image separation based on transformer-guided GAN," *Sensors,* vol. 23, no. 10, p. 4638 (2023).

[40] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multimodal medical image synthesis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614 (Oct. 2022).

[41] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2019, no. 12, p. 124020 (2019).

[42] B. C. Geiger, "On information plane analyses of neural network classifiers—A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7039–7051 (Dec. 2022).

[43] R. Shwartz-Ziv, A. Painsky, and N. Tishby, "Representation compression and generalization in deep neural networks," OpenReview (2019). [Online]. Available: https://openreview.net/forum?id=SkeL6sCqK7

[44] H. Hafez-Kolahi, S. Kasaei, and M. Soleymani-Baghshah, "Sample complexity of classification with compressed input," *Neurocomputing*, vol. 415, pp. 286–294 (2020).

[45] K. Kawaguchi, Z. Deng, X. Ji, and J. Huang, "How does information bottleneck help deep learning?," in Proc. 40th Int. Conf. Mach. Learn. (ICML), Honolulu, Hawaii, USA, PMLR 202, pp. 1–48 (2023). [Online]. Available: https://arxiv.org/abs/2305.18887

[46] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1125–1134 (2017).

[47] Z. Zhou, R. Bai, J. Munasinghe, L. Nie, and X. Chen, "T1-T2 dual-modal magnetic resonance imaging: From molecular basis to contrast agents," *ACS Nano,* vol. 11, no. 6, pp. 5227–5232 (Jun. 2017), doi: 10.1021/acsnano.7b03075.

[48] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024 (Oct. 2014).

[49] T. Nyholm, E. Nyberg, L. Karlsson, and A. Carlsson Tedgren, "MR and CT data with multiobserver delineations of organs in the pelvic area—Part of the gold atlas project," *Medical Physics,* vol. 45, no. 3, pp. 1295–1300 (Mar. 2018).

[50] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Medical Image Analysis,* vol. 5, no. 2, pp. 143–156 (Jun. 2001).

[51] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, et al., "Attention U-Net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999 (2018).

[52] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in Proc. Int. Conf. Machine Learning (ICML), vol. 97, pp. 7354–7363 (2019).

[53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 6629–6640 (2017).

[54] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce texture," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 7880–7889 (2020).

[55] L.-H. Chen, C. G. Bampis, Z. Li, C. Chen, and A. C. Bovik, "Convolutional block design for learned fractional downsampling," arXiv preprint arXiv:2105.09999 (2021).

[56] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On convergence and stability of GANs," arXiv preprint arXiv:1705.07215 (2017).

[57] Y. Luo, Y. Gao, L. Zhang, Z. Zhang, M. Jiang, and S. Hu, "Edge-preserving MRI image synthesis via adversarial network with iterative multi-scale fusion," *Neurocomputing*, vol. 452, pp. 63–77 (Sep. 2021).

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 1–11 (2017).

Received June, 2024