Conditional Generation of Medical Images via Disentangled Adversarial Inference

Mohammad Havaei[1(✉)], Ximeng Mao[2], Yipping Wang[1], and Qicheng Lao[1,2]

[1] Imagia, Montreal, Canada
mohammad@imagia.com
[2] Montréal Institute for Learning Algorithms (MILA), Université de Montréal, Montreal, Canada

Abstract. 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 conditional generation process. 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. We conduct extensive qualitative and quantitative assessments on two publicly available medical imaging datasets (LIDC and HAM10000) and test for conditional image generation and style-content disentanglement. We also show that our proposed model (DRAI) achieves the best disentanglement score and has the best overall performance.

1 Introduction

In recent years, conditional generation of medical images has become a popular area for research using conditional Generative Adversarial Networks (cGAN) [20, 36]. One common pitfall of cGAN is that the conditioning codes are extremely high-level and do not cover nuances of the data. This challenge is exacerbated 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,

M. Havaei and X. Mao—Equal contribution.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-88210-5_5) contains supplementary material, which is available to authorized users.

S. Engelhardt et al. (Eds.): DGM4MICCAI 2021/DALI 2021, LNCS 13003, pp. 45–66, 2021.
https://doi.org/10.1007/978-3-030-88210-5_5

location, background information, etc. Learning disentangled representation of content and style allows us to control the detailed nuances of the generation process.

In this work, we consider two types of information to preside over the image domain: *content* and *style*, which by definition, are independent 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.

We impose two novel disentanglement constraints to facilitate this process: Firstly, we introduce a novel application of the *Gradient Reverse Layer* (GRL) [16] to minimize the shared information between the two variables. Secondly, we present a new type of self-supervised regularization to further enforce disentanglement; using content-preserving transformations, we *attract* matching content information, while *repelling* different style information.

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. 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.

2 Method

2.1 Overview

Let $\boldsymbol{t}$ be the conditioning vector associated with image $\boldsymbol{x}$. Using the pairs $\{(\boldsymbol{t}_i, \boldsymbol{x}_i)\}, i = 1, \ldots, N$, 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 $\boldsymbol{c}$ and style $\boldsymbol{z}$ 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 $\boldsymbol{t}$ but also the style/content disentanglement. An Illustration of DRAI is made in Fig. 1

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 Sect. B.2).

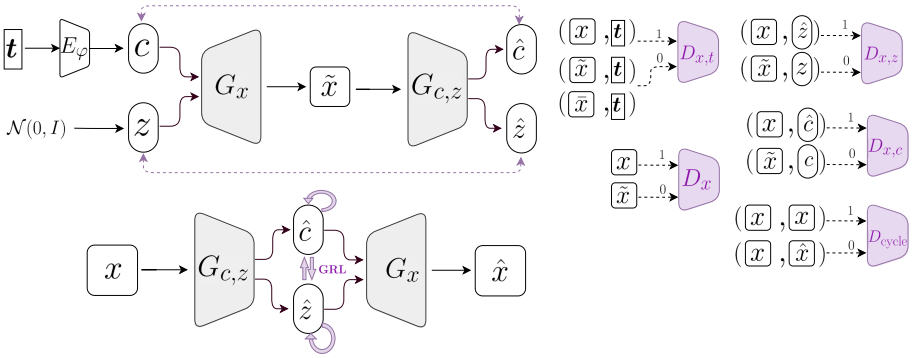


Fig. 1. Overview of DRAI. The dashed purple arrows mark the cycle consistency between features implemented via ℓ_1 norm, while the solid purple arrows show the imposed disentanglement constrains. On the right hand side of the figure we show all the discriminators used for training. $\hat{\boldsymbol{c}}$ represents the inferred content, $\hat{\boldsymbol{z}}$ the inferred style, $\hat{\boldsymbol{x}}$ the reconstructed input image and $\bar{\boldsymbol{x}}$ the image with mismatched conditioning.

2.2 Dual Adversarial Inference (DAI)

We follow the formulation of [30] for Dual Adversarial Inference (DAI) which is a conditional generative model that uses bidirectional adversarial inference [14,15] to learn content and style variables from the image data. To impose alignment between conditioning vector $\boldsymbol{t}$ and the generated image $\tilde{\boldsymbol{x}}$, we seek to match $p(\tilde{\boldsymbol{x}}, \boldsymbol{t})$ with $p(\boldsymbol{x}, \boldsymbol{t})$. To do so, we adopt the matching-aware discriminator proposed by [40]. For this discriminator—denoted as $D_{x,t}$—the positive sample is the pair of real image and its corresponding conditioning vector $(\boldsymbol{x}, \boldsymbol{t})$, whereas the negative sample pairs consist of two groups; the pair of real image with mismatched conditioning $(\bar{\boldsymbol{x}}, \boldsymbol{t})$, and the pair of synthetic image with corresponding conditioning $(G_x(\boldsymbol{z}, \boldsymbol{c}), \boldsymbol{t})$. 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:

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

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

We use adversarial inference to infer style and content codes from the image. 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 $\boldsymbol{z}$, can use the bidirectional adversarial inference formulation individually for style and content. This dual adversarial inference objective is thus formulated as:

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

To improve the stability of training, we include image-cycle consistency ($V_{\text{image-cycle}}$) [51] and latent code cycle consistency ($V_{\text{code-cycle}}$) objectives [12].

2.3 Disentanglement Constrains

The dual adversarial inference (DAI) encourages disentanglement through the independence assumption of style and content. However, it does not explicitly penalize entanglement. We introduce two constraints to impose style-content disentanglement. Refer to the Appendix for details.

Content-Style Information Minimization: We propose a novel application of the Gradient Reversal Layer (GRL) strategy [16] to *explicitly* minimize the shared information between style and content. 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. This constrains the content feature generation to disregard style features and the style feature generation to disregard content features.

Self-supervised Regularization: We incorporate a self-supervised regularization such that the content is invariant to content-preserving transformations (such as a rotation, horizontal or vertical flip) while the style is sensitive to such transformations. More formally, we maximize the similarity between the inferred contents of $\boldsymbol{x}$ and the transformed $\boldsymbol{x}'$ while minimizing 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 the transformation attributes.

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

$$\min_G \max_{D,F} V_{\text{t2i}} + V_{\text{dALI}} + V_{\text{image-cycle}} + V_{\text{code-cycle}} + V_{\text{GRL}} + V_{\text{self}} + \lambda D_{KL}(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})) + \lambda D_{KL}(q(\boldsymbol{c}|\boldsymbol{x}) \,||\, \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})) + \lambda D_{KL}(q(\boldsymbol{c}|\boldsymbol{t}) \,||\, \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})). \quad (2)$$

3 Experiments

We conduct experiments on two publicly available medical imaging datasets: LIDC [4] and HAM10000 [46] (see Appendix for details on these datasets). 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 [12] referred to as cInfoGAN and a conditional adversarial variational Autoencoder (cAVAE). We also compare DRAI to Dual Adversarial Inference (DAI) [30] and show how using our proposed disentanglement constraints together with latent code cycle-consistency can significantly boost performance. See Appendix for more details on various baselines. Finally, we conduct rigorous ablation studies to evaluate the impact of each component in DRAI.

3.1 Generation Evaluation

To evaluate the quality and diversity of the generated images, we measure FID and IS (see Appendix Sect. D.3) for the proposed DRAI model and various double and single latent variable baselines described in Appendix Sect. D. The results are reported in Table 1 for both LIDC and HAM10000 datasets. For the LIDC dataset, we observe all methods have comparable IS score while DRAI and DAI have significantly lower FID compared to 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.

Table 1. Comparison of image generation metrics (FID, IS) and disentanglement metric(CIFC) on HAM10000 and LIDC datasets for single and double variable baselines. CIFC is only evaluated for double variable baselines.

Method	HAM10000			LIDC		
	FID (↓)	IS (↑)	CIFC (↓)	FID (↓)	IS (↑)	CIFC (↓)
cInfoGAN	1.351 ± 0.33	1.326 ± 0.03	–	0.283 ± 0.06	1.366 ± 0.02	–
cAVAE	3.566 ± 0.56	1.371 ± 0.01	–	0.181 ± 0.03	$\mathbf{1.424 \pm 0.01}$	–
D-cInfoGAN	1.684 ± 0.42	$\mathbf{1.449 \pm 0.03}$	1.201 ± 0.17	0.333 ± 0.06	1.342 ± 0.09	1.625 ± 0.11
D-cAVAE	4.893 ± 0.99	1.321 ± 0.01	1.354 ± 0.03	0.378 ± 0.03	1.371 ± 0.04	1.944 ± 0.02
DAI [30]	1.327 ± 0.06	1.304 ± 0.01	0.256 ± 0.01	$\mathbf{0.106 \pm 0.02}$	$\mathbf{1.423 \pm 0.05}$	1.096 ± 0.28
DRAI	$\mathbf{1.224 \pm 0.05}$	1.300 ± 0.01	$\mathbf{0.210 \pm 0.01}$	$\mathbf{0.089 \pm 0.02}$	$\mathbf{1.422 \pm 0.03}$	$\mathbf{0.456 \pm 0.06}$

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 [5] and thus qualitative assessment is needed. We use the provided conditioning vector for the generation process and only sample the style variable $\boldsymbol{z}$. The generated samples are visualized in Fig. 2. 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.

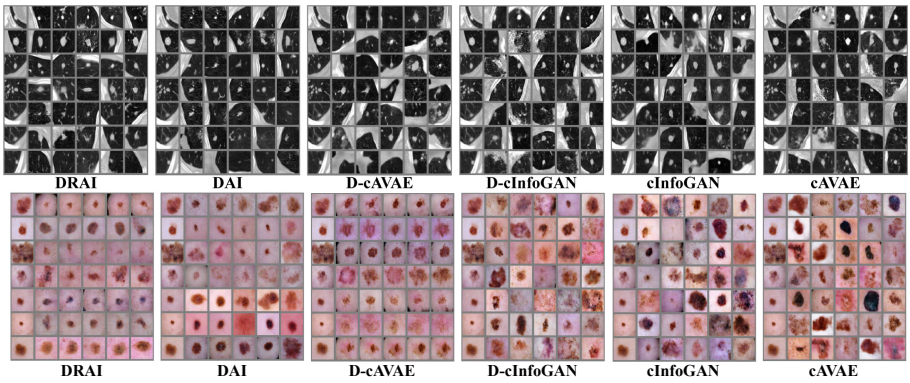


Fig. 2. Conditional generations on LIDC and HAM10000. The images are generated by keeping the content code ($\boldsymbol{c}$) fixed and only sampling the style codes ($\boldsymbol{z}$).

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.*, $\boldsymbol{c}$ 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.

3.2 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.

As a quantitative metric, we introduce the disentanglement error CIFC (refer to Appendix for details). Table 1 shows results on this metric. As seen from this table, in both HAM10000 and LIDC datasets, 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 the learned features but also facilitates disentanglement. In the ablation studies (Sect. 3.3), we investigate the effect of the individual components of DRAI on disentanglement.

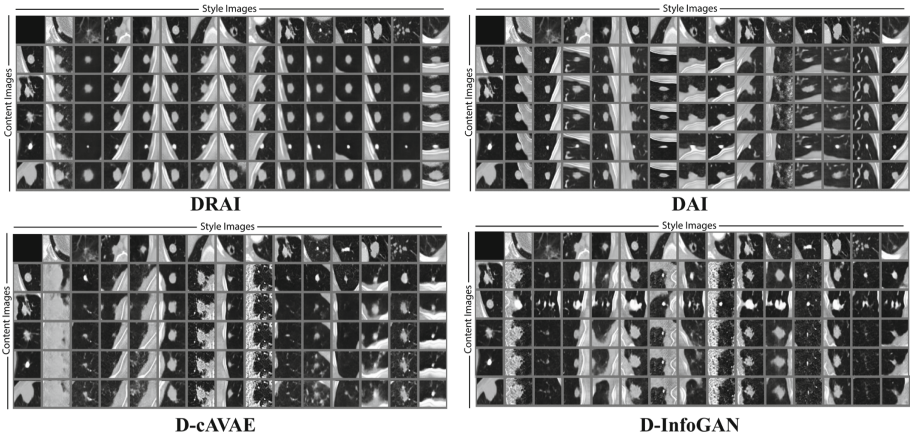


Fig. 3. Qualitative evaluation of style-content disentanglement through hybrid image generation on LIDC dataset. In every sub-figure, images in the first row present style image references and those in the first column present content image references. Hybrid images are generated by using the style and content codes inferred from style and content reference images respectively.

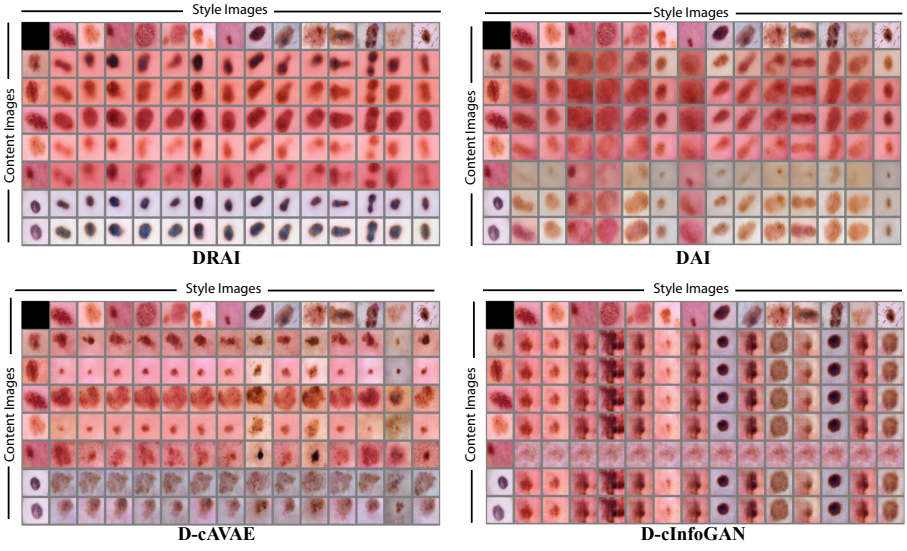


Fig. 4. Qualitative evaluation of style-content disentanglement through hybrid image generation on HAM10000 dataset. In every sub-figure, images in the first row represent style image references and those in the first column represent content image references. Hybrid images are generated by using the style and content codes inferred from style and content reference images respectively.

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 (See Appendix for details on hybrid images). 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 3 and Fig. 4 show these results on the two datasets. 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. 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. Overall, it is evident from the qualitative experiments that the proposed disentanglement regularizations help to decouple the style and content variables.

3.3 Ablation Studies

In this section, we perform ablation studies to evaluate the effect of each component on disentanglement using the CIFC metric. Ablated models use the same architecture with the same amount of parameters. The quantitative assessment is presented in Table 2. 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.

Table 2. Quantitative ablation study on LIDC and HAM10000 datasets

Method	LIDC		HAM10000	
	FID (↓)	CIFC (↓)	FID (↓)	CIFC (↓)
DAI [30]	$\mathbf{0.106 \pm 0.02}$	1.096 ± 0.284	1.327 ± 0.06	0.256 ± 0.01
DRAI = DAI+selfReg+MIReg+featureCycle	$\mathbf{0.089 \pm 0.02}$	$\mathbf{0.456 \pm 0.069}$	$\mathbf{1.224 \pm 0.05}$	$\mathbf{0.210 \pm 0.01}$
DAI+selfReg+MIReg	0.176 ± 0.06	0.554 ± 0.185	1.350 ± 0.12	0.233 ± 0.01
DAI+featureCycle	0.221 ± 0.07	0.913 ± 0.074	1.367 ± 0.12	0.311 ± 0.01
DAI+MIReg	0.154 ± 0.04	0.747 ± 0.226	$\mathbf{1.298 \pm 0.12}$	$\mathbf{0.228 \pm 0.01}$
DAI+selfReg	0.208 ± 0.05	0.781 ± 0.203	1.347 ± 0.14	$\mathbf{0.219 \pm 0.04}$

4 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 [30], we show the impact of imposing the proposed disentanglement constraints over the content and style variables.

A Disentanglement Constrains

Lao et al. [30] 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.

A.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 [16], 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:

$$\begin{aligned} \min_G \max_F V_{\text{GRL}}(F_z, F_c, G_{c,z}) \\ = -\mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x}),(\hat{\boldsymbol{z}},\hat{\boldsymbol{c}})\sim q(\boldsymbol{z},\boldsymbol{c}|\boldsymbol{x})}[\|\hat{\boldsymbol{z}} - F_z(\hat{\boldsymbol{c}})\| + \|\hat{\boldsymbol{c}} - F_c(\hat{\boldsymbol{z}})\|]. \end{aligned} \tag{3}$$

This constrains the content feature generation to disregard style features and the style feature generation to disregard content features. Figure 5b shows a visualization of this module.

We can show that Eq. (3) minimizes 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

$$I(\boldsymbol{z};\boldsymbol{c}) = H(\boldsymbol{z}) - H(\boldsymbol{z}|\boldsymbol{c}). \tag{4}$$

Once again, following [2], we define a variational lower bound on $I(\boldsymbol{z};\boldsymbol{c})$ by rewriting the conditional entropy in (4) as:

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

and by extension:

$$I(\boldsymbol{z};\boldsymbol{c}) = H(\boldsymbol{z}) + \max_{F_z} \mathbb{E}_{\hat{\boldsymbol{c}}\sim q(\boldsymbol{c}|\boldsymbol{x})}[\log q(\boldsymbol{z}|\hat{\boldsymbol{c}})], \tag{5}$$

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 constant for F_z and $||\hat{\boldsymbol{z}} - F_z(\hat{\boldsymbol{c}})||$ corresponds to $-\log q(\boldsymbol{z}|\hat{\boldsymbol{c}})$, minimization of mutual information can be written as:

$$\min_G I(\boldsymbol{z};\boldsymbol{c}) = \min_G \max_{F_z} -\mathbb{E}_{\hat{\boldsymbol{c}}\sim q(\boldsymbol{c}|\boldsymbol{x}),\hat{\boldsymbol{z}}\sim q(\boldsymbol{z}|\boldsymbol{x})}[||\hat{\boldsymbol{z}} - F_z(\hat{\boldsymbol{c}})||], \tag{6}$$

which corresponds to Eq. (3).

A.2 Self-supervised Regularization

Self-supervised learning has shown great potential in unsupervised representation learning [11,21,39]. 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 Fig. 5a. The objective function for the self-supervised regularization is defined as:

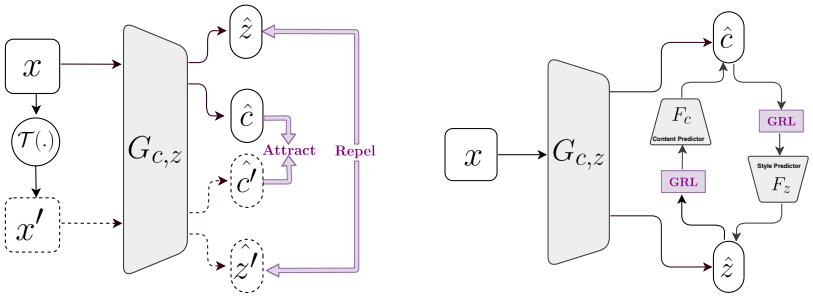
$$\min_G V_{\text{self}}(G_{c,z}) = \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\|\hat{\boldsymbol{c}} - \hat{\boldsymbol{c}}'\| - \|\hat{\boldsymbol{z}} - \hat{\boldsymbol{z}}'\|], \tag{7}$$

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

B Implementation Details

B.1 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



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

(b) Content-Style information minimization. For a given image $\boldsymbol{x}$, F_c is trained to predict the content $\hat{\boldsymbol{c}}$ from the style $\hat{\boldsymbol{z}}$. By reversing the direction of the gradients, the GRL penalizes $G_{c,z}$ to minimize the content information in the style variable $\boldsymbol{z}$. The same procedure is carried out to minimize style information in the content variable $\boldsymbol{c}$.

Fig. 5. Proposed disentanglement constraints.

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) [34] 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 (2) since other than the KL terms, they were all relatively on the same scale. As for the KL 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 [27] with initial learning rate $1e^{-5}$.

For IS and 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.

B.2 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 $\boldsymbol{t}$, 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 $\boldsymbol{t}_i$ and a style image $\boldsymbol{x}_j$: In this setup, we use the conditioning factor $\boldsymbol{t}_i$ as the content and the inferred $\hat{\boldsymbol{z}}_j$ from the style image $\boldsymbol{x}_j$ as the style:

$$\begin{aligned} \boldsymbol{c}_i &= E_\varphi(\boldsymbol{t}_i) \\ \hat{\boldsymbol{z}}_j, \hat{\boldsymbol{c}}_j &= G_{c,z}(\boldsymbol{x}_j) \\ \tilde{\boldsymbol{x}}_{ij} &= G_x(\hat{\boldsymbol{z}}_j, \boldsymbol{c}_i). \end{aligned}$$

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

$$\begin{aligned} \hat{\boldsymbol{z}}_i, \hat{\boldsymbol{c}}_i &= G_{c,z}(\boldsymbol{x}_i) \\ \hat{\boldsymbol{z}}_j, \hat{\boldsymbol{c}}_j &= G_{c,z}(\boldsymbol{x}_j) \\ \tilde{\boldsymbol{x}}_{ij} &= G_x(\hat{\boldsymbol{z}}_j, \hat{\boldsymbol{c}}_i) \end{aligned}$$

The generation of hybrid images is graphically explained in Fig. 6 for the aforementioned two scenarios.

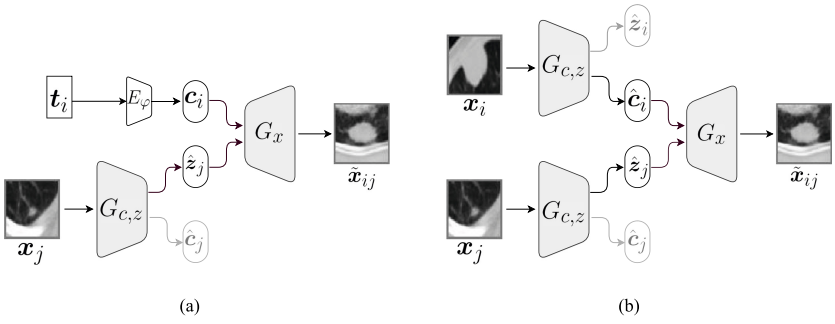


Fig. 6. Hybrid image generation: (a) via the conditioning factor $\boldsymbol{t}_i$ (representing content) and the style code $\hat{\boldsymbol{z}}_j$ inferred from the style reference image. (b) via the content code $\hat{\boldsymbol{c}}_i$ inferred from the content reference image and the style code $\hat{\boldsymbol{z}}_j$ inferred from the style reference image.

C Datasets

C.1 HAM10000

Human Against Machine (HAM10000) [46], 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.

C.2 LIDC

The Lung Image Database Consortium image collection (LIDC-IDRI) consists of lung CT scans from 1018 clinical cases [4]. 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 [3], 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. [43] 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.

D 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 [12] 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) [30] 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.

D.1 Conditional InfoGAN

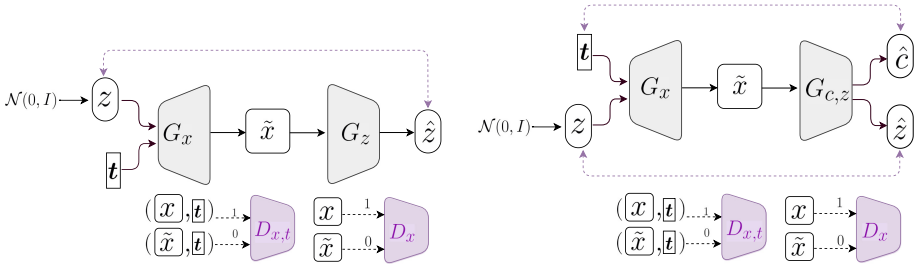
InfoGAN is a variant of generative adversarial network that aims to learn unsupervised disentangled representations. In order to do so, InfoGAN modifies the original GAN in two ways. First, it adds an additional input $\boldsymbol{c}$ to the generator. Second, using an encoder network Q, it predicts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

$$\min_{G,Q} \max_{D} V_{\text{InfoGAN}}(D, G, Q) = V_{\text{GAN}}(D, G) - \lambda(\mathbb{E}_{G(\boldsymbol{z},\boldsymbol{c}),p(\boldsymbol{c})}[\log Q(\boldsymbol{c}|\boldsymbol{x})] + H(\boldsymbol{c})). \quad (8)$$

The variable $\boldsymbol{c}$ can follow a discrete categorical distribution or a continuous distribution such as the normal distribution. InfoGAN is an unsupervised model popular for learning disentangled factors of variation [29,38,47].

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 Fig. 7a, 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 Fig. 7b where in addition to inferring $\hat{\boldsymbol{z}}$ we also infer $\hat{\boldsymbol{c}}$ through cycle consistency using the ℓ_1 norm.



(a) Conditional InfoGAN (cInfoGAN). (b) Dual conditional InfoGAN (D-cInfoGAN).

Fig. 7. InfoGan baselines.

D.2 cAVAE

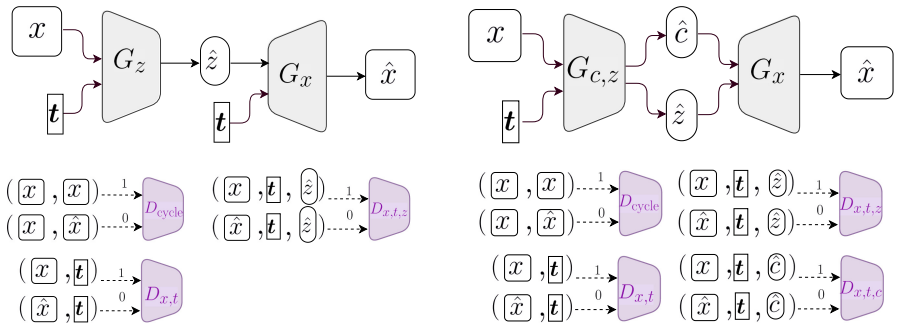
Variational Auto-Encoders (VAEs) [28] are latent variable models commonly used for inferring disentangled factors of variation governing the data distribution. Let $\boldsymbol{x}$ be the random variable over the data distribution and $\boldsymbol{z}$ the random variable over the latent space. VAEs are 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 $p(\boldsymbol{x})$:

$$\max_{G} V_{\text{VAE}}(G_x, G_z) \quad = \quad \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}|\boldsymbol{z})] \quad - \quad D_{KL}[q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z})]. \quad (9)$$

Kingma and Welling [28] 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. [31,35] and [50].

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 [32,35], 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 (Fig. 8a) and a double latent variable version D-cAVAE (Fig. 8b), 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 $D_{x,t,c}$.



(a) Conditional Adversarial VAE (cAVAE). (b) Dual conditional Adversarial VAE (D-cAVAE).

Fig. 8. Adversarial VAE baselines

D.3 Evaluation Metrics

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

Measure of Disentanglement (CIFC). Multiple methods have been proposed to measure the degree of disentanglement between variables [23]. 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 $(\hat{\boldsymbol{c}}_a, \hat{\boldsymbol{z}}_b)$ and $(\hat{\boldsymbol{c}}_b, \hat{\boldsymbol{z}}_a)$ 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:

$$CIFC = \mathbb{E}_{(I_a, I_b)\sim q_{\text{test}}(\boldsymbol{x})}[\|\hat{\boldsymbol{z}}_a - \hat{\boldsymbol{z}}_{aa}\| + \|\hat{\boldsymbol{c}}_a - \hat{\boldsymbol{c}}_{aa}\| + \|\hat{\boldsymbol{z}}_b - \hat{\boldsymbol{z}}_{bb}\| + \|\hat{\boldsymbol{c}}_b - \hat{\boldsymbol{c}}_{bb}\|], \tag{10}$$

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

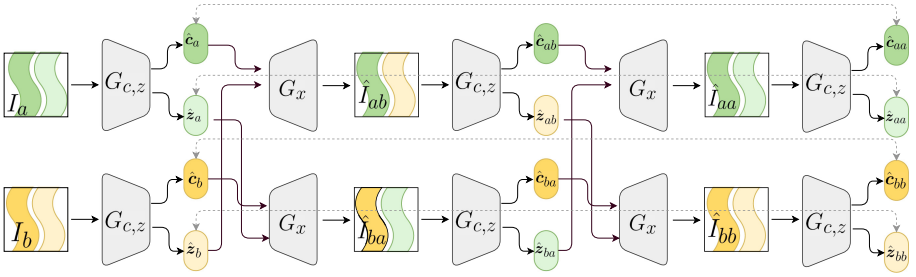


Fig. 9. Cross Image Feature Consistency (CIFC) error. CIFC is computed by first generating hybrid images of mixed style and content across two different images and then reconstructing the original images. For a more robust evaluation, CIFC is measured in the feature space.

FID. The Frechet inception distance (FID) score [22] 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 [45] up to layer "*mixed3*" (initialized with pretrained ImageNet weights), and append a global average pooling layer followed by a dense layer.

Inception Score. Inception Score (IS) [41] 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 [44] is used to calculate both FID and IS.

E Related Work

E.1 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 [13,13,19,26,37]. Jin et al. [26] 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. [19] 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. [13] 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 [37] 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. [8] uses the semantic segmentation of skin lesions and generate high resolution images. Their model combines the pix2pix framework [25] 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. [24] 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. [6] 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.

E.2 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. [18] 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. [49] aim to disentangle anatomical information and modality information in order to improve on a downstream liver segmentation task.

Ben-Cohen et al. [7] 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 [7] 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. [48] 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. [9] 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. [33], Given artifact affected CT images, metal-artifact reduction (MAR) is performed by disentangling the metal-artifact representations from the underlying CT images.

Sarhan et al. [42] use β-TCAV [10] 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. [17] 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.

References

1. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous systems (2015). https://www.tensorflow.org/
2. Agakov, D.B.F.: The IM algorithm: a variational approach to information maximization. Adv. Neural. Inf. Process. Syst. **16**, 201 (2004)
3. The Cancer Imaging Archive. Lung image database consortium - reader annotation and markup - annotation and markup issues/comments (2017). https://wiki.cancerimagingarchive.net/display/public/lidc-idri
4. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)

5. Barratt, S., Sharma, R.: A note on the inception score. arXiv preprint arXiv:1801.01973 (2018)
6. Baur, C., Albarqouni, S., Navab, N.: Generating highly realistic images of skin lesions with GANs. In: Stoyanov, D., et al. (eds.) CARE/CLIP/OR 2.0/ISIC - 2018. LNCS, vol. 11041, pp. 260–267. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01201-4_28
7. Ben-Cohen, A., Mechrez, R., Yedidia, N., Greenspan, H.: 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), pp. 886–889. IEEE (2019)
8. Bissoto, A., Perez, F., Valle, E., Avila, S.: Skin lesion synthesis with generative adversarial networks. In: Stoyanov, D., et al. (eds.) CARE/CLIP/OR 2.0/ISIC - 2018. LNCS, vol. 11041, pp. 294–302. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01201-4_32
9. Chartsias, A., et al.: Disentangled representation learning in cardiac image analysis. Med. Image Anal. **58**, 101535 (2019)
10. Chen, R.T.Q., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31, pp. 2610–2620. Curran Associates Inc. (2018)
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
12. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In: NIPS (2016)
13. Costa, P., et al.: Towards adversarial retinal image synthesis. arXiv preprint arXiv:1701.08974 (2017)
14. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. In: ICLR (2017)
15. Dumoulin, V., et al.: Adversarially learned inference. In: ICLR (2017)
16. Ganin, Y., et al.: Domain-adversarial training of neural networks. CoRR, abs/1505.07818 (2015)
17. Garcia1, M., Orgogozo, J.-M., Clare, K., Luck, M.: Towards autism detection on brain structural MRI scans using deep unsupervised learning models. In: Proceedings of Medical Imaging meets NeurIPS Workshop (2019)
18. Gonzalez-Garcia, A., van de Weijer, J., Bengio, Y.: Image-to-image translation for cross-domain disentanglement. In: NIPS (2018)
19. Guibas, J.T., Virdi, T.S., Li, P.S.: Synthetic medical images from dual generative adversarial networks. arXiv preprint arXiv:1709.01872 (2017)
20. Havaei, M., Mao, X., Wang, Y., Lao, Q.: Conditional generation of medical images via disentangled adversarial inference. Med. Image Anal. 102106 (2021)
21. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS (2017)
23. Higgins, I., et al.: Beta-VAE: learning basic visual concepts with a constrained variational framework. In: ICLR (2017)

24. Hu, X., Chung, A.G., Fieguth, P., Khalvati, F., Haider, M.A., Wong, A.: ProstateGAN: mitigating data bias via prostate diffusion imaging synthesis with generative adversarial networks. arXiv preprint arXiv:1811.05817 (2018)
25. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
26. Jin, D., Xu, Z., Tang, Y., Harrison, A.P., Mollura, D.J.: CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 732–740. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_81
27. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
28. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
29. Kurutach, T., Tamar, A., Yang, G., Russell, S.J., Abbeel, P.: Learning plannable representations with causal InfoGAN. In: Advances in Neural Information Processing Systems, pp. 8733–8744 (2018)
30. Lao, Q., Havaei, M., Pesaranghader, A., Dutil, F., Di Jorio, L., Fevens, T.: Dual adversarial inference for text-to-image synthesis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7567–7576 (2019)
31. Larsen, A.B.L., Kaae Sønderby, S., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: ICML (2016)
32. Li, C., et al.: ALICE: towards understanding adversarial learning for joint distribution matching. In: NIPS (2017)
33. Liao, H., Lin, W.-A., Zhou, S.K., Luo, J.: ADN: artifact disentanglement network for unsupervised metal artifact reduction. IEEE Trans. Med. Imaging **39**(3), 634–643 (2019)
34. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802 (2017)
35. Mescheder, L., Nowozin, S., Geiger, A.: Adversarial variational bayes: unifying variational autoencoders and generative adversarial networks. In: International Conference on Machine Learning, pp. 2391–2400. PMLR (2017)
36. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
37. Mok, T.C.W., Chung, A.C.S.: Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11383, pp. 70–80. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11723-8_7
38. Ojha, U., Singh, K.K., Hsieh, C.-J., Lee, Y.J.: Elastic-InfoGAN: unsupervised disentangled representation learning in imbalanced data. arXiv preprint arXiv:1910.01112 (2019)
39. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
40. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016)
41. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: NIPS (2016)

42. Sarhan, M.H., Eslami, A., Navab, N., Albarqouni, S.: Learning interpretable disentangled representations using adversarial VAEs. In: Wang, Q., et al. (eds.) DART/MIL3ID -2019. LNCS, vol. 11795, pp. 37–44. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33391-1_5
43. Shen, S., Han, S.X., Aberle, D.R., Bui, A.A.T., Hsu, W.: An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Syst. Appl. **128**, 84–95 (2019)
44. Shor, J.: TensorFlow-GAN (TF-GAN): a lightweight library for generative adversarial networks (2017). https://github.com/tensorflow/gan
45. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016)
46. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**, 1–9 (2018)
47. Wang, N., et al.: Unsupervised classification of street architectures based on InfoGAN (2019)
48. Yang, J., Dvornek, N.C., Zhang, F., Chapiro, J., Lin, M.D., Duncan, J.S.: Unsupervised domain adaptation via disentangled representations: application to cross-modality liver segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 255–263. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_29
49. Yang, J., et al.: Domain-agnostic learning with anatomy-consistent embedding for cross-modality liver segmentation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2019)
50. Yu, X., Zhang, X., Cao, Y., Xia, M.: VAEGAN: a collaborative filtering framework based on adversarial variational autoencoders. In: IJCAI, pp. 4206–4212 (2019)
51. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)