Conditional Generation of Medical Images via Disentangled Adversarial Inference

Mohammad Havaei*, Ximeng Mao*, Yiping Wang, Qicheng Lao

Abstract—Synthetic medical image generation has a huge potential for improving healthcare through many applications, from data augmentation for training machine learning systems to preserving patient privacy. Conditional Adversarial Generative Networks (cGANs) use a conditioning factor to generate images and have shown great success in recent years. Intuitively, the information in an image can be divided into two parts: 1) *content* which is presented through the conditioning vector and 2) *style* which is the undiscovered information missing from the conditioning vector. Current practices in using cGANs for medical image generation, only use a single variable for image generation (*i.e.*, content) and therefore, do not provide much flexibility nor control over the generated image.

In this work we propose DRAI—a dual adversarial inference framework with augmented disentanglement constraints—to learn from the image itself, disentangled representations of style and content, and use this information to impose control over the generation process. In this framework, style is learned in a fully unsupervised manner, while content is learned through both supervised learning (using the conditioning vector) and unsupervised learning (with the inference mechanism). We undergo two novel regularization steps to ensure content-style disentanglement. First, we minimize the shared information between content and style by introducing a novel application of the gradient reverse layer (GRL); second, we introduce a self-supervised regularization method to further separate information in the content and style variables.

For evaluation, we consider two types of baselines: single latent variable models that infer a single variable, and double latent variable models that infer two variables (style and content). We conduct extensive qualitative and quantitative assessments on two publicly available medical imaging datasets (LIDC and HAM10000) and test for conditional image generation, image retrieval and style-content disentanglement. We show that in general, two latent variable models achieve better performance and give more control over the generated image. We also show that our proposed model (DRAI) achieves the best disentanglement score and has the best overall performance.

I. INTRODUCTION

SUPERVISED deep neural networks have shown great success in many applications, including those in medical imaging [21, 27, 29, 36]. However, previous works have demonstrated the data hungry nature of these methods. While there are a lot of medical images being scanned inside health centers everyday, there are various factors which prohibit training of large scale models capable of achieving expert level performance. Among others, these factors include patient privacy, difficulty in collecting diverse and unbiased datasets with expert level annotations [53, 28]. Using synthetic data as a mean to circumvent the aforementioned challenges is a fascinating research venue for the medical imaging community. With the advent of newer algorithms and computational power, the desire to synthesize medical images that resemble real data is closer than ever to becoming a possibility.

Generative Adversarial Networks (GANs) [25] are generative models based on artificial neural networks which have proven successful in many applications [58, 79, 69, 35, 6, 76, 77, 49], including in the medical imaging domain [73, 34, 72, 18, 9]. They can be thought of as transformation functions that map a sample from a prior distribution (*e.g.*, the normal distribution) to a random sample from the learned data distribution ($p_{\text{model}}(\boldsymbol{x})$). However, GANs, as introduced in [25], are unconditioned generative models, and therefore, there is no control on the data being generated. Conditional generation can be very helpful as we can be more selective on which part of the learned distribution ($p_{\text{model}}(x)$)to generate data from. This is particularly interesting in the medical imaging domain, where datasets are known to have long-tail distributions with the majority of the mass around common diseases. Therefore, it's very difficult to collect real patient data from rare diseases which lie on the lower ends of the distribution tail. In this respect, conditional generative models can be used to sample the learned data manifold in areas of interest; the generator can be conditioned on some factors which we care about, *e.g.*, malignancy of a tumor, age group, ethnicity. Images generated in this way can be subsequently used for data augmentation, medical staff training, etc. Conditional GAN (cGAN) [54] is a generative adversarial network where the model is conditioned during training by additional information in order to direct the generation process. This auxiliary information could be, in theory, any type of data, such as a class label, a set of tags, a text description, or even another image.

One common pitfall of cGAN is that the conditioning codes are extremely high-level and do not cover nuances of the data. For instance, the conditioning factor could be in the form of a sparse vector representing crude class information, such as class labels. Since the class information varies significantly

M. Havaei is with Imagia, Montréal, Canada.

X. Mao is with Montréal Institute for Learning Algorithms (MILA), Université de Montréal, Montréal, Canada. This work was carried out while the author was interning at Imagia.

Y. Wang is with Imagia, Montréal, Canada; and University of Victoria, Victoria, Canada.

Q. Lao is with Imagia, Montréal, Canada; Montréal Institute for Learning Algorithms (MILA), Université de Montréal, Montréal, Canada; and West China Biomedical Big Data Center, West China Hospital of Sichuan University, Chengdu, China.

This research was carried out in Imagia.

* Authors contributed equally.

Correspondence to M. Havaei (e-mail:mohammad@imagia.com) and Q. Lao (e-mail:qicheng.lao@gmail.com).

within each class, the class label alone does not provide the freedom to control the individual factors of variation governing each class. This challenge is exemplified in the medical imaging domain where insufficient label granularity is a common occurrence. We refer to the factors of variation that depend on the conditioning vector as *content*.

Another challenge in conditional image generation is that the image distribution also contains factors of variation that are agnostic to the conditioning code. These types of information are shared among different classes or different conditioning codes. In this work we refer to such information as *style*, which depending on the task, could correspond to position, orientation, location, background information, etc.

Having distinct control over both attributes of content and style is very appealing when generating medical images. For example, for privacy concerns, we may want to refrain from using real patient images and rely on synthetically generated images instead. In such cases, having full control over the image generation is crucial for preserving patient privacy, *i.e.*, removing patient information (style) while preserving disease information (content). Learning disentangled representation of content and style allow us to control the detailed nuances of the generation process.

Another venue where inferring disentangled content and style can be of interest is unsupervised style or content based image retrieval, where we want the retrieved image to respect the style or the content of the query image. In this case, if there is no disentanglement of content and style, the shared information between the two variables, would result in an undesirable outcome. (see Section III-E3 for experiments on this application).

In this work, we consider two types of information to preside over the image domain: $i)$ *content*, which refers to the information in the conditioning vector for image generation and $ii)$ *style*, which encompasses any information not covered by the conditioning vector. By definition, these two types of information are independent from one another and this independence criteria should be taken into account when training a model. By explicitly constraining the model to disentangle content and style, we ensure their independence and prevent information leakage between them. To achieve this goal, we introduce Dual Regularized Adversarial Inference (DRAI), a conditional generative model that leverages unsupervised learning and novel disentanglement constraints to learn disentangled representations of content and style, which in turn enables more control over the generation process.

Since there is no supervision on the style, we use an adversarial inference mechanism and learn to infer style in an unsupervised way from the image. For the content information however, we have access to both the conditioning vector and the image. The content can thus be learned both in a supervised way through the conditioning vector and unsupervised way from the image itself.

We impose two novel disentanglement constraints to facilitate the separation of content and style: Firstly, we introduce a novel application of the *Gradient Reverse Layer* (GRL) [22] to minimize the shared information between the two variables. Secondly, we present a new type of self-supervised regu-

larization to further enforce disentanglement; using content-preserving transformations, we *attract* matching content information, while *repelling* different style information.

An important feature of our model is that, in contrary to most conditional generation methods that require the conditioning vector at test time, our model has the flexibility to use either a conditioning vector or a content code inferred from a reference image. In addition, we also allow generating hybrid images by mixing the inferred style and content codes from multiple sources (see more details in Section II-F).

We compare the proposed method with multiple baselines on two datasets. We show the advantage of using two latent variables to represent style and content for conditional image generation. To quantify style-content disentanglement, we introduce a disentanglement measure and show the proposed regularizations can improve the separation of style and content information. We also demonstrate the use case of DRAI in the style or content based image retrieval.

The contributions of this work can be summarized as follows:

- To the best of our knowledge, this is the first time disentanglement of content and style has been explored in the context of medical image generation.
- We introduce a novel application of GRL that penalizes shared information between content and style in order to achieve better disentanglement.
- We introduce a self-supervised regularization that encourages the model to learn independent information as content and style.
- we introduce a quantitative content-style disentanglement measure that does not require any content or style labels. This is especially useful in real world scenarios where attributes contributing to content and style are not available.

II. METHOD

Our proposed framework, DRAI, has two main components: a conditional image generation module and an inference module. In what follows, we explain each component individually. Note that the two modules are not independent since training is end-to-end. To ensure disentanglement between the two inferred variables, *i.e.*, style and content, we impose disentanglement constraints which are also introduced in this section.

A. Overview

We start by describing the objectives of this paper. Let t be the conditioning vector associated with image x, where t could be a text description, class identity, meta data or any other piece of information for which our image generation is based on. Using the pairs $\{(t_i, x_i)\}, i = 1, \ldots, N$, where N denotes the size of the dataset, we train an inference model $G_{c,z}$ and a generative model G_x such that (i) the inference model $G_{c,z}$ infers content c and style z in a way that they are disentangled from each other and (ii) the generator G_x can generate realistic images that not only visually respect the conditioning vector t but also the style/content disentanglement.
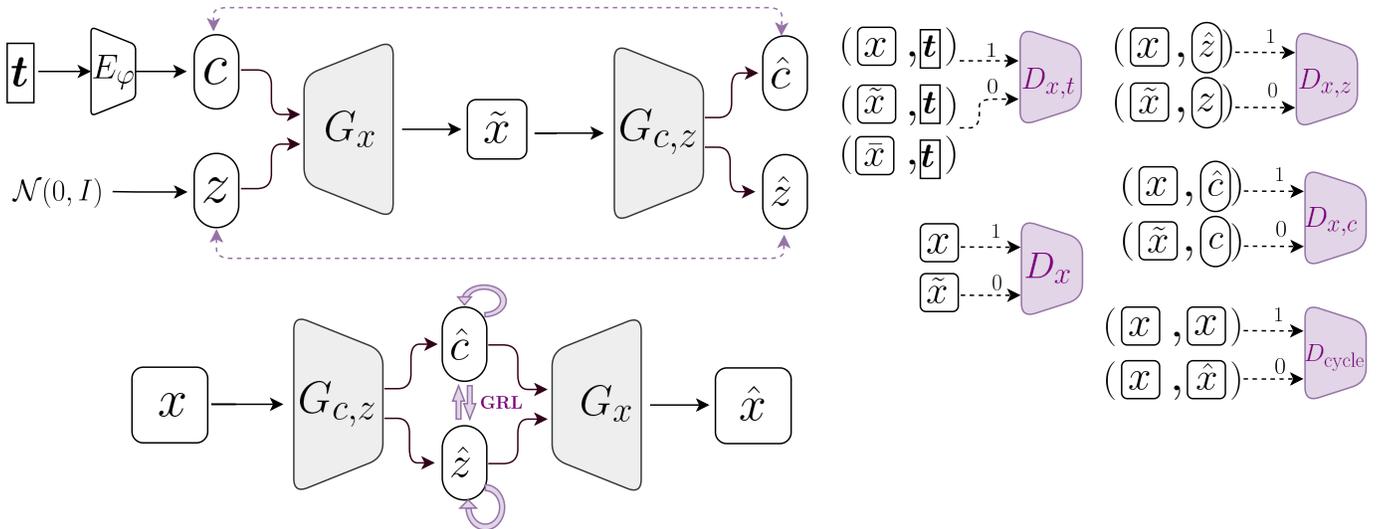
Fig. 1: Overview of DRAI. The encoder E_φ maps the conditioning vector t to a higher dimensional content space represented by the random variable c with distribution $q(c)$. The generator G_x takes two inputs: a sample from the content random variable c, and a sample from the style random variable z to generate $\tilde{x}$. The dashed purple arrows mark the cycle consistency between features implemented via ℓ_1 norm, while the solid purple arrows show the imposed disentanglement constrains. On the right hand side of the figure we show all the discriminators used for training. $\hat{c}$ represents the inferred content, $\hat{z}$ the inferred style, $\hat{x}$ the reconstructed input image and $\bar{x}$ the image with mismatched conditioning.

This framework allows for great flexibility in image generation. On one hand, conditional image generation is possible by conditioning on vector t to generate x. In such cases, the content information is conveyed through the conditioning vector t. On the other hand, if t is not available, we can use the inference module to infer style and content from a reference image. The inferred codes can then be used to generate images that resemble the content and/or style of the reference image. This is in contrast to prior works which operate on the strong assumption that the conditioning vector t is always available. It is worth noting that our generative module is *not* constrained to require a style image. Having a probabilistic generative model allows us to sample the style code from the style prior distribution and generate images with random style attributes.

The framework also allows us to generate hybrid images by mixing style and content from various sources (details can be found in Section II-F).

In the remaining of this section, we will describe each component of the proposed DRAI model. An illustration of DRAI is made in Figure 1.

B. Conditional image generation

We follow the formulation of Lao et al. [44]. Our image generation module takes two vectors as input: a noise vector z (representing style) sampled from the style prior distribution and a vector c (representing content) which is an encoded representation of the conditioning vector t. In order to learn richer representations for the content, we sample c from a Gaussian distribution $q_\varphi(c|t) = \mathcal{N}(\mu_t, \Sigma_t)$, where μ and Σ are functions of t parameterized by a neural network with parameters φ and are trained end-to-end with the rest of the network. To simplify the notations, we denote $q_\varphi(c|t)$ as $q(c)$ throughout the paper.

In this work, we use adversarial training [25] in multiple fronts for which we will explain in the following. Adversarial training was introduced by [25] where a discriminator network is trained to distinguish between real data examples (positive class) and fake examples (negative class) generated by the generator. The generator on the other hand is optimized towards fooling the discriminator. Through the adversarial game between the generator and discriminator, the distribution of the generated fake examples moves towards the distribution of real data, resulting in the generator producing realistic images. Adversarial training provides powerful implicit loss functions and has shown to be very powerful in matching complex distributions. In order to improve the alignment between conditioning vector t and the generated image $\tilde{x}$, we seek to match $p(\tilde{x}, t)$ with $p(x, t)$. To do so, we adopt the matching-aware discriminator proposed by [60]. For this discriminator—denoted as $D_{x,t}$—the positive sample is the pair of real image and its corresponding conditioning vector (x, t), whereas the negative sample pairs consist of two groups; the pair of real image with mismatched conditioning $(\bar{x}, t)$, and the pair of synthetic image with corresponding conditioning $(G_x(z, c), t)$. Borrowing intuition from [10], we can show that $D_{x,t}$ measures the mutual information between t and x and assigns higher scores to (x, t) pairs with higher mutual information. In order to retain the fidelity of the generated images, we also train a discriminator D_x that distinguishes between real and generated images. The loss

function imposed by $D_{x,t}$ and D_x is as follows:

$$\min_G \max_D V_{\text{t2i}}(D_x, D_{x,t}, G_x) =$$
$$\mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}}[\log D_x(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z}), \boldsymbol{c} \sim q(\boldsymbol{c})}[\log(1 - D_x(G_x(\boldsymbol{z}, \boldsymbol{c})))] +$$
$$\mathbb{E}_{(\boldsymbol{x},t) \sim p_{\text{data}}}[\log D_{x,t}(\boldsymbol{x}, t)] +$$
$$\frac{1}{2}\Big\{ \mathbb{E}_{(\bar{\boldsymbol{x}},t) \sim p_{\text{data}}}[\log(1 - D_{x,t}(\bar{\boldsymbol{x}}, t))] +$$
$$\mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z}), \boldsymbol{c} \sim q(\boldsymbol{c}), t \sim p_{\text{data}}}[\log(1 - D_{x,t}(G_x(\boldsymbol{z}, \boldsymbol{c}), t))] \Big\}, \quad (1)$$

where $\tilde{\boldsymbol{x}} = G_x(\boldsymbol{z}, \boldsymbol{c})$ is the generated image and $(\bar{\boldsymbol{x}}, t)$ designates a mis-matched pair.

C. Adversarial Inference

Latent variable models provide an efficient way to perform approximate inference in order to discover factors of variations governing the data distribution. This allows the model to reason about the data on an abstract level. While data generation is carried out through mapping the latent space $\boldsymbol{z}$ to the data space $\boldsymbol{x}$, an inference mechanism learns the inverse mapping function from $\boldsymbol{x}$ to $\boldsymbol{z}$. Bidirectional GAN *a.k.a* Adversarially Learned Inference (ALI) [19, 20] is a GAN based latent variable model that performs approximate inference by training a bidirectional discriminator to distinguish between two joint distributions: real data sample and its inferred latent code $(\boldsymbol{x}, \hat{\boldsymbol{z}})$, and real latent code and its generated data sample $(\boldsymbol{z}, \tilde{\boldsymbol{x}})$.

1) Single variable adversarial inference: For a single latent variable model, let $q(\boldsymbol{x})$ represent the empirical data distribution and $p(\boldsymbol{z})$ the marginal distribution of the latent variable, specified as a simple random distribution, *e.g.*, the standard normal distribution $\mathcal{N}(\boldsymbol{0}, \mathbf{I})$. ALI aims to match the two joint distributions $q(\boldsymbol{x}, \boldsymbol{z}) = q(\boldsymbol{z}|\boldsymbol{x})q(\boldsymbol{x})$ and $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$, which in turn implies that $q(\boldsymbol{z}|\boldsymbol{x})$ matches $p(\boldsymbol{z}|\boldsymbol{x})$. To achieve this, an encoder $G_z(\boldsymbol{x}) : \hat{\boldsymbol{z}} = G_z(\boldsymbol{x}), \boldsymbol{x} \sim q(\boldsymbol{x})$ is introduced in the generation phase, in addition to the standard generator $G_x(\boldsymbol{z}) : \tilde{\boldsymbol{x}} = G_x(\boldsymbol{z}), \boldsymbol{z} \sim p(\boldsymbol{z})$. The discriminator D is trained to distinguish between the joint pairs $(\boldsymbol{x}, \hat{\boldsymbol{z}})$ and $(\tilde{\boldsymbol{x}}, \boldsymbol{z})$. The minimax objective of adversarial inference can be written as:

$$\min_G \max_D V(D, G_x, G_z) =$$
$$\mathbb{E}_{\boldsymbol{x} \sim q(\boldsymbol{x}), \hat{\boldsymbol{z}} \sim q(\boldsymbol{z}|\boldsymbol{x})}[\log D(\boldsymbol{x}, \hat{\boldsymbol{z}})] +$$
$$\mathbb{E}_{\tilde{\boldsymbol{x}} \sim p(\boldsymbol{x}|\boldsymbol{z}), \boldsymbol{z} \sim p(\boldsymbol{z})}[\log(1 - D(\tilde{\boldsymbol{x}}, \boldsymbol{z}))]. \quad (2)$$

2) Double variable adversarial inference: ALI and its variants encode all the information in a single latent variable. Following [44], we augment this framework to support two independent variables representing style ($\boldsymbol{z}$) and content ($\boldsymbol{c}$) which allows us to encode disjoint information in each variable and ultimately disentangle style and content information. In this augmented framework, given a sample $\boldsymbol{x}$ from empirical data distribution $q(\boldsymbol{x})$, the posterior distribution over style and content is formulated as $q(\boldsymbol{z}, \boldsymbol{c}|\boldsymbol{x})$. Using the adversarial inference framework, we are interested in matching the conditional $q(\boldsymbol{z}, \boldsymbol{c}|\boldsymbol{x})$ to the posterior $p(\boldsymbol{z}, \boldsymbol{c}|\boldsymbol{x})$. Given the Independence assumption of $\boldsymbol{c}$ and $\boldsymbol{z}$, we have the following factorisation:

$$q(\boldsymbol{z}, \boldsymbol{c} \mid \boldsymbol{x}) = q(\boldsymbol{z} \mid \boldsymbol{x})q(\boldsymbol{c} \mid \boldsymbol{x}), \quad (3)$$
$$p(\boldsymbol{z}, \boldsymbol{c} \mid \boldsymbol{x}) = p(\boldsymbol{z} \mid \boldsymbol{x})p(\boldsymbol{c} \mid \boldsymbol{x}). \quad (4)$$

This formulation allows us to match $q(\boldsymbol{z}|\boldsymbol{x})$ to $p(\boldsymbol{z}|\boldsymbol{x})$ and $q(\boldsymbol{c}|\boldsymbol{x})$ to $p(\boldsymbol{c}|\boldsymbol{x})$, respectively. We achieve this by matching the two pairs of joint distributions:

$$q(\boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{z}, \boldsymbol{x}), \quad (5)$$
$$q(\boldsymbol{c}, \boldsymbol{x}) = p(\boldsymbol{c}, \boldsymbol{x}). \quad (6)$$

In the dual adversarial inference process, the feature generator $G_{c,z}$, encodes the image to infer style and content; $\hat{\boldsymbol{z}}, \hat{\boldsymbol{c}} = G_{c,z}(\boldsymbol{x}), \boldsymbol{x} \sim q(\boldsymbol{x})$, while the image generator G_x decodes samples from the style and content prior distributions into an image; $\tilde{\boldsymbol{x}} = G_x(\boldsymbol{z}, \boldsymbol{c}), \boldsymbol{z} \sim p(\boldsymbol{z}), \boldsymbol{c} \sim p(\boldsymbol{c})$. To compete with the generators (*i.e.*, G_x and $G_{c,z}$), we train two discriminators: $D_{x,z}$ to discriminate between the pairs $(\boldsymbol{x}, \hat{\boldsymbol{z}})$ and $(\tilde{\boldsymbol{x}}, \boldsymbol{z})$ (sampled from $q(\boldsymbol{x}, \boldsymbol{z})$ and $p(\boldsymbol{x}, \boldsymbol{z})$), and $D_{x,c}$ to discriminate between the pairs $(\boldsymbol{x}, \hat{\boldsymbol{c}})$ and $(\tilde{\boldsymbol{x}}, \boldsymbol{c})$ (sampled from $q(\boldsymbol{x}, \boldsymbol{c})$ and $p(\boldsymbol{x}, \boldsymbol{c})$). The dual adversarial inference objective can be thus framed as:

$$\min_G \max_D V_{\text{dALI}}(D_{x,z}, D_{x,c}, G_x, G_{c,z}) =$$
$$\mathbb{E}_{\boldsymbol{x} \sim q(\boldsymbol{x}), \hat{\boldsymbol{z}}, \hat{\boldsymbol{c}} \sim q(\boldsymbol{z}, \boldsymbol{c}|\boldsymbol{x})}[\log D_{x,z}(\boldsymbol{x}, \hat{\boldsymbol{z}}) + \log D_{x,c}(\boldsymbol{x}, \hat{\boldsymbol{c}})] +$$
$$\mathbb{E}_{\tilde{\boldsymbol{x}} \sim p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c}), \boldsymbol{z} \sim p(\boldsymbol{z}), \boldsymbol{c} \sim p(\boldsymbol{c})}[\log(1 - D_{x,z}(\tilde{\boldsymbol{x}}, \boldsymbol{z})) + \log(1 - D_{x,c}(\tilde{\boldsymbol{x}}, \boldsymbol{c}))]. \quad (7)$$

3) Image cycle-consistency: Matching the two joint distributions alone as done in ALI, does not identify the relationship between the latent codes (*i.e.*, $\boldsymbol{z}$ and $\boldsymbol{c}$) and the data (*i.e.*, $\boldsymbol{x}$). This results in ALI generating realistic looking images, but having poor reconstructions. Li et al. [47] address this issue and describe the non-identifiability problem of ALI in the single latent variable setup. To impose correlation between the latent code $\boldsymbol{z}$ and image $\boldsymbol{x}$, they propose to incorporate a loss function which enforces cycle-consistency between data sample $\boldsymbol{x}$ and the generated image from the inferred code $\hat{\boldsymbol{z}}$. More specifically, a discriminator D_{cycle} is trained to distinguish between $\boldsymbol{x}$ and its reconstruction $\hat{\boldsymbol{x}}$. They show that in addition to achieving better reconstructions, using cycle consistency stabilizes the training of ALI. The cycle-consistency loss was introduced in [79] for unpaired image to image translation using the CycleGAN architecture and has since then been used in many different applications [3, 74, 40].

In this work, we adopt cycle-consistency in a similar way as [47]. We train a discriminator D_{cycle} to distinguish between pairs $(\boldsymbol{x}, \boldsymbol{x})$ and $(\boldsymbol{x}, \hat{\boldsymbol{x}})$ with $\hat{\boldsymbol{x}}$ being the reconstruction for $\boldsymbol{x}$; $\hat{\boldsymbol{x}} = G_x(\hat{\boldsymbol{z}}, \hat{\boldsymbol{c}})$, where $\hat{\boldsymbol{z}}, \hat{\boldsymbol{c}} = G_{c,z}(\boldsymbol{x})$. The cycle-consistency loss is denoted by V_{cycle} as follows:

$$\min_G \max_D V_{\text{image-cycle}}(D_{cycle}, G_x, G_{c,z}) =$$
$$\mathbb{E}_{\boldsymbol{x} \sim q(\boldsymbol{x})}[\log D_{cycle}(\boldsymbol{x}, \boldsymbol{x})] +$$
$$\mathbb{E}_{\boldsymbol{x} \sim q(\boldsymbol{x}), (\hat{\boldsymbol{z}}, \hat{\boldsymbol{c}}) \sim q(\boldsymbol{z}, \boldsymbol{c}|\boldsymbol{x})}[\log(1 - D_{cycle}(\boldsymbol{x}, G_x(\hat{\boldsymbol{z}}, \hat{\boldsymbol{c}})))]. \quad (8)$$

We also experimented with the ℓ_1 loss as the objective function for cycle-consistency. In practice, the adversarial cycle consistency was slightly better.

4) Latent code cycle-consistency: To further preserve information from a pair of style-content codes in the generated image, we infer the latent code from the generated image and

apply cycle consistency between the inferred and the original codes.

$$\min_{G} V_{\text{code-cycle}}(G_x, G_{c,z}) =$$
$$\mathbb{E}_{(\boldsymbol{z}',\boldsymbol{c}')\sim q(\boldsymbol{z},\boldsymbol{c}|\tilde{\boldsymbol{x}}),\boldsymbol{z}\sim p(\boldsymbol{z}),\boldsymbol{c}\sim q(\boldsymbol{c})}[\|\boldsymbol{z}'-\boldsymbol{z}\|+\|\boldsymbol{c}'-\boldsymbol{c}\|], \quad (9)$$

where $\tilde{\boldsymbol{x}} \sim p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})$.

We can show that Equation (9) maximizes the mutual information between the content variable and the generated image as well as the mutual information between the style variable and the generated image. Let $I(\boldsymbol{c}; G_x(\boldsymbol{c},\boldsymbol{z}))$ denote the mutual information between the content variable and the generated image, where

$$I(\boldsymbol{c}; G_x(\boldsymbol{c},\boldsymbol{z})) = H(\boldsymbol{c}) - H(\boldsymbol{c}|\tilde{\boldsymbol{x}}). \quad (10)$$

Following [2], we define a variational lower bound on $I(\boldsymbol{c}; \tilde{\boldsymbol{x}})$ by rewriting the conditional entropy in Equation (10) as:

$$-H(\boldsymbol{c}|\tilde{\boldsymbol{x}}) = \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})}[\log q(\boldsymbol{c}|\tilde{\boldsymbol{x}}) + D_{KL}(p(\boldsymbol{c}|\tilde{\boldsymbol{x}})\|q(\boldsymbol{c}|\tilde{\boldsymbol{x}}))],$$

and by extension:

$$I(\boldsymbol{c};\tilde{\boldsymbol{x}}) = H(\boldsymbol{c}) + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})}[\log q(\boldsymbol{c}|\tilde{\boldsymbol{x}}) + D_{KL}(q(\boldsymbol{c}|\tilde{\boldsymbol{x}})\|p(\boldsymbol{c}|\tilde{\boldsymbol{x}}))].$$

Using the non-negativity of $H(\boldsymbol{c})$ and $D_{KL}(p(\boldsymbol{c}|\tilde{\boldsymbol{x}})\|q(\boldsymbol{c}|\tilde{\boldsymbol{x}})))$, we obtain the following lower bound on the mutual information:

$$I(\boldsymbol{c};\tilde{\boldsymbol{x}}) \geq \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})}[\log q(\boldsymbol{c}|\tilde{\boldsymbol{x}})].$$

Similar derivation can be made to show

$$I(\boldsymbol{z};\tilde{\boldsymbol{x}}) \geq \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})}[\log q(\boldsymbol{z}|\tilde{\boldsymbol{x}})].$$

The objective function in Equation (9), maximizes the log likelihoods $\mathbb{E}_{\tilde{\boldsymbol{x}}\sim p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})}[\log q(\boldsymbol{c}|\tilde{\boldsymbol{x}})]$ and $\mathbb{E}_{\tilde{\boldsymbol{x}}\sim p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})}[\log q(\boldsymbol{z}|\tilde{\boldsymbol{x}})]$ and by extension the variational lower bound on the terms $I(\boldsymbol{c};\tilde{\boldsymbol{x}})$ and $I(\boldsymbol{z};\tilde{\boldsymbol{x}})$.

D. Disentanglement constraints

Lao et al. [44] use double variable ALI as a criterion for disentanglement. However, ALI does approximate inference and does not necessarily guarantee disentanglement between variables. To further impose disentanglement between style and content, we propose additional constrains and regularization measures.

1) Content-Style information minimization: The content should not include any information of the style and vice versa. We seek to *explicitly* minimize the shared information between style and content. For this, we propose a novel application of the Gradient Reversal Layer (GRL) strategy. First introduced in [22], the GRL strategy is used in domain adaptation methods to learn domain-agnostic features, where it acts as the identity function in the forward pass but reverses the direction of the gradients in the backward pass. In domain adaptation literature, GRL is used with a domain classifier. Reversing the direction of the gradients coming from the domain classification loss has the effect of minimizing the information between the representations and domain identity, thus, learning domain invariant features. Inspired by the literature on domain adaptation, we use GRL to minimize the information between

style and content. More concretely, for a given example $\boldsymbol{x}$, we train an encoder F_c to predict the content from style and use GRL to minimize the information between the two. The same process is done for predicting style from content through F_z, resulting in the following objective function:

$$\min_{G} \max_{F} V_{\text{GRL}}(F_z, F_c, G_{c,z}) =$$
$$-\mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x}),(\hat{\boldsymbol{z}},\hat{\boldsymbol{c}})\sim q(\boldsymbol{z},\boldsymbol{c}|\boldsymbol{x})}[\|\hat{\boldsymbol{z}} - F_z(\hat{\boldsymbol{c}})\| + \|\hat{\boldsymbol{c}} - F_c(\hat{\boldsymbol{z}})\|]. \quad (11)$$

This constrains the content feature generation to disregard style features and the style feature generation to disregard content features. Figure 2 shows a visualization of this module.

We can show that Equation (11) minimizes a lower bound on the mutual information between the style variable and the content variable. Here, we only provide the proof for using GRL with F_z to predict style from content. Similar reasoning can be made for using GRL with F_c. Let $I(\boldsymbol{z};\boldsymbol{c})$ denote the mutual information between the inferred content and the style variables, where

$$I(\boldsymbol{z};\boldsymbol{c}) = H(\boldsymbol{z}) - H(\boldsymbol{z}|\boldsymbol{c}). \quad (12)$$

Once again, following [2], we define a variational lower bound on $I(\boldsymbol{z};\boldsymbol{c})$ by rewriting the conditional entropy in (12) as:

$$-H(\boldsymbol{z}|\boldsymbol{c}) = \mathbb{E}_{\hat{\boldsymbol{c}}\sim q(\boldsymbol{c}|\boldsymbol{x})}[\log q(\boldsymbol{z}|\hat{\boldsymbol{c}}) + D_{KL}(p(\boldsymbol{z}|\hat{\boldsymbol{c}})\|q(\boldsymbol{z}|\hat{\boldsymbol{c}}))]],$$

and by extension:

$$I(\boldsymbol{z};\boldsymbol{c}) = H(\boldsymbol{z}) + \max_{F_z}\mathbb{E}_{\hat{\boldsymbol{c}}\sim q(\boldsymbol{c}|\boldsymbol{x})}[\log q(\boldsymbol{z}|\hat{\boldsymbol{c}})], \quad (13)$$

where the maximum is achieved when

$$D_{KL}(p(\boldsymbol{z}|\hat{\boldsymbol{c}})\|q(\boldsymbol{z}|\hat{\boldsymbol{c}})) = 0.$$

Since $H(\boldsymbol{z})$ is non-negative and $\|\hat{\boldsymbol{z}} - F_z(\hat{\boldsymbol{c}})\|$ corresponds to $-\log q(\boldsymbol{z}|\hat{\boldsymbol{c}})$, minimization of lower bound on mutual information can be written as:

$$\min_{G} \max_{F_z} -\mathbb{E}_{\hat{\boldsymbol{c}}\sim q(\boldsymbol{c}|\boldsymbol{x}),\hat{\boldsymbol{z}}\sim q(\boldsymbol{z}|\boldsymbol{x})}[\|\hat{\boldsymbol{z}} - F_z(\hat{\boldsymbol{c}})\|], \quad (14)$$

which corresponds to Equation (11).

2) Self-supervised regularization: Self-supervised learning has shown great potential in unsupervised representation learning [57, 30, 15]. To provide more control over the latent variables $\boldsymbol{c}$ and $\boldsymbol{z}$, we incorporate a self-supervised regularization such that the content is invariant to content-preserving transformations while the style is sensitive to such transformations. The proposed self-supervised regularization constraints the feature generator $G_{c,z}$ to encode different information for content and style. More formally, let $\mathcal{T}$ be a random content-preserving transformation such as a rotation, horizontal or vertical flip. For every example $\boldsymbol{x} \sim q(\boldsymbol{x})$, let $\boldsymbol{x}'$ be its transformed version; $\boldsymbol{x}' = T_i(\boldsymbol{x})$ for $T_i \sim p(\mathcal{T})$. We would like to maximize the similarity between the inferred contents of $\boldsymbol{x}$ and $\boldsymbol{x}'$ and minimize the similarity between their inferred styles. This constrains the content feature generation to focus on the content of the image reflected in the conditioning vector and the style feature generation to focus on other attributes. This regularization procedure is visualized in Figure 3. The
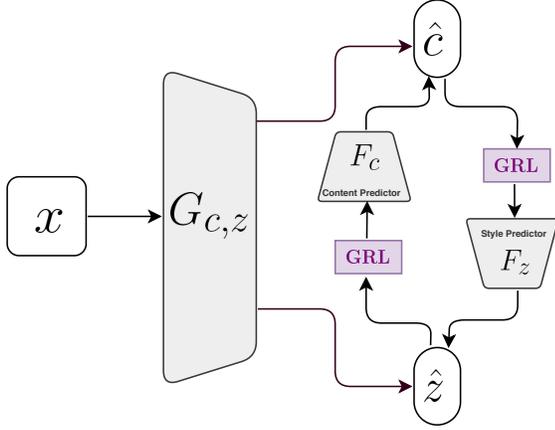
Fig. 2: Content-Style information minimization. For a given image x, F_c is trained to predict the content $\hat{c}$ from the style $\hat{z}$. By reversing the direction of the gradients, the GRL penalizes $G_{c,z}$ to minimize the content information in the style variable z. The same procedure is carried out to minimize style information in the content variable c.
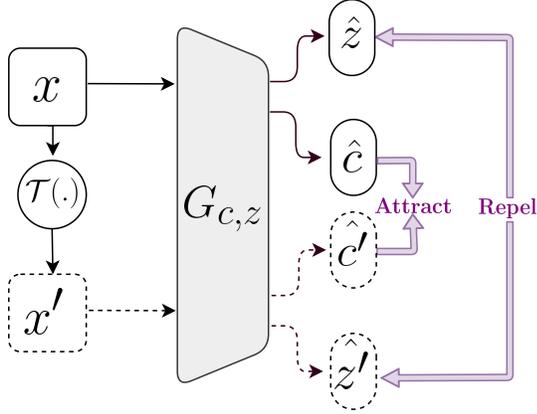

Fig. 3: Self-Supervised regularization. Given x and its transformed version x', their corresponding content codes c and $\hat{c}$ form a positive pair and the disparity between them is minimized (*i.e.*, attract each other) while their corresponding style codes z and $\hat{z}$ form a negative pair and the disparity between them is maximized (*i.e.*, repel each other).

objective function for the self-supervised regularization is defined as:

$$\min_G V_{\text{self}}(G_{c,z}) = \mathbb{E}_{x \sim q(x)}[\|\hat{c} - \hat{c}'\| - \|\hat{z} - \hat{z}'\|], \quad (15)$$

where $(\hat{z}, \hat{c}) \sim q(z, c|x)$ and $(\hat{z}', \hat{c}') \sim q(z, c|x')$.

E. Full Objective

DRAI is a probabilistic model which requires reparameterization trick to sample from the approximate posteriors $q(z|x)$, $q(c|x)$ and $q(c|t)$. We use KL divergence in order to regularize these posteriors to follow the normal distribution $\mathcal{N}(0, I)$. Taking that into account, the complete objective criterion for
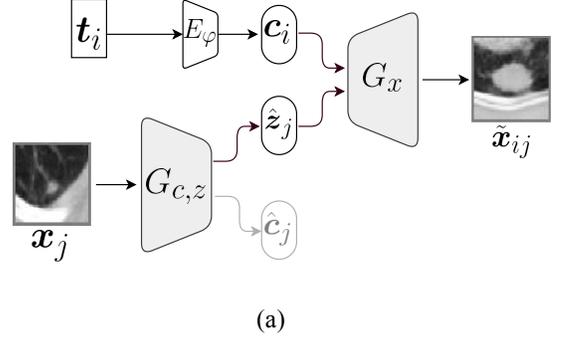
(a)
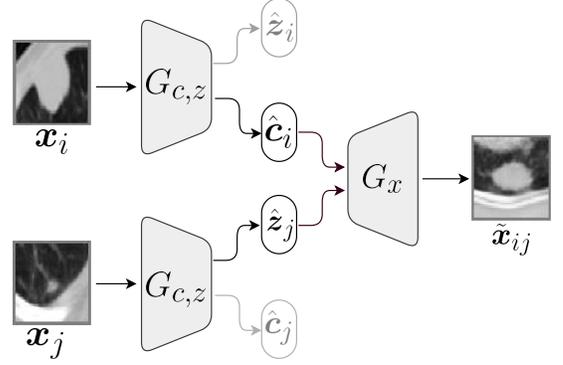
(b)

Fig. 4: Hybrid image generation: (a) via the conditioning factor t_i (representing content) and the style code $\hat{z}_j$ inferrnd from the style reference image. (b) via the content code $\hat{c}_i$ inferred from the content reference image and the style code $\hat{z}_j$ inferred from the style reference image.

DRAI is:

$$
\begin{aligned}
\min_G \max_{D,F} V_{\text{t2i}}(D_x, D_{x,t}, G_x) \ + \\
V_{\text{dALI}}(D_{x,z}, D_{x,c}, G_x, G_{c,z}) \ + \\
V_{\text{image-cycle}}(D_{cycle}, G_x, G_{c,z}) \ + \\
V_{\text{code-cycle}}(G_x, G_{c,z}) \ + \\
V_{\text{GRL}}(F_z, F_c, G_{c,z}) \ + \\
V_{\text{self}}(G_{c,z}) \ + \\
\lambda D_{KL}(q(z|x) \| \mathcal{N}(0, I)) \ + \\
\lambda D_{KL}(q(c|x) \| \mathcal{N}(0, I)) \ + \\
\lambda D_{KL}(q(c|t) \| \mathcal{N}(0, I)).
\end{aligned}
\quad (16)
$$

Figure 1 illustrates various discriminators and objective functions used to train DRAI.

F. Generating Hybrid Images

Thanks to our encoder that is able to infer disentangled codes for style and content and also our generator that does not have a hard constraint on requiring the conditioning embedding t, we can generate hybrid images where we mix style and content from different image sources. Let i and j be the indices of two different images. There are two ways in which DRAI can generate hybrid images:

1) Using a conditioning vector t_i and a style image x_j: In this setup, we use the conditioning factor t_i as the

content and the inferred $\hat{z}_j$ from the style image x_j as the style:

$$c_i = E_\varphi(t_i)$$
$$\hat{z}_j, \hat{c}_j = G_{c,z}(x_j)$$
$$\tilde{x}_{ij} = G_x(\hat{z}_j, c_i).$$

2) Using a content image x_i and a style image x_j: In this setup we do not rely on the conditioning factor t. Instead, we infer codes for both style and content (*i.e.*, $\hat{z}_j$ and $\hat{c}_i$) from style and content source images respectively.

$$\hat{z}_i, \hat{c}_i = G_{c,z}(x_i)$$
$$\hat{z}_j, \hat{c}_j = G_{c,z}(x_j)$$
$$\tilde{x}_{ij} = G_x(\hat{z}_j, \hat{c}_i)$$

The generation of hybrid images is graphically explained in Figure 4 for the aforementioned two scenarios.

III. EXPERIMENTS

A. Datasets

We conduct experiments on two publicly available medical imaging datasets.

1) HAM10000: Human Against Machine (HAM10000) [66], contains approximately 10000 training images, includes 10015 dermatoscopic images of seven types of skin lesions and is widely used as a classification benchmark. One of the lesion types, "Melanocytic nevi" (nv), occupies around 67% of the whole dataset, while the two lesion types that have the smallest data size, namely, "Dermatofibroma" (df) and "Vascular skin lesions" (vasc), have only 115 and 143 images respectively. Such data imbalance is undesirable for our purpose since limitations on the data size lead to severe lack of image diversity of the minority classes. For our experiments, we select the three largest skin lesion types, which in order of decreasing size are: "nv" with 6705 images; "Melanoma" (mel) with 1113 images; and "Benign keratosis-like lesions" (bkl) with 1099 images. Patches of size 48×48 centered around the lesion are extracted and then resized to 64×64. To balance the dataset, we augment mel and bkl three times with random flipping. We follow the train-test split provided by the dataset, and the data augmentation is done only on the training data.

2) LIDC: The Lung Image Database Consortium image collection (LIDC-IDRI) consists of lung CT scans from 1018 clinical cases [5]. In total, 7371 lesions are annotated by one to four radiologists, of which 2669 are given ratings on nine nodule characteristics: "malignancy", "calcification", "lobulation", "margin", "spiculation", "sphericity", "subtlety", "texture" and "internal structure". We take the following pre-processing steps for LIDC: *a)* We normalize the data such that it respects the Hounsfield units (HU), *b)* the volume size is converted to $256 \times 256 \times 256$, *c)* areas around the lungs are cropped out. For our experiments, we extract a subset of 2D patches composing nodules with consensus from at least three radiologists. Patches of size 48×48 centered around the nodule are extracted and then resized to

64×64. Furthermore, we compute the inter-observer median of the malignancy ratings and exclude those with malignancy median of 3 (out of 5). This is to ensure a clear separation between benign and malignant classes presented in the dataset. The conditioning factor for each nodule is a 17-dimensional vector, coming from six of its characteristic ratings, as well as the nodule size. Note that "lobulation" and "spiculation" are removed due to known annotation inconsistency in their ratings [4], and "internal structure" is removed since it has a very imbalanced distribution. We quantize the remaining characteristics to binary values following the same procedure of Shen et al. [63] and use the one-hot encoding to generate a 12-dimensional vector for each nodule. The remaining five dimensions are reserved for the quantization of the nodule size, ranging from 2 to 12 with an interval of 2. Following the above described procedure, the nodules with case index less than 899 are included in the training dataset while the nodules of the remaining cases are considered as the test set. By augmenting the label in such way, we exploit the richness of each nodule in LIDC, which proves to be beneficial for training.

B. Baselines

To evaluate the quality of generation, inference, and disentanglement, we consider two types of baselines. To show the effectiveness of dual variable inference, we compare our framework with single latent variable models. For this, we introduce a conditional adaptation of InfoGan [16] referred to as cInfoGAN and a conditional adversarial variational Autoencoder (cAVAE), both of which are explained in this section.

To compare our approach to dual latent variable inference methods, we extend InfoGAN and cAVAE to dual variables which we denote as D-cInfoGAN and D-cAVAE respectively.

We also compare DRAI to Dual Adversarial Inference (DAI) [44] and show how using our proposed disentanglement constraints together with latent code cycle-consistency can significantly boost performance. Finally, we conduct rigorous ablation studies to evaluate the impact of each component in DRAI.

1) conditional InfoGAN: InfoGAN is a variant of generative adversarial network that aims to learn unsupervised disentangled representations. In order to do so, InfoGAN modifies the original GAN in two ways. First, it adds an additional input c to the generator. Second, using an encoder network Q, it predicts c from the generated image and effectively maximizes a lower bound on the mutual information between the input code c and the generated image $\tilde{x}$. The final objective is the combination of the original GAN objective plus that of the inferred code $\hat{c} \sim Q(c|x)$:

$$\min_{G,Q} \max_D V_{\text{InfoGAN}}(D, G, Q) =$$
$$V_{\text{GAN}}(D, G) - \lambda(\mathbb{E}_{x \sim G(z,c), c \sim p(c)}[\log Q(c|x)] + H(c)).$$
(17)

The variable c can follow a discrete categorical distribution or a continuous distribution such as the normal distribution.
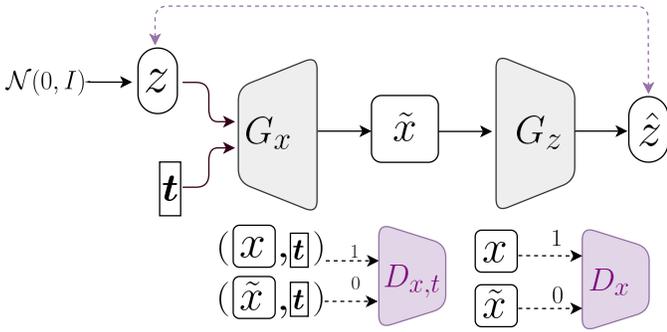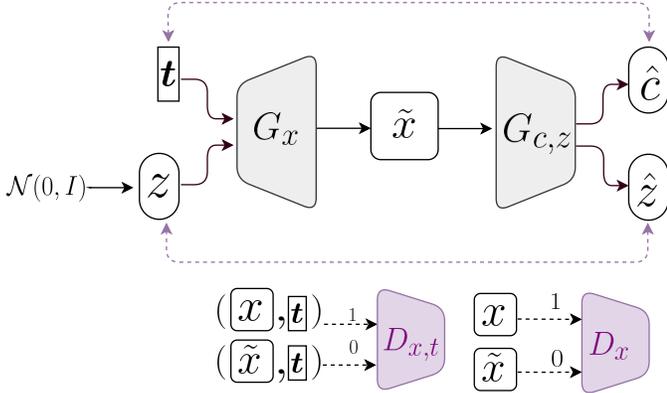
Fig. 5: Conditional InfoGAN (cInfoGAN).

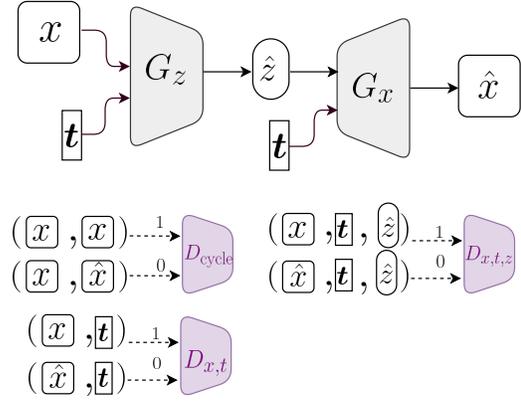Fig. 6: Dual conditional InfoGAN (D-cInfoGAN).

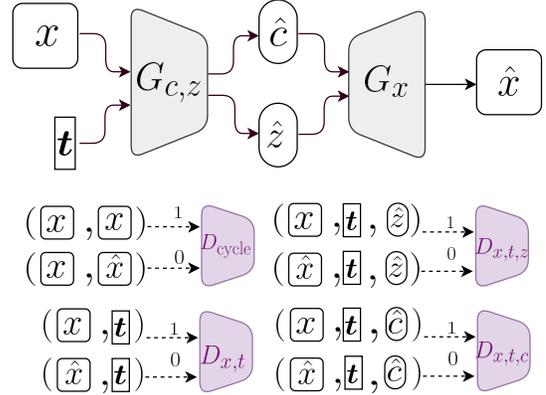Fig. 7: Conditional Adversarial VAE (cAVAE).

Fig. 8: Dual conditional Adversarial VAE (D-cAVAE).

InfoGAN is an unsupervised model popular for learning disentangled factors of variation [68, 43, 56].

We adopt a conditional version of InfoGAN–denoted by cInfoGAN–which is a conditional GAN augmented with an inference mechanism using the InfoGAN formulation. We experiment with two variants of cInfoGAN; a single latent variable model (cInfoGAN) shown in Figure 5, where the discriminator D_x is trained to distinguish between real ($\boldsymbol{x}$) and fake ($\tilde{\boldsymbol{x}}$) images while the discriminator $D_{x,t}$ distinguishes between the positive pair ($\boldsymbol{x}, \boldsymbol{t}$) and the corresponding negative pair ($\tilde{\boldsymbol{x}}, \boldsymbol{t}$), where $\tilde{\boldsymbol{x}} = G_x(\boldsymbol{z}, \boldsymbol{t})$ and $\boldsymbol{t}$ is the conditioning vector representing content. With the help of G_z, InfoGAN's mutual information objective is applied on $\boldsymbol{z}$ which represents the unsupervised style.

We also present a double latent variable model of InfoGAN (D-cInfoGAN) shown in Figure 6 where in addition to inferring $\hat{\boldsymbol{z}}$ we also infer $\hat{\boldsymbol{c}}$ through cycle consistency using the ℓ_1 norm.

2) cAVAE: Variational Auto-Encoders (VAEs) [42] are latent variable models commonly used for inferring disentangled factors of variation governing the data distribution. Let $\boldsymbol{x}$ be the random variable over the data distribution and $\boldsymbol{z}$ the random variable over the latent space. VAEs are trained by alternating between two phases, an inference phase where an encoder G_z is used to map a sample from the data to the latent space and infer the posterior distribution $q(\boldsymbol{z}|\boldsymbol{x})$ and a generation phase where a decoder G_x reconstructs the original image using samples of the posterior distribution with likelihood $p(\boldsymbol{x}|\boldsymbol{z})$.

VAEs maximize the evidence lower bound (ELBO) on the likelihood $p(\boldsymbol{x})$:

$$\max_G V_{\text{VAE}}(G_x, G_z) =$$
$$\mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}[q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z})]. \quad (18)$$

Kingma and Welling [42] also introduced a conditional version of VAE (cVAE) where $p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})$ is guided by both the latent code $\boldsymbol{z}$ and conditioning factor $\boldsymbol{c}$. There have also been many attempts in combining VAEs and GANs. Notable efforts are that of Larsen et al. [46], [52] and [75].

Conditional Adversarial Variational Autoencoder (cAVAE) is very similar to conditional Variational AutoEncoder (cVAE) but uses an adversarial formulation for the likelihood $p(x|z, c)$. Following the adversarial formulation for reconstruction [52, 47], a discriminator D_{cycle} is trained on positive pairs ($\boldsymbol{x}, \boldsymbol{x}$) and negative pairs ($\boldsymbol{x}, \hat{\boldsymbol{x}}$), where $\hat{\boldsymbol{x}} \sim p(x|t, \hat{\boldsymbol{z}})$ and $\hat{\boldsymbol{z}} \sim q(z|x)$. For the conditional generation we train a discriminator $D_{x,t}$ on positive pairs ($\boldsymbol{x}, \boldsymbol{t}$) and negative pairs ($\hat{\boldsymbol{x}}, \boldsymbol{t}$), where $\boldsymbol{t}$ is the conditioning factor. We empirically discover that adding an additional discriminator $D_{x,t,z}$ which also takes advantage of the latent code $\hat{\boldsymbol{z}}$ improves inference. Similar to cInfoGAN, we use two versions of cAVAE: a single latent variable version denoted by cAVAE (Figure 7) and a double latent variable version D-cAVAE (Figure 8), where in addition to the style posterior $q(\boldsymbol{z}|\boldsymbol{x})$, we also infer the content posterior $q(\boldsymbol{c}|\boldsymbol{x})$. Accordingly, to improve inference on the content variable, we add the discriminator $D_{x,t,c}$.

C. Evaluation Metrics

We explain in detail various evaluation metrics used in our experiments.

1) Measure of disentanglement: Multiple methods have been proposed to measure the degree of disentanglement between variables [32]. In this work, we propose a measure which evaluates the desired disentanglement characteristics of both the feature generator and the image generator. To have good feature disentanglement, we desire a feature generator (*i.e.*, encoder) that separates the information in an image in two disjoint variables of style and content in such a way that 1) the inferred information is consistent across images. *e.g.*, position and orientation is encoded the same way for all images; and 2) every piece of information is handled by *only* one of the two variables, meaning that the style and content variables do not share features. In order to measure these properties, we propose Cross Image Feature Consistency (CIFC) error where we measure the model's ability to first generate hybrid images of mixed style and content inferred from two different images and then its ability to reconstruct the original images. Figure 9 illustrates this process. As seen in this figure, given two images I_a and I_b, hybrid images I_{ab} and I_{ba} are generated using the pairs $(\hat{c}_a, \hat{z}_b)$ and $(\hat{c}_b, \hat{z}_a)$ respectively. By taking another step of hybrid image generation, I_{aa} and I_{bb} are generated as reconstructions of I_a and I_b respectively. To make the evaluation robust with respect to high frequency image details, we compute the reconstruction error in the feature space. In retrospect, the disentanglement measure is computed as:

$$CIFC = \mathbb{E}_{(I_a, I_b) \sim q_{\text{test}}(\boldsymbol{x})}[\|\hat{z}_a - \hat{z}_{aa}\| + \|\hat{c}_a - \hat{c}_{aa}\| + \\ \|\hat{z}_b - \hat{z}_{bb}\| + \|\hat{c}_b - \hat{c}_{bb}\|], \quad (19)$$

where $q_{\text{test}}(\boldsymbol{x})$ represents the empirical distribution of the test images.

2) FID: The Frechet inception distance (FID) score [31] measures the distance between the real and generated data distributions. An inception model is required for calculating FID, but since the conventional inception model used for FID is pretrained on colored natural images, it is not suitable to be used with LIDC which consists of single channel CT scans. Consequently, we train an inception model on the LIDC dataset to classify benign and malignant nodules. We use InceptionV3 [65] up to layer "*mixed3*" (initialized with pretrained ImageNet weights), and append a global average pooling layer followed by a dense layer.

3) Inception Score: Inception Score (IS) [61] is another quantitative metric on image generation which is commonly used to measure the diversity of the generated images. We use the same inception model described above to calculate IS. The TensorFlow-GAN library [64] is used to calculate both FID and IS.

4) Conditional Generation Accuracy: Since the inception model is pretrained on image labels, we can evaluate the classification accuracy of the generated images to measure the performance of the conditional generation. This metric is referred to as CGAcc through out this text.

TABLE I: Comparison of image generation metrics on LIDC dataset of single and double variable baselines

Method	LIDC		
	FID($\downarrow$)	IS($\uparrow$)	CGAcc($\uparrow$)
cInfoGAN	0.283 ± 0.06	1.366 ± 0.02	0.740 ± 0.02
cAVAE	0.181 ± 0.03	1.424 ± 0.01	0.809 ± 0.02
D-cInfoGAN	0.333 ± 0.06	1.342 ± 0.09	0.645 ± 0.04
D-cAVAE	0.378 ± 0.03	1.371 ± 0.04	0.597 ± 0.07
DAI [44]	0.106 ± 0.02	1.423 ± 0.05	0.773 ± 0.03
DRAI	0.089 ± 0.02	1.422 ± 0.03	0.773 ± 0.02

5) Image retrieval scores: Quantitative Image retrieval tests are conducted on LIDC dataset, where given a query image, the curated test set[1] is searched to find the closest match.

In our experiments, we construct a query dataset from the nodules in the original test set which were excluded from our curated test set (those with median malignancy equals to 3). To quantitatively evaluate the performance of image retrieval, we construct an attribution vector for each image, which consists of the conditioning vector and the nodule size. The nodule size is calculated from the nodule segmentation maps provided in the dataset.

For each query image, the image retrieval test requires the model to find top-N nearest neighbors (we set N equal to 3 in our experiments) in the test set. The model's searching criterion is the distance between two images, which is defined by the mean absolute error (MAE) between their inferred latent representations.

We introduce two quantitative metrics to evaluate the image retrieval performance. One is the "disagreement divergence" that measures the average disagreement—computed via mean square error (MSE)—between the label of the query image and those of the top-N retrieved images. The other measure is the percentage of the label from the ground truth image found in the top-N nearest neighbors. For every query image, the ground truth image is defined as the image with the smallest disagreement divergence in the test set. It is important to note that since the conditioning vectors used for the evaluation of the image retrieval performance is tied with the content variable, this quantitative measure can only evaluate the content based image retrieval performance. We resort to qualitative assessments to evaluate the style based image retrieval experiments.

D. Generation evaluation

To evaluate the quality and diversity of the generated images, we measure FID and IS (see Section III-C) for the proposed DRAI model and various double and single latent variable baselines described in Section III-B. The results are reported in Table I and Table II for LIDC and HAM10000 datasets respectively. We also report in these tables the CGAcc score which measures how well the generated images match the conditioning factors. For the LIDC dataset, we observe all methods have comparable IS and CGAcc score while DRAI and DAI have significantly lower FID compared to

---

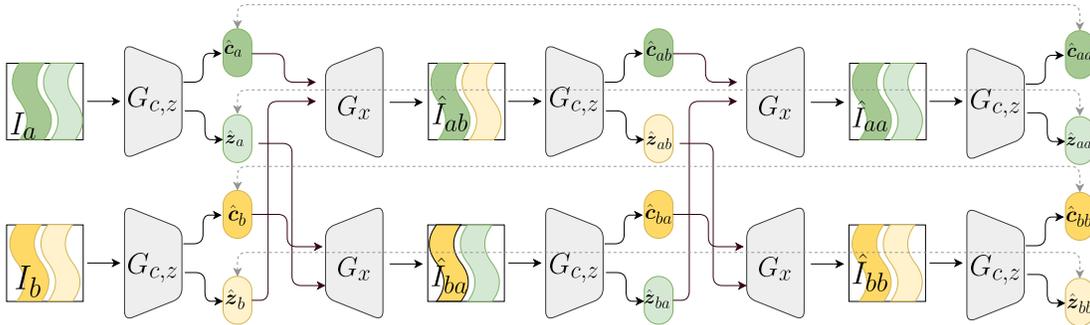[1]See Section III-A2 for details on how the data is curated

Fig. 9: Cross Image Feature Consistency (CIFC) error. CIFC is computed by first generating hybrid images of mixed style and content across two different images and then reconstructing the original images. For a more robust evaluation, CIFC is measured in the feature space.

TABLE II: Comparison of image generation metrics on HAM10000 dataset of single and double variable baselines.

Method	HAM		
	FID($\downarrow$)	IS($\uparrow$)	CGAcc($\uparrow$)
cInfoGAN	1.351 ± 0.33	1.326 ± 0.03	0.647 ± 0.04
cAVAE	3.566 ± 0.56	1.371 ± 0.01	0.651 ± 0.02
D-cInfoGAN	1.684 ± 0.42	1.449 ± 0.03	0.654 ± 0.06
D-cAVAE	4.893 ± 0.99	1.321 ± 0.01	0.578 ± 0.01
DAI [44]	1.327 ± 0.06	1.304 ± 0.01	0.656 ± 0.02
DRAI	1.224 ± 0.05	1.300 ± 0.01	0.630 ± 0.02

other baselines, with DRAI having better performance. For the HAM10000 dataset, DRAI once again achieves the best FID score while D-cInfoGAN achieves the best IS. All methods seem to perform on par regarding the conditional generation accuracy CGAcc. It is interesting to note that compared with other baselines, DAI and DRAI achieve lower prediction intervals, which indicate more stable training process. For each model, the prediction interval is computed across four training sessions.

We highlight that while FID and IS are the most common metrics for the evaluation of GAN based models, they do not provide the optimum assessment [7] and thus qualitative assessment is needed. For the qualitative evaluation, we visualize samples generated by each model. We use the provided conditioning vector for the generation process and only sample the style variable z. The generated samples are visualized in Figure 10 and Figure 11 for LIDC and HAM10000 datasets respectively. In every sub-figure, the first column represents the reference image corresponding to the conditioning vector used for the image generation, and the remaining columns represent synthesized images. Studying these figures, we make the following observations:

- For LIDC, DRAI and DAI generate images with higher fidelity and less artifacts compared to other baselines. For HAM, all compared models generate realistic looking images. This could be because HAM10000 is a larger dataset.
- There is evidently more content consistency in double variable models compared to single variable models. This validates the hypothesis that using two latent variables to infer style and content allows for more control over the generation process. Among double variable models,

DRAI has the best content consistency.
- By fixing the content and sampling the style variable, we can discover the types of information that are encoded as style and content for each dataset. We observe that the learned content information are color and lesion size for HAM10000, and nodule size for LIDC; while the learned style information are location, orientation and lesion shape for HAM10000 and background for LIDC. We also observe that DRAI is very successful in preserving the content information when there is no stochasticity in the content variable (*i.e.*, c is fixed). As for other baselines, sampling style results in changing the content information of the generated images, which indicates information leak from the content variable to the style variable. The results show that compared to DAI and other baselines, DRAI achieves better separation of style and content.

E. Evaluation of Style-Content Disentanglement

Achieving good style-content disentanglement in both inference and generation phases is the main focus of this work. We conduct multiple quantitative and qualitative experiments to asses the quality of disentanglement in DRAI (our proposed method) as well as the competing baselines.

1) Quantitative evaluation using CIFC: As a quantitative metric, we introduce the disentanglement error CIFC (refer to III-C for details). Table III shows results on this metric. As seen from this table, in both HAM10000 and LIDC datasets, DRAI and DAI achieve significantly lower disentanglement error compared to other baselines, indicating the importance of dual adversarial inference. The dual adversarial inference formulation with style and content independence assumption facilitates their separation, and allows for better disentanglement compared to D-cInfoGAN and D-cAVAE where the latent space is merely divided into two variables with separate encoders and separate discriminators. This experiment also shows that model architecture alone is not enough for style-content separation and proper objective functions are needed to guide the learning process. We observe that DRAI improves over DAI by a notable margin, which demonstrates the advantage of the proposed disentanglement regularizations; on one hand, the information regularization objective through GRL minimizes the shared information between style and content variables, and on the other hand, the self-supervised regularization objective not only allows for better control of
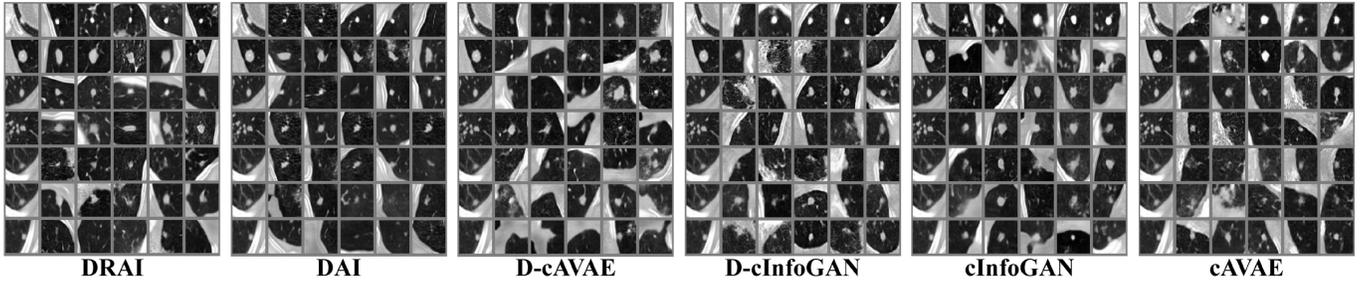
Fig. 10: Conditional generations on LIDC. In every sub-graph, the first column represents the real image corresponding to the conditioning vector. The images are generated by keeping the content code (c) fixed and only sampling the style codes (z).
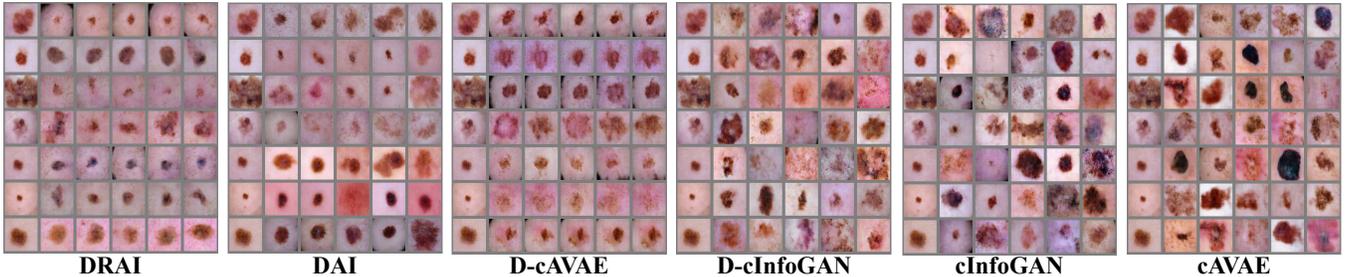

Fig. 11: Conditional generations HAM10000. In every sub-graph, the first column represents the real image corresponding to the conditioning vector. The images are generated by keeping the content code (z) fixed and only sampling the style codes (z).

TABLE III: Content-style disentanglement evaluation using CIFC error.

Method	CIFC($\downarrow$)	
	HAM10000	LIDC
D-cInfoGAN	1.201 ± 0.17	1.625 ± 0.11
D-cAVAE	1.354 ± 0.03	1.944 ± 0.02
DAI [44]	0.256 ± 0.01	1.096 ± 0.28
DRAI	0.210 ± 0.01	0.456 ± 0.06

the learned features but also facilitates disentanglement. In the ablation studies (Section III-F), we investigate the effect of the individual components of DRAI on disentanglement.

2) Qualitative evaluation: To have a more interpretable evaluation, we qualitatively assess the style-content disentanglement through generating hybrid images by combining style and content information from different sources. We can then evaluate the extent to which the style and content of the generated images respect the corresponding style and content of the source images. Figure 12 and Figure 13 show these results on the two datasets. Following these figures, we make the following observations:

- For the LIDC dataset, DAI and DRAI learn CT image background as style and nodule as content. This is due to the fact that the nodule characteristics such as nodule size is included in the conditioning factor and thus the content tends to focus on those attributes. Since the background of the CT images is not well represented in the conditioning label of the LIDC dataset, the model tends to learn features controlling the background in an unsupervised way as style. For style features, the boundary between high and low density tissues is often captured. The model also inclines towards ignoring small

artifacts in the background, which suggests that it infers based on higher-level features.

- Thanks to the added disentanglement regularizations, DRAI has the best content-style separation compared to all other baselines and demonstrates clear decoupling of the two variables. Because of the self-supervised regularization objective, DRAI assigns more emphases on capturing nodule characteristics as part of the content and background as part of the style. DAI follows closely behind but is not as successful as DRAI in inferring the correct background or nodule size. Overall, it is evident from the qualitative experiments that the proposed disentanglement regularizations help to decouple the style and content variables.
- DRAI generates better reconstructed images compared to other baselines. The inferred content is mostly constrained to nodule size and density, which are correlated. It is likely that the nodule size and malignancy dominate the content, while other LIDC characteristics seem less discriminative in comparison. A richer and finer conditioning factor could help alleviate this.
- For HAM10000, style encodes features shared across lesion types such as shape, orientation and location of the skin lesion, while the content constitutes lesion type, size and color information of both the lesion and background.
- The conditioning factor for HAM10000 is only comprised of lesion type, so learning color and lesion size as part of the content, is the result of inductive biases caused by inferring *two* latent variables to learn style and content, and also the result of objective functions such as the *self supervised regularization*. The unsupervised learning of content (in addition to the supervised learning) helps bring about a richer content representation.
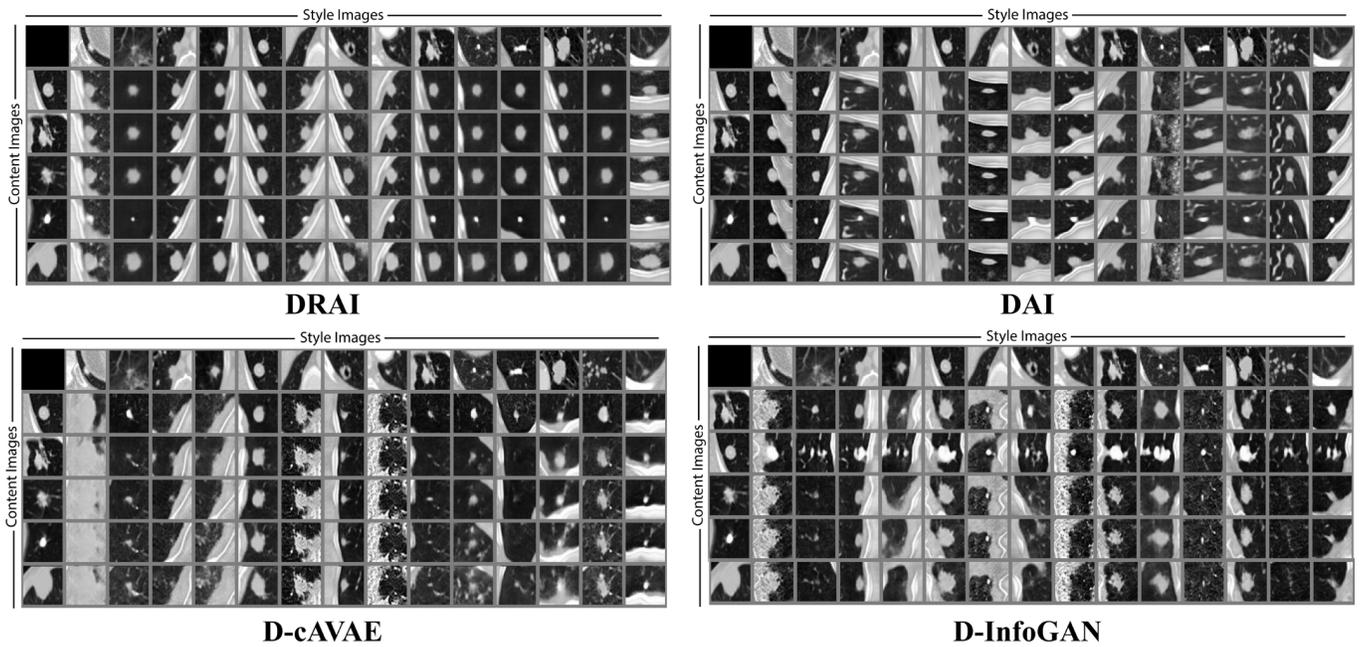- DRAI also shows clear separation of style and content

Fig. 12: Qualitative evaluation of style-content disentanglement through hybrid image generation on LIDC dataset. In every sub-figure, images in the first row present style image references and those in the first column present content image references. Hybrid images are generated by using the style and content codes inferred from style and content reference images respectively.
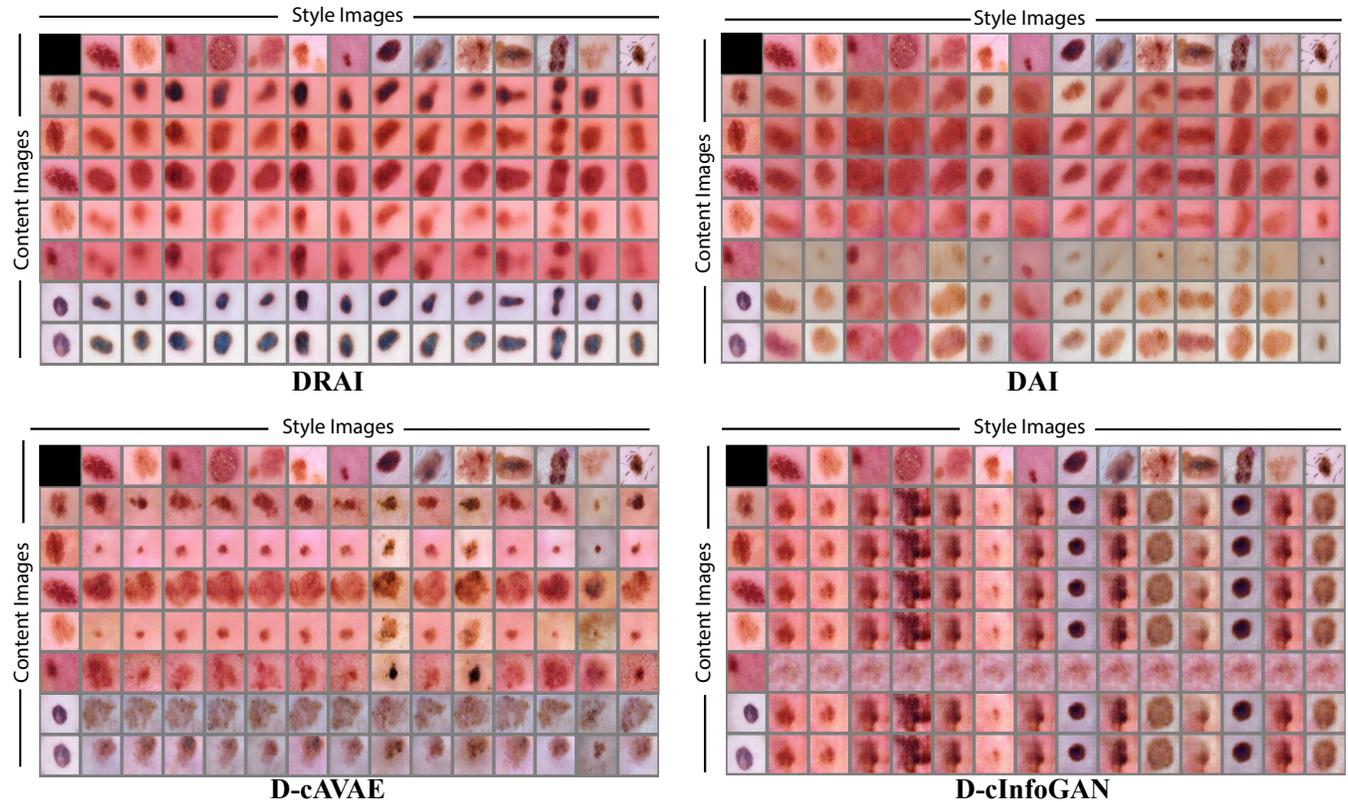

Fig. 13: Qualitative evaluation of style-content disentanglement through hybrid image generation on HAM10000 dataset. In every sub-figure, images in the first row represent style image references and those in the first column represent content image references. Hybrid images are generated by using the style and content codes inferred from style and content reference images respectively.

on HAM10000 dataset. DAI fails to disentangle the two vectors and some features such as color are captures by both the style and content variables. As for D-cInfoGAN and D-cAVAE models, the hybrid image generation is mostly dominated by either the style (for D-cInfoGAN) or the content (for D-cAVAE).

• HAM10000 is a harder dataset in terms of the vast lesion variability. Using separate style and content vectors

without an independence assumption between them (*e.g.*, D-cInfoGAN and D-cAVAE), results in the coupling of the two representations which as seen in Figure 12 and Figure 13, is detrimental.

- Requiring disentanglement adds an additional regularization effect on the generation task, which introduces a trade-off between disentanglement and generation. This is evident in the HAM10000 dataset where we observe images generated by double variable models (especially DRAI and DAI where there is more regularization involved) are blurrier than those generated by single variable models. Careful tuning of the hyper-parameters is required to generate high quality images which also satisfy the disentanglement criteria for style and content.

3) Image retrieval: While the above experiments evaluate disentanglement on both inference and generation, we would also like to further assess how the inferred style and content features are disentangled from each other in the retrieval test. For a given query image, we can retrieve its closest matches in a reference dataset, in terms of style and content separately. The style-content disentanglement can then be evaluated by investigating the style and content of the retrieved images.

We resort to qualitative evaluation of style based image retrieval since there is no label on style. We could however, evaluate content based image retrieval, based on the conditioning vector (nodule size and other characteristics) associated with each image. This allows us to evaluate how well each model can infer correct content information.Quantitative results on content based image retrieval are presented in Table IV where we present the disagreement divergence and label overlap. As seen in this table, while DAI and DRAI have relatively similar performance, other baselines (*i.e.*, D-cINfoGAN and D-cAVAE) significantly under perform. Interestingly, "Pixel Matched", which is the nearest neighbor to the query image computed in the image space, focuses only on the background (*i.e.*, the dominant factor in the image). This is also quantitatively shown in Table IV; Pixel Matched performs poorly on both metrics.

For the quantitative evaluation, we resort to using only the LIDC dataset since it has a richer conditioning label compared to HAM10000 dataset.

For the qualitative evaluation, the top nearest neighbors to the query image with respect to its inferred style and content codes are retrieved. Figure 14 and Figure 15 showcase the qualitative evaluation of style based and content based image retrieval on LIDC and HAM10000 datasets respectively. In each figure, the horizontal images next to the query image represent the retrieved images based on the style information, while the vertical images represent the retrieved images based on the content information. For single variable baselines *i.e.*, cAVAE and cInfoGAN, image retrieval is performed based on the inferred single code. We observe that DRAI manages to retrieve images based on style or content without allowing interference from the other variable (*i.e.*, content or style respectively). As for the other baselines, since the style and content information are intertwined, the retrieved images have both style and content information which is undesirable.

TABLE IV: Quantitative evaluation on the top 3 nearest neighbors for content based image retrieval on LIDC dataset.

Method	disagreement divergence(↓)	label-overlap(↑)
Pixel Matched	2.42	35.27
D-cInfoGAN	2.566 ± 0.04	33.850 ± 1.19
D-cAVAE	2.961 ± 0.08	35.010 ± 0.96
DAI+selfReg+MIReg	1.826 ± 0.05	52.325 ± 3.87
DAI+featureCycle	1.458 ± 0.11	59.173 ± 3.16
DAI+MIReg	1.928 ± 0.26	54.909 ± 1.82
DAI+selfReg	1.791 ± 0.26	56.589 ± 3.59
DAI [44]	1.687 ± 0.32	54.780 ± 5.02
DRAI	1.575 ± 0.04	53.359 ± 2.94

F. Ablation studies

As discussed in Section II, our proposed method DRAI is composed of different components to ensure the quality of inference and style-content disentanglement. In this section, we perform ablation studies to evaluate the effect of each component, both quantitatively and qualitatively. Ablated models use the same architecture with the same amount of parameters. The quantitative assessment is presented in Table V and Table VI where we study the performance of each component with respect to FID, IS, CGAcc and CIFC metrics. We observe that on both LIDC and HAM10000, each added component improves over DAI, while the best performance is achieved when these components are combined together to form DRAI.

We also provide qualitative analysis of ablation studies which are visualized in Figure 16 and Figure 17 for LIDC and HAM10000 datasets respectively. Our qualitative results also support the quantitative findings and show each added component in DRAI namely self-supervised regularization ("selfReg"), shared information minimization ("MIReg") and latent code cycle consistency ("featureCycle") can improve DAI. We observe selfReg and MIReg have significant impact on disentanglement when combined together. However, the model may suffer from over regularization especially in the case of LIDC dataset where some content information *e.g.*, nodule size is under estimated. This effect is alleviated when the featureCycle term is integrated. We achieve the best disentanglement with style-content preservation in DRAI where we integrate all 3 components to DAI.

a) Latent space interpolation: For our best performing model DRAI, as a sanity check for memorization and overfitting, we look at latent space interpolations between testset examples as shown in Figure 18 for both LIDC and HAM10000. We conduct 3 tests for each dataset: (i) interpolating the content while maintaining the style of the source image, (ii) interpolating the style while maintaining the content of the source image, (iii) interpolating both style and content.

In all 3 tests, we observe smooth transitions between pairs of examples with plausible intermediary image generations. This experiment also demonstrates the disentanglement power of DRAI; in the content interpolation experiment, DRAI maintains the style of the source image while the content gradually transitions from the source image to the target image, and in the style interpolation test, the content is maintained and the style transitions.

Fig. 14: Style-Content image retrieval on LIDC dataset. The "Target image" is the image in the testset with the lowest disagreement divergence to the query image. the "Pixel Matched" image is the nearest neighbor to the query image in the pixel space. In every sub-figure, images to the right of the query image represent the nearest neighbors computed via the style code, while images at the bottom, represent the nearest neighbors computed via the content code.

b) t-SNE plots for style and content latent spaces: To better understand how DRAI clusters style and content information, we provide t-SNE [67] plots of the style and content variables for both datasets in 2 dimensions. As seen in Figure 19, for LIDC, the style clusters similar backgrounds together while the content focuses more on the size of the nodules. Whereas for HAM10000, the style focuses more on the shape, orientation and position of the lesion while the content focuses more on size of the lesions and color.

G. Implementation details

In this section, we provide the important implementation details of DRAI. Firstly, to reduce the risk of information leak between style and content, we use completely separate encoders to infer the two variables. For the same reason, the dual adversarial discriminators are also implemented separately for style and content. The data augmentation includes random flipping and cropping. To enable self-supervised regularization, each batch is trained twice, first with the original images and then with the transformed batch. The transformations include rotations of 90, 180, and 270 degrees, as well as horizontal and vertical flipping. LSGAN (Least Square GAN) [51] loss is used for all GAN generators and discriminators, while ℓ_1 loss is used for the components related to disentanglement constraints, *i.e.*, GRL strategy and self-supervised regularization. In general, we found that "Image cycle-consistency" and "Latent code cycle-consistency" objectives improve the stability of training. This is evident by DRAI achieving lower prediction intervals (*i.e.*, standard deviation across multiple runs with different seeds) in our experiments.

We did not introduce any coefficients for the loss components in Equation (16) since other than the KL terms, they were all relatively on the same scale. As for the KL

Fig. 15: Style-Content image retrieval on HAM10000. In every sub-figure, images to the right of the query image represent the nearest neighbors computed via the style code, while images at the bottom, represent the nearest neighbors computed via the content code.
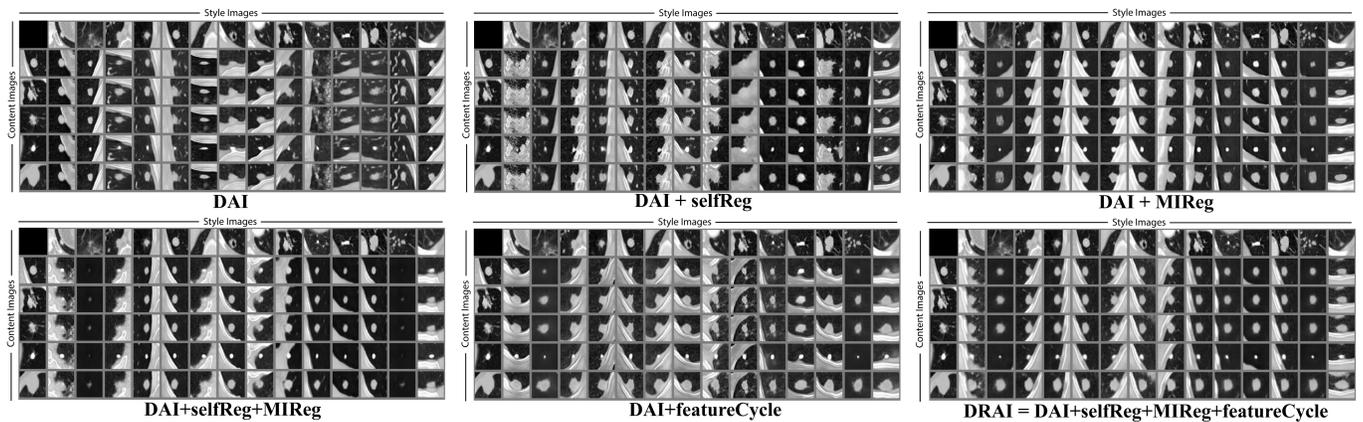

Fig. 16: Qualitative ablation study of style-content disentanglement through hybrid image generation on LIDC dataset.

co-efficients λ, we tried multiple values and qualitatively evaluated the results. Since the model was not overly sensitive to KL, we used a coefficient of 1 for all KL components.

All models including the baselines are implemented in TensorFlow [1] version 2.1, and the models are optimized via Adam [41] with initial learning rate $1e^{-5}$. For IS and
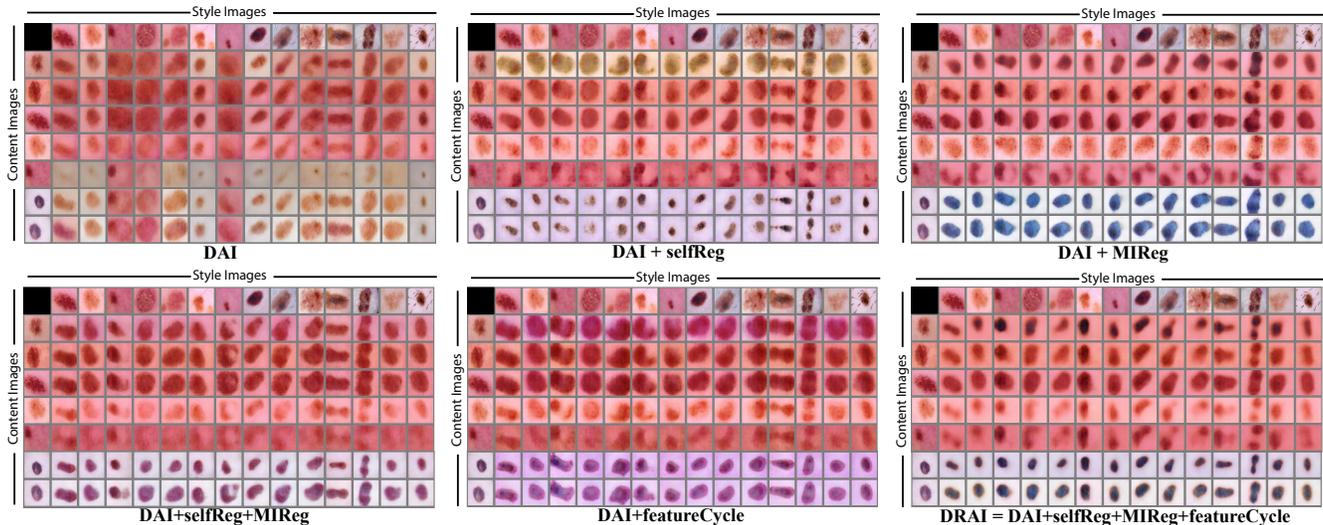
Fig. 17: Qualitative ablation study of style-content disentanglement through hybrid image generation on HAM10000 dataset.

TABLE V: Quantitative ablation study on LIDC dataset

Method	LIDC			
	FID($\downarrow$)	IS($\uparrow$)	CGAcc($\uparrow$)	CIFC($\downarrow$)
DAI [44]	0.106 ± 0.02	1.423 ± 0.052	0.773 ± 0.031	1.096 ± 0.284
DRAI=DAI+selfReg+MIReg+featureCycle	0.089 ± 0.02	1.422 ± 0.030	0.773 ± 0.016	0.456 ± 0.069
DAI+selfReg+MIReg	0.176 ± 0.06	1.433 ± 0.018	0.760 ± 0.015	0.554 ± 0.185
DAI+featureCycle	0.221 ± 0.07	1.383 ± 0.039	0.708 ± 0.037	0.913 ± 0.074
DAI+MIReg	0.154 ± 0.04	1.411 ± 0.028	0.746 ± 0.041	0.747 ± 0.226
DAI+selfReg	0.208 ± 0.05	1.433 ± 0.033	0.780 ± 0.014	0.781 ± 0.203

TABLE VI: Quantitative ablation study on HAM10000 dataset.

Method	HAM			
	FID($\downarrow$)	IS($\uparrow$)	CGAcc($\uparrow$)	CIFC($\downarrow$)
DAI [44]	1.327 ± 0.06	0.656 ± 0.02	1.304 ± 0.01	0.256 ± 0.01
DRAI=DAI+selfReg+MIReg+featureCycle	1.224 ± 0.050	0.618 ± 0.011	1.300 ± 0.01	0.210 ± 0.01
DAI+selfReg+MIReg	1.350 ± 0.12	0.683 ± 0.02	1.299 ± 0.01	0.233 ± 0.01
DAI+featureCycle	1.367 ± 0.12	0.690 ± 0.01	1.296 ± 0.01	0.311 ± 0.01
DAI+MIReg	1.298 ± 0.12	0.647 ± 0.04	1.290 ± 0.02	0.228 ± 0.01
DAI+selfReg	1.347 ± 0.14	0.653 ± 0.03	1.295 ± 0.01	0.219 ± 0.04

FID computation, we fine-tune the inception model on a 5 way classification on nodule size for LIDC and a 7 way classification on lesion type for HAM10000. FID and IS are computed over a set of 5000 generated images.

IV. RELATED WORK

a) Connection to other conditional GANs in medical imaging: While adversarial training has been used extensively in the medical imaging domain, most work uses adversarial training to improve image segmentation and domain adaptation. The methods that use adversarial learning for image generation can be divided into two broad categories; the first group are those which use image-to-image translation as a proxy to image generation. These models use an image mask as the conditioning factor, and the generator generates an image which respects the constraints imposed by the mask [38, 26, 17, 17, 55]. Jin et al. [38] condition the generative adversarial network on a 3D mask, for lung nodule generation. In order to embed the nodules within their background context, the GAN is conditioned on a volume of interest whose central part containing the nodule has been erased. A favored approach for generating synthetic fundus retinal images is to use vessel segmentation maps as the conditioning factor. Guibas et al. [26] uses two GANs in sequence to generate fundus images. The first GAN generates vessel masks, and in stage two, a second GAN is trained to generate fundus retinal images from the vessel masks of stage one. Costa et al. [17] first use a U-Net based model to generate vessel segmentation masks from fundus images. An adversarial image-to-image translation model is then used to translate the mask back to the original image.

In Mok and Chung [55] the generator is conditioned on a brain tumor mask and generates brain MRI. To ensure correspondence between the tumour in the generated image and the mask, they further forced the generator to output the tumour boundaries in the generation process. Bissoto et al. [12] uses the semantic segmentation of skin lesions and generate
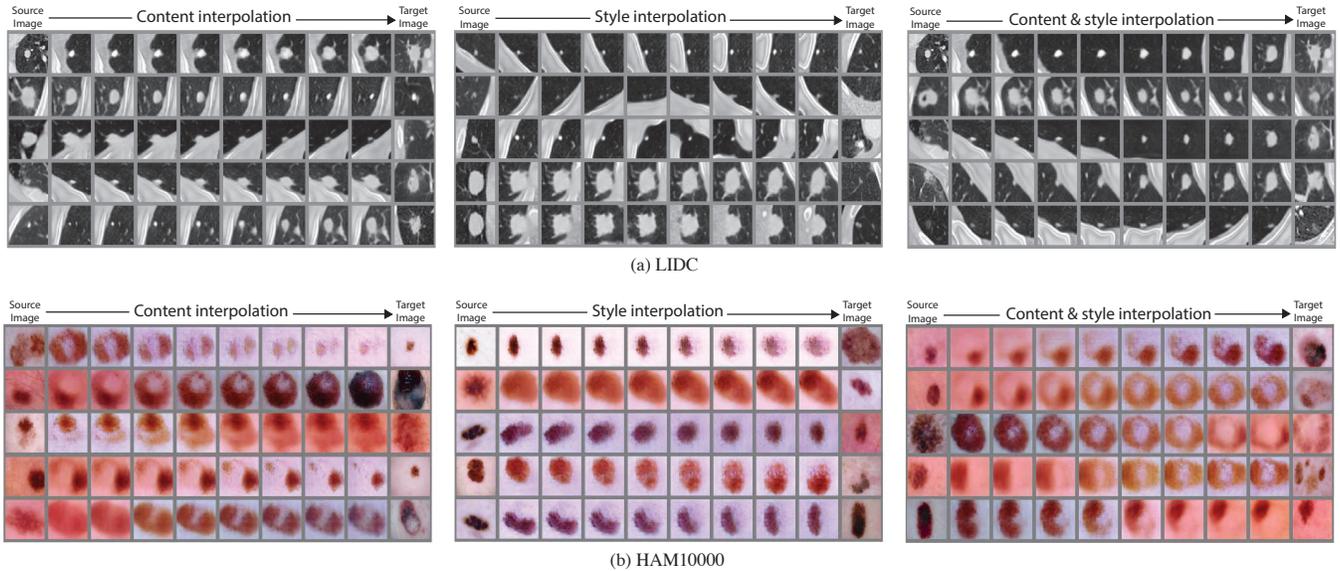
(a) LIDC

(b) HAM10000

Fig. 18: Latent space interpolation between testset examples for (a) LIDC and (b) HAM10000. For each dataset, we show (left) content only interpolation from source to target images while maintaining style of the source image, (middle) style only interpolation from source to target images while maintaining content of the source image and (right) both style and content interpolation.

high resolution images. Their model combines the pix2pix framework [35] with multi-scale discriminators to iteratively generate coarse to fine images.

While methods in this category give a lot of control over the generated images, the generator is limited to learning domain information such as low level texture and not higher level information such as shape and composition. Such information is presented in the mask which requires an additional model or an expert has to manually outline the mask which can get tedious for a lot of images.

The second category of methods are those which use high level class information in the form of a vector as the conditioning factor. Hu et al. [33] takes Gleason score vector as input to the conditional GAN to generate synthetic prostate diffusion imaging data corresponding to a particular cancer grade. Baur et al. [8] used a progressively growing model to generate high resolution images of skin lesions.

As mentioned in the introduction one potential pitfall of such methods is that by just using the class label as conditioning factor, it is hard to have control over the nuances of every class. While our proposed model falls within this category, our inference mechanism allows us to overcome this challenge by using the image data itself to discover factors of variation corresponding to various nuances of the content.

b) Disentangled representation learning: In the literature, disentanglement of style and content is primarily used for domain translation or domain adaptation. Content is defined as domain agnostic information shared between the domains, while style is defined as domain specific information. The goal of disentanglement to preserve as much content as possible and to prevent leakage of style from one domain to another. Gonzalez-Garcia et al. [24] used adversarial disentanglement for image to image translation. In order to prevent exposure of style from domain A to domain B, a Gradient Reversal Layer (GRL) is used to penalize shared information between the

generator of domain B and style of domain A. In contrast, our proposed DRAI, uses GRL to minimize the shared information between style and content. In the medical domain, Yang et al. [71] aim to disentangle anatomical information and modality information in order to improve on a downstream liver segmentation task.

Ben-Cohen et al. [11] used adversarial learning to infer content agnostic features as style. Intuitively their method is similar to using GRL to minimize leakage of content information into a style variable. However, while [11] prevents leakage of content into style, it does not prevent the reverse effect which is leakage of style into content and thus does not guarantee disentanglement.

Yang et al. [70] use disentangle learning of modality agnostic and modality specific features in order to facilitate cross-modality liver segmentation. They use a mixture of adversarial training and cycle consistency loss to achieve disentanglement. The cycle-consistency component is used for in-domain reconstruction and the adversarial component is used for cross-domain translation. The two components encourage the disentanglement of the latent space, decomposing it into modality agnostic and modality specific sub-spaces.

To achieve disentanglement between modality information and anatomical structures in cardiac MR images, Chartsias et al. [13] use an autoencoder with two encoders: one for the modality information (style) and another for anatomical structures (content). They further impose constraints on the anatomical encoder such that every encoded pixel of the input image has a categorical distribution. As a result, the output of the anatomical encoder is a set of binary maps corresponding to cardiac substructures.

Disentangled representation learning has also been used for denoising of medical images. In Liao et al. [48], Given artifact affected CT images, metal-artifact reduction (MAR) is performed by disentangling the metal-artifact representations

(a) HAM10000 content
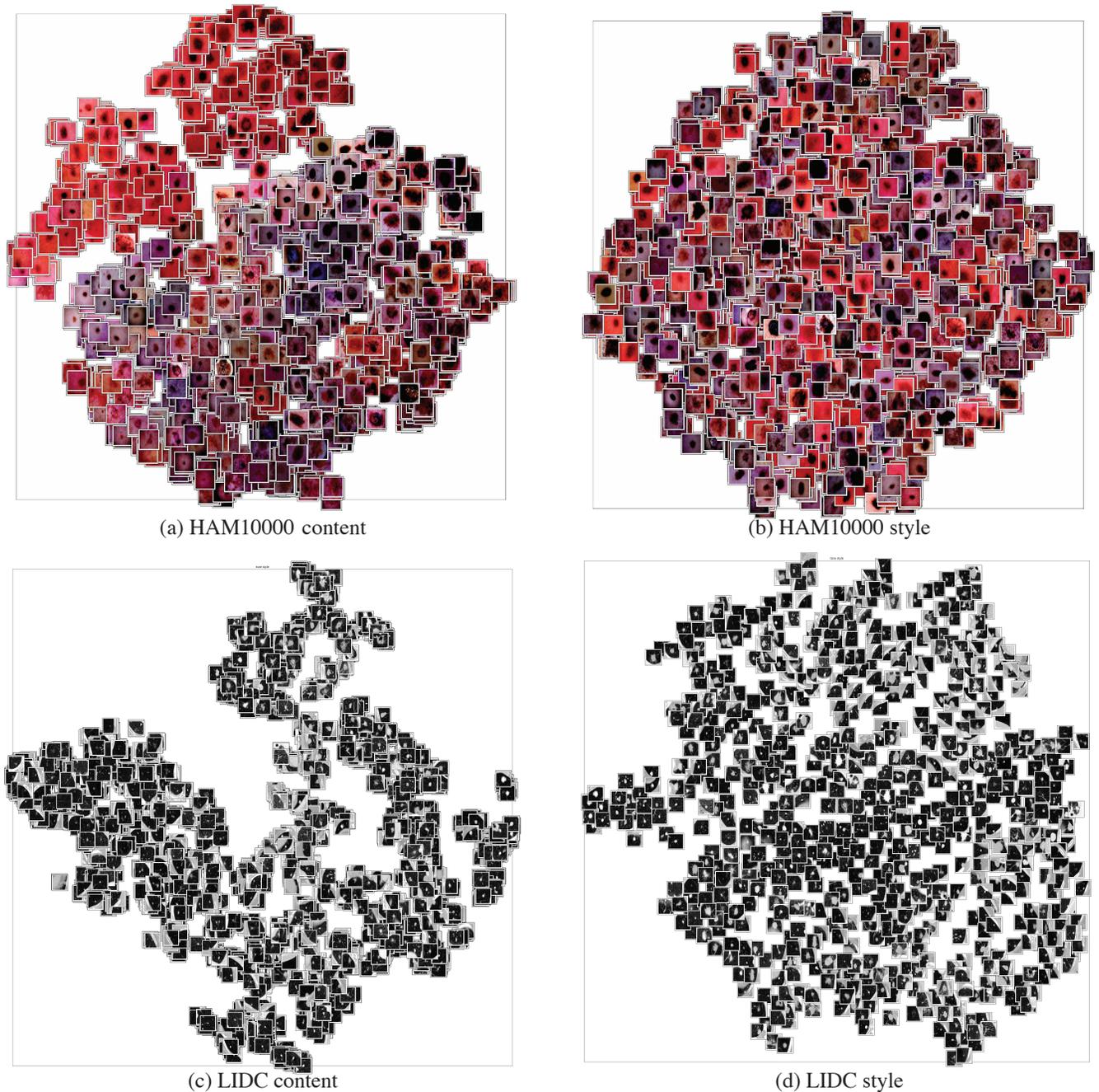
(b) HAM10000 style

(c) LIDC content

(d) LIDC style

Fig. 19: t-SNE plots of the inferred content (a and c) and style (b and d) codes for LIDC and HAM10000. In the case of HAM10000, the content variable focuses mostly on color and size whereas the style variable focuses on lesion location and orientation. In the case of LIDC, the content variable focuses mostly on nodule size whereas the style variable focuses on background information.

from the underlying CT images.

Sarhan et al. [62] use β-TCAV [14] to learn disentangled representations on an adversarial variation of the VAE. Their proposed model differs fundamentally from our work; its is a single variable model, without a conditional generative process, and does not infer separate style and content information.

Garcia1 et al. [23] used ALI (single variable) on structured MRI to discover regions of the brain that are involved in Autism Spectrum Disorder (ASD).

In contrast to previous work, we use style-content disentanglement to control features for conditional image generation.

To the best of our knowledge this is the first time such attempt has been made in the context of medical imaging.

c) Gradient Reverse Layer (GRL): Introduced in Ganin et al. [22], GRL uses the identity function in the forward pass, but reverses the sign of the gradient in the backward pass. GRL was initially intended for domain adaptation with the goal of learning domain invariant representations. Since then, GRL has been used extensively in various domain adaptation approaches [78, 37, 45]. For image translation, Gonzalez-Garcia et al. [24] used a model based on VAE-GAN and GRL to disentangle the attributes of paired data into shared

and exclusive representations. Raff and Sylvester [59] used GRL for *fairness* with the goal of learning a model whose outcome is invariant to sensitive attributes such as ethnicity or gender. It is worth mentioning that some fairness and invariant representation learning methods, rely on adversarial training which in spirit is similar to using GRL [39, 50].

Different from previous work, we introduce a novel use case for GRL to explicitly disentangle the inferred style and content information.

d) Connection to CycleGAN approaches: Zhu et al. [79] introduced CycleGAN as an image-translation (domain-translation) model which preserves the content and changes the style from one domain to another domain. At test time, the mapping between the domains is deterministic and one-to-one. Almahairi et al. [3] augmented CycleGAN with latent variables which allowed many to many mapping between the two domains. In this paper we do not tackle the problem of image domain translation and rather focus on conditional generation of desired content and style through dual adversarial inference. In our approach, the cycle-consistency is used to stabilize the adversarial inference. That being said, it would be possible to apply our framework to image domain translation problem as well. We leave exploration in this direction to future work.

V. CONCLUSION

We introduce DRAI, a frame work for generating synthetic medical images which allows control over the style and content of the generated images.

DRAI uses adversarial inference together with conditional generation and disentanglement constraints to learn content and style variables from the dataset. We compare DRAI quantitatively and qualitatively with multiple baselines and show its superiority in image generation in terms of quality, diversity and style-content disentanglement. Through ablation studies and comparisons with DAI [44], we show the impact of imposing the proposed disentanglement constraints over the content and style variables.

It is important to note that DRAI learns style in a completely unsupervised fashion, allowing for its practical use case in real world scenarios where information contributing to style is not available. Also, contrary to previous methods that assume a high-level conditioning vector for generation, in addition to using supervised learning to learn the content, our framework uses unsupervised learning to discover factors of variation not present in the conditioning vector.

The proposed model has a wide range of potential applications from improving the training of deep models by means of data augmentation to generating rare image cases for training of medical personnel.

In future work we would like to explore DRAI's ability to improve generalization of supervised learning through conditional synthetic image generation. It is also interesting to explore whether DRAI's ability to disentangle style and content can help to learn better representations of under-represented subgroups.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

[2] David Barber Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16:201, 2004.

[3] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *ICML*, pages 23–45, 2018.

[4] The Cancer Imaging Archive. Lung image database consortium - reader annotation and markup - annotation and markup issues/comments, https://wiki.cancerimagingarchive.net/display/public/lidc-idri, 2017.

[5] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, and et. al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical Physics*, 38(2):915–931, 2011.

[6] Duhyeon Bang and Hyunjung Shim. Improved training of generative adversarial networks using representative features. In *ICML*, 2018.

[7] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[8] Christoph Baur, Shadi Albarqouni, and Nassir Navab. Generating highly realistic images of skin lesions with gans. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 260–267. Springer, 2018.

[9] Neslihan Bayramoglu, Mika Kaakinen, Lauri Eklund, and Janne Heikkila. Towards virtual h&e staining of hyperspectral lung histology images using conditional generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 64–71, 2017.

[10] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540, 2018.

[11] Avi Ben-Cohen, Roey Mechrez, Noa Yedidia, and Hayit Greenspan. Improving cnn training using disentanglement for liver lesion classification in ct. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 886–889. IEEE, 2019.

[12] Alceu Bissoto, Fábio Perez, Eduardo Valle, and Sandra Avila. Skin lesion synthesis with generative adversarial networks. In *OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis*, pages 294–302. Springer, 2018.

[13] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical image analysis*, 58:101535, 2019.

[14] Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2610–2620. Curran Associates, Inc., 2018.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[16] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

[17] Pedro Costa, Adrian Galdran, Maria Inês Meyer, Michael David Abràmoff, Meindert Niemeijer, Ana Maria Mendonça, and Aurélio Campilho. Towards adversarial retinal image synthesis. *arXiv preprint arXiv:1701.08974*, 2017.

[18] Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abràmoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2017.

[19] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017.

[20] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *ICLR*, 2017.

[21] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[22] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *CoRR*, abs/1505.07818, 2015.

[23] Mélanie Garcia1, Jean-Marc Orgogozo1, Kelly Clare, and Margaux Luck. Towards autism detection on brain structural mri scans using deep unsupervised learning models. In *Proceedings of Medical Imaging meets NeurIPS workshop*, 2019.

[24] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *NIPS*, 2018.

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[26] John T Guibas, Tejpal S Virdi, and Peter S Li. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*, 2017.

[27] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477. Springer, 2016.

[28] Mohammad Havaei, Nicolas Guizard, Hugo Larochelle, and Pierre-Marc Jodoin. Deep learning trends for focal brain pathology segmentation in mri. In *Machine Learning for Health Informatics*, pages 125–148. Springer, 2016.

[29] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[30] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[32] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[33] Xiaodan Hu, Audrey G Chung, Paul Fieguth, Farzad Khalvati, Masoom A Haider, and Alexander Wong. Prostategan: Mitigating data bias via prostate diffusion imaging synthesis with generative adversarial networks. *arXiv preprint arXiv:1811.05817*, 2018.

[34] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K Moyo, Michael R Savona, Richard G Abramson, and Bennett A Landman. Synseg-net: Synthetic segmentation without target

modality ground truth. *IEEE transactions on medical imaging*, 38(4):1016–1025, 2018.

[35] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[36] Xiang Jiang, Liqiang Ding, Mohammad Havaei, Andrew Jesson, and Stan Matwin. Task adaptive metric space for medium-shot medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 147–155. Springer, 2019.

[37] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 4816–4827. PMLR, 2020.

[38] Dakai Jin, Ziyue Xu, Youbao Tang, Adam P Harrison, and Daniel J Mollura. Ct-realistic lung nodule simulation from 3d conditional generative adversarial networks for robust lung segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 732–740. Springer, 2018.

[39] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

[40] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017.

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[42] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[43] Thanard Kurutach, Aviv Tamar, Ge Yang, Stuart J Russell, and Pieter Abbeel. Learning plannable representations with causal infogan. In *Advances in Neural Information Processing Systems*, pages 8733–8744, 2018.

[44] Qicheng Lao, Mohammad Havaei, Ahmad Pesaranghader, Francis Dutil, Lisa Di Jorio, and Thomas Fevens. Dual adversarial inference for text-to-image synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7567–7576, 2019.

[45] Qicheng Lao, Xiang Jiang, Mohammad Havaei, and Yoshua Bengio. Continuous domain adaptation with variational domain-agnostic feature replay. *arXiv preprint arXiv:2003.04382*, 2020.

[46] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016.

[47] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. Alice: Towards understanding adversarial learning for joint distribution matching. In *NIPS*, 2017.

[48] Haofu Liao, Wei-An Lin, S Kevin Zhou, and Jiebo Luo. Adn: Artifact disentanglement network for unsupervised metal artifact reduction. *IEEE Transactions on Medical Imaging*, 39(3):634–643, 2019.

[49] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[50] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[51] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[52] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, pages 2391–2400. PMLR, 2017.

[53] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[54] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. In *arXiv preprint arXiv:1411.1784*, 2014.

[55] Tony CW Mok and Albert CS Chung. Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks. In *International MICCAI Brainlesion Workshop*, pages 70–80. Springer, 2018.

[56] Utkarsh Ojha, Krishna Kumar Singh, Cho-Jui Hsieh, and Yong Jae Lee. Elastic-infogan: Unsupervised disentangled representation learning in imbalanced data. *arXiv preprint arXiv:1910.01112*, 2019.

[57] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[58] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[59] Edward Raff and Jared Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198. IEEE, 2018.

[60] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

[61] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016.

[62] Mhd Hasan Sarhan, Abouzar Eslami, Nassir Navab, and Shadi Albarqouni. Learning interpretable disentangled representations using adversarial vaes. In Qian Wang,

Fausto Milletari, Hien V. Nguyen, Shadi Albarqouni, M. Jorge Cardoso, Nicola Rieke, Ziyue Xu, Konstantinos Kamnitsas, Vishal Patel, Badri Roysam, Steve Jiang, Kevin Zhou, Khoa Luu, and Ngan Le, editors, *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 37–44, Cham, 2019. Springer International Publishing.

[63] Shiwen Shen, Simon X Han, Denise R Aberle, Alex AT Bui, and William Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert systems with applications*, 128:84–95, 2019.

[64] Joel Shor. Tensorflow-gan (tf-gan): A lightweight library for generative adversarial networks, 2017. URL https://github.com/tensorflow/gan.

[65] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[66] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 2018.

[67] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[68] Ning Wang, Xianhan Zeng, Renjie Xie, Zefei Gao, Yi Zheng, Ziran Liao, Junyan Yang, and Qiao Wang. Unsupervised classification of street architectures based on infogan. 2019.

[69] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018.

[70] Junlin Yang, Nicha C Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 255–263. Springer, 2019.

[71] Junlin Yang, Nicha C Dvornek, Fan Zhang, Juntang Zhuang, Julius Chapiro, MingDe Lin, and James S Duncan. Domain-agnostic learning with anatomy-consistent embedding for cross-modality liver segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[72] Xin Yi and Paul Babyn. Sharpness-aware low-dose ct denoising using conditional generative adversarial network. *Journal of digital imaging*, 31(5):655–669, 2018.

[73] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

[74] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017.

[75] Xianwen Yu, Xiaoning Zhang, Yang Cao, and Min Xia. Vaegan: A collaborative filtering framework based on adversarial variational autoencoders. In *IJCAI*, pages 4206–4212, 2019.

[76] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[77] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. In *arXiv preprint arXiv:1710.10916*, 2017.

[78] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019.

[79] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.