Manipulating Medical Image Translation with Manifold Disentanglement

1st Siyu Liu  
School of EECS  
The University of Queensland  
Queensland, Australia

2nd Jason A. Dowling  
The Australian e-Health Research Centre  
CSIRO Health and Biosecurity  
Queensland, Australia

3rd Craig Engstrom  
School of EECS  
The University of Queensland  
Queensland, Australia

4th Peter B. Greer  
School of Mathematical and Physical Sciences  
University of Newcastle  
New South Wales, Australia

5th Stuart Crozier  
School of EECS  
The University of Queensland  
Queensland, Australia

6th Shekhar S. Chandra  
School of EECS  
The University of Queensland  
Queensland, Australia

Abstract—Most current medical image translation Generative Adversarial Networks (GANs) can only generate one definitive output for each input. However, in reality, there can be many simultaneously valid translations. For example, when synthesising medical images from segmentation, diverse tissue structures, contrasts, textures, and even modalities are possible based on the same segmentation. In this work, we propose Manifold Disentanglement Generative Adversarial Network (MDGAN), a style-based network capable of capturing this output diversity. The mechanism enabling output diversity is a style-based manifold, which is learnt from image data, and can be sampled to “style” the input into diverse outputs. We train MDGAN for segmentation-to-MR-and-CT translation and show that the manifold i) can learn distinct clusters that control the output modality (CT or MR), ii) can be traversed to smoothly alter features within each modality (such as tissue structures, contrasts in MR) and iii) is disentangled such that the input’s anatomical structures are faithfully preserved when generating diverse images based on the same segmentation map.

Index Terms—GAN, medical image translation, multi-modal

I. INTRODUCTION

Cross-modality medical image translation has the potential to aid disease treatment procedures. For example, in magnetic resonance (MR)-only radiotherapy treatment [1], a conditional GAN can be used to “retrieve” (generate) missing computed tomography (CT) images from other available imaging modalities. Indeed, in areas facing data scarcity, medical image translation provides a ready avenue be used for data augmentation by artificially expanding the training set size [2], [3].

Conditional GANs [4] have been widely applied to medical image translation as they provide a controllable way to synthesise realistic images. A typical image translation GAN uses a fully convolutional generator $G$ to perform the source-to-target mapping. In this process, $G$ must correctly synthesise images with i) the correct anatomical features from the input and ii) valid target domain features (style). In most GANs for medical image translation, these anatomical features and target-domain features are intertwined and, thus, cannot be individually modified. As such, for each input, only one definitive translation could be acquired [5], [6]. However, in reality, there can be multiple equally valid translations for the same input. For example, one CT image could be mapped to multiple valid MR images (with different soft-tissue structures but with the same anatomical structures as shown in Figure 1). This is because CT has inherently lower contrast for soft-tissue expressions (compared to MR). Hence, a good GAN framework should be capable of generating diverse outputs which capture this natural ambiguity.

![Fig. 1. Paired pelvic MR and CT images with varying soft tissue mappings.](image)

The StyleGAN [7], [8] framework is a natural candidate for such diverse medical image translation tasks. A StyleGAN generator relies on style code injection to manipulate the output, and the style codes essentially form a manifold that controls the output. We can logically formulate our objective as a stylisation problem. As in Figure 2, a shared convolutional encoder-decoder generator network could be used to extract and retain anatomical features from the input. At the same time, a learnt manifold could supply style codes to “style” (translate) the input into diverse outputs. Based on this idea, we propose MDGAN, a powerful style-based generative framework for medical image translation. The contribution of this framework can be summarised as:

- We harness the StyleGAN framework to learn a medical image manifold of domain features from segmentations. These features are encoded within the style code network,
and the sampled style codes are used to diversify the output medical images.

- The sampled style codes can diversify not only the modality (e.g., CT or MR) of the output but also the fine features (e.g., tissue variations) within each modality.
- Anatomical structures and target-domain features are disentangled such that output diversity is achieved while adhering to the input anatomical features.
- The manifold is naturally smooth and explorable. We can interpolate between two style codes to smoothly transition between their corresponding outputs.

We use MDGAN to translate from segmentation maps to diverse CT and MR images. A shared generator is used to preserve the anatomical features from the segmentation maps, and a style code network is used to encode diverse CT and MR features as a manifold. Performing dimensionality reduction on sampled style codes to reveal two well-formed manifold clusters (for MR and CT). By interpolating between two style codes from the MR manifold cluster, we observed smooth and systematic transitions in tissue appearances. Throughout the transition, the anatomical structures of the segmentation map remain unaltered.

II. RELATED WORK

A. Image Synthesis in Medical Imaging Analysis

Both conditional GANs and unconditional GANs have been widely adopted for medical imaging synthesis [9]. Medical image synthesis using GANs have been shown to be effective for data augmentation, which may enhance existing deep learning models facing data scarcity issues. For example, [2] uses GAN-generated CT scans to improve segmentation accuracy, and [3] uses three generative networks to synthesise three types of liver lesions to improve classification accuracy. Conditional GAN is useful for many applications beyond data augmentation, thanks to its versatility. The most common application of conditional GANs is image domain translation [10]. For example, segmentation map to medical image [11], [12]. MR reconstruction [13], [14], image denoising [15], [16] and cross-modality translation [17]–[19]. Many of these methods are based on the popular Pix2Pix framework, which relies on paired data across domains. When pair-wise labels are not available, CycleGAN [6] and UNIT [20] are used for semi-supervised learning on unpaired images [21]–[25].

As GAN techniques keep improving, most of GANs in recent years are capable of realistic medical images that are difficult to distinguish from the real ones. However, most methods, especially in medical image analysis overlook image diversity. In most cases, the mapping between two modalities is not unique as they are adept at capturing different features. As such, a GAN framework capable of controlling image diversity is desirable. It is worth noting that while diffusion models have been [26], [27] have been rising in popularity for conditional and unconditional image generation, the image generation process tends to be too chaotic for medical image translation - they still cannot reliably generate consistent video frames [28] not to mention consistent anatomical structures.

B. Style-based GAN

Style-based GAN [7] is an unconventional GAN approach to image synthesis. It rethink image synthesis from the perspective of style transfer. Huang et al. [29] explored the profound effect of activation normalisation in Convolutional Neural Network (CNN) and proposed Adaptive Instance Normalisation (AdaIN) as a means of real-time style transfer. In comparison to traditional gradient-based style transfer methods [30], AdaIN exhibits superior versatility and control of the output style while keeping the content of the image consistent. StyleGAN [7] was the first to harness the power of AdaIN in a generative adversarial framework. Instead of starting with a latent noise vector, the StyleGAN generator applies AdaIN at various points of the network to inject a learnt latent (style) vector. This alternative approach to image generation achieves unprecedented control of the generated image at all scales (from global structure to local details). A very important property of the style vectors is that they form a high dimensional surface (manifold). Interpolating between any two points on this manifold can smoothly transform between the two associated output images. Other technical features include a progressively growing training scheme, mini-batch standard deviation [31], path length regularisation, and style-mixing regularisation to optimise performance and stabilise training. The successor to StyleGAN, StyleGAN2 [8], was published with various significant refinements to StyleGAN including i) First, AdaIN was removed in favour of modulated convolution to alleviate normalisation artefacts in the output, ii) the progressively growing networks were simplified to residual architectures for easier training and iii) a new path length regularisation was used to improve image quality and network invertibility. The main limitation of the original StyleGAN and StyleGAN2 is their lack of an image input, which is unsuitable for image translation tasks. Recently, Pixel2Style2Pixel [32] proposed an image encoder network as an extension to the
StyleGAN framework. The encoder network maps the inputs to codes, which enables domain translation using the StyleGAN framework. Another prominent conditional extension to StyleGAN is StarGAN [33], which employs AdaIN and cycle consistency to achieve unpaired multi-domain translation.

C. Limitations of current works

Compared to conventional a GAN, the StyleGAN framework introduces a more powerful approach to manipulating outputs using an external style input. As described in the Introduction, it fits naturally within our objective of creating a versatile medical image translation framework based on feature disentanglement. However, the original StyleGAN framework is fundamentally unconditional, and its manifold does not provide disentanglement of domain-related features. While Pixel2Style2Pixel is one step closer to our objective due to the addition of a conditional input and its diverse outputs, it does not provide explicit disentangled of anatomical features and target-domain manifold for independent control. At the same time, it is not designed for multi-modal applications. StarGAN has also been considered a candidate for our task as it is a powerful multi-domain image translation network. However, it is more tailored towards imposing the style of one image onto another, which does not provide the best control of the output.

Most of the methods in the medical imaging context only perform one-to-one mappings on a given input, which does not capture the true dynamics of image translation tasks. For example, one segmentation map can be theoretically mapped to infinitely many valid target images. We are also not aware of other work that formulates multi-domain medical image translation as a general stylisation problem with disentangled manifolds. The closest use of a style-based GAN in medical translation as a general stylisation problem with disentangled features. UNIT can be more desirable for some applications, but to the best of our knowledge, we are also not aware of any work that formulates multi-domain medical image translation as a general stylisation problem with disentangled manifolds.

III. METHODS

The objective of the proposed method is to achieve disentangled representations of anatomical features and target-domain features, where the anatomical features are translated to the target domain according to style codes sampled from learnt manifold clusters. In this section, we formulate a framework for one-to-many medical image translation using this idea.

A. Proposed Framework

Improving upon the foundation of StyleGAN, the proposed MDGAN consists of four networks: conditional encoder (E), style-based image synthesiser (G), style code network (S) and discriminator (D). E and G are shared networks that encapsulate anatomical information (such as anatomical structure and shape information). Most of S is also shared, but for each target domain (modality) T, a modality vector embedding e_T and a discriminator (D_T) are learnt to control the output modality.

The network component interactions are illustrated in Figure 3. Given a source domain input x and a desired target domain T, E produces a shared latent representation of the input w = E(x) which learns anatomical features, G is supplied both w and a target-domain style code s_T = S(z_{noise} ∼ N(0,1), e_T) as conditional inputs to synthesise the translated image y_T = G(w, s_T). s_T is conditioned on the domain embedding e_T and a random noise vector z which control the output modality and the diversity within each modality, respectively. It is also important to note that w and s_T are deliberately separated to disentangle shared features and target-domain features. The role of s_T is to modulate the convolutional weights in G to achieve the desired output style. Compared to a random noise vector, s_T is more interpretable to G as it is fundamentally a manifold cluster of the features specific to domain T. Finally, the discriminator for each target domain is a binary classifier D_T which aims to distinguish real data y_T from the fake y_T.

For each training step, we randomly sample a target domain t ∈ T and update the weights in e_t, D_t along with the shared image synthesiser G and conditional encoder E.

B. Network Architecture Details

S is a fully-connected network with four 384-unit hidden layers. It learns a mapping from a modality embedding (or class label) to a modality-specific style code s_T. E is a fully-convolutional network and D is a re-implemented StyleGAN2 [8] with conditional inputs. E contains four convolutional blocks (16, 16, 32 and 64 filters), which progressively down-sample the input while expanding the feature depths. G contains convolutional blocks of depths 256, 128, 64 and 48, and its structure mirrors that of E to recover the original scale of the input. Residual connections [8] are used in both E and G to improve the connectivity between neighbouring blocks. Like the original StyleGAN, we also incorporate noise feature maps in G to introduce fine-grained variations. D_T uses a similar structure to E but the filter depths are increased to 48, 64, 128 and 256. The final feature maps of D_T are mapped to a confidence score using a densely connected layer.

All of the convolutional layers in G are modulated convolution as used in StyleGAN2. As below, modulated convolution performs weight re-normalisation based on some affine-transformed external style vector. In the proposed framework, this re-normalisation procedure provides the mechanism for freely switching among multiple target domains as well as producing diverse outputs.

\[ w'_{ijk} = (s_i \cdot w_{ijk})/\sqrt{\sum_{i,k} (s_i \cdot w_{ijk})^2 + \epsilon} \]

Like the modulated convolution in StyleGAN2, s_i comes from an external style input, i, j and k enumerate the input feature maps, output feature maps and spatial dimensions, respectively.
Fig. 3. Network architectures for the proposed framework. The generative networks $E$, $S$ and $G$ are shared. $G$ uses modulated convolution (denoted $M$) for stylisation based on some external style code. The multiple domain embeddings control the final output modality (e.g. MR or CT). The noise vector add further diversity within each domain.

C. Proposed Training Procedure

The primary losses of the proposed framework include a non-saturating adversarial loss $L_{adv}$ [35], a perceptual reconstruction loss $L_{rec}$ and an $H_1$ gradient penalty $L_{gp}$ [36] term:

$$L_{adv} = \mathbb{E}[\log D_T(y_T)] - \mathbb{E}[\log (D_T(G(E(x), z, e_T)))$$

$$L_{rec} = \sum_{i=3}^n ||\phi_i(G(E(x), z, e_T)) - \phi_i(y_T))||^2$$

$$L_{gp} = \mathbb{E}[||\nabla D_T(y_T)||^2]$$

When computing $L_{rec}$, the perceptual network for target domain $T$ is its corresponding discriminator $D_T$. As Figure 4, $L_{rec}$ captures the perceptual difference based on the intermediate outputs (denoted $\phi$) of $D_T$ from the third convolution layer onward.

Most GAN frameworks used for image translation are one-to-one mapping networks, which arguably resemble the undesirable effect of mode collapse. With the MDGAN framework, we can achieve one-to-many image translation by ensuring the output is diverse and valid. This is done by imposing an additional diversification regulariser (similar to [37]) as below to ensure the output is well-conditioned on the style codes. Our experiments show that the model tends to collapse and produce similar outputs without this regulariser.

$$L_{div} = - ||G(E(x), z_1, e_T)) - G(E(x), z_2, e_T))$$

The total loss is finally defined as follows, $\mu$ and $\lambda$ are scaling factors for the reconstruction loss and gradient penalty term, respectively. $\lambda_{div}$ is the scaling factor for the diversification loss. We use $\lambda_{div} = 1$, and unlike StarGAN2 [33], we avoid using decay as it results in mode collapse in our case.

$$L_{total} = L_{adv} + \mu L_{rec} + \lambda L_{gp} + \lambda_{div} L_{div}$$

For training, we use Adam [38] optimiser with a learning rate of 0.0001 for $G$ and $D$, and a learning rate of 0.000001 for all the other networks. The models are trained for 72 hours on an NVIDIA P100 GPU equipped with 16GB of VRAM to process a batch of 8 images at a time.

D. Experiment

We test the MDGAN framework by performing domain translation from (multi-object) image-based segmentation(s) to MR and CT scans. The dataset is a manually segmented 3D prostate dataset with 211 MR and 42 CT scans. The scans are from a prostate cancer treatment study of 42 patients over the
course of 8 weeks [1]. The 3D images are $256 \times 256 \times 128$ in resolution and is manually labelled with five foreground classes: body, bone, bladder, rectum and prostate. During training, we randomly sample $256 \times 128$ images from the centre 40 slices of the coronal plane. All the input images are prepossessed by mapping their pixel intensity ranges to $[0, 1]$. We train two instances of MDGAN on the coronal and axial slices, respectively to cross verify any findings. For comparison, we compare against baseline methods including CycleGAN, UNIT [20], MUNIT [39] and StarGANv2.

Our framework takes the segmentation maps (one-hot encoded) as the input and $G$ stylises them to arrive at the target domains $T = \{CT, MR\}$. As described in III-A, the generative part of the framework only requires two separately learnt embeddings $e_{CT}$ and $e_{MR}$, which are inexpensive to train. The expensive components $E$, $G$ and $S$ are fully shared. $D_{CT}$ and $D_{MR}$ are also separately trained networks, but they do not contribute to inference and can be discarded after training.

### IV. RESULTS

In this section, we present the results for the segmentation-to-MR-CT translation task.

#### A. Quantitative Results

The two main quality metrics we focus on are image quality (measured with FID [40] and PSNR [1]) and image diversity (measured with pixel-wise standard deviation (Avg Pixel Std) [2]). These metrics of MDGAN are compared with the baseline methods in Table I. The results suggest that MDGAN produces comparable image quality compared to other state-of-the-art GAN methods. The STD values indicate MDGAN produced substantially more diverse outputs than the baselines especially StarGAN v2. MUNIT also produced relatively more diverse outputs. However, it comes at performance trade-offs in FID (Frechet Inception Distance) and PSNR (peak signal-to-noise ratio).

#### B. Qualitative Results

To test MDGAN for output diversification, we perform image translation on a given input segmentation in combination with different style codes. The explicit disentanglement of anatomical and target-domain features allows us to “edit” the tissues in the generated MR while keeping the mutual shape information intact. And in our experiments, the types of diversification enabled by MDGAN includes modality (MR and CT) and diversification within each modality. We also observe that scaling factor $\lambda_{div}$ has a positive association with the magnitude of diversification with larger values of $\lambda_{div}$ taking significantly longer to converge and may occasionally produce invalid outputs.

Figure 5 presents example CT and MR images generated using MDGAN. All the generated images are acquired using a shared generator, and only the style code network input label was altered to change the output modality. It can be seen that the generated MR and CT outputs retain consistent shape information from the input segmentation maps. MDGAN was also able to generate diverse images within each domain of MR and CT. Figure 6 shows example diverse MRs generated from the same segmentation maps (we choose to show MR as the features are more distinct for visualisation). It can be seen that the anatomical structures are maintained while ambiguous features like tissues and contrast are free to change. Figure 7 shows similar results but in a different imaging plane (axial), where similar diverse results were acquired. In Figure 7, we further compare MDGAN against StarGANv2 and MUNIT (baselines). According to the pixel-wise STD maps, MDGAN produced the most diverse images, and the most diverse regions are in soft tissues (bright regions in the STD maps). MUNIT also produced relatively more diverse outputs. However, it comes at performance trade-offs in FID (Frechet Inception Distance) and PSNR (peak signal-to-noise ratio).

#### C. Manifold of MR Features

Finally, we perform dimensionality reduction on the style codes to explore the learnt manifold of each domain. This is done by sampling 10,000 styles codes from $S_{MR}$ and $S_{CT}$ each and mapping them to 2D space using UMAP [41] (minimum distance of 0.2 and 5 neighbours). Manifold interpolation was performed on the MR style codes (instead of CT because MR scans contain more complex features) to observe visual transitions in the MR domain. For a given segmentation input, a geodesic path with 36 points across the manifold is selected, and 36 images are generated.

Figure 8 presents the results of the MR manifold geodesic path or “walk”. As shown, the style codes of MR (orange) and CT (blue) are embedded as two separate manifold clusters on the manifold of the generator. The clear separation between the two clusters acts as a boundary to explicitly prevent “feature mix-ups” between the two exclusive domains. This suggests the proposed style-based generator is capable of learning and interpreting multiple medical imaging manifold clusters for different imaging modalities.

| Metrics | PSNR | FID | Avg Pixel STD | Diversity |
|---------|------|-----|---------------|-----------|
| CycleGAN [6] | 12.7 | 72.9 | N/A | No |
| UNIT [20] | 18.4 | 49.3 | N/A | No |
| MUNIT [39] | 17.5 | 93.5 | 5.9 | Yes |
| StarGAN v2 [33] | 20.1 | 39.5 | 3.1 | Yes |
| MDGAN | 21.2 | 23.19 | 7.7 | Yes |

$^1$MR and CT combined

$^2$For each validation input segmentation, the pixel-wise STD was averaged over 100 diverse outputs. These pixel-wise STD maps were then averaged again across all the samples to produce the diversity metric in Table I.
Traversing the chosen geodesic path in the MR manifold (green), the images generated using the style code sequence (and a chosen segmentation map) show smooth and systematic transitions. We observe that these transitions result in consistent and meaningful changes in tissue structure for all valid segmentation maps. Examples of this finding are shown in Figure 9, and the changes before and after the transition are highlighted in colour. The style codes appear to have a similar global influence on all segmentation maps, even for the heavily distorted case (Figure 7E) that was not part of the training set, though the features are localised differently in each image due to the shape and validity constraints. $L_{div}$ plays a profound role in forming these well-distributed manifold clusters. Our experiments without this term resulted in mode collapse, where all style codes produce the same output for a given input and chosen target domain. Future work will involve using the proposed manifold disentanglement to construct meaningful manifolds in order to understand human diseases via MR and CT images from large studies. For example, we could train a MDGAN on an osteoarthritis dataset and explore how disease features are distributed on the manifold.

V. CONCLUSION

In this paper, we introduce MDGAN as a style-based framework for medical image domain translation. Besides its robust generative performance, the framework explicitly models anatomical features and target-domain features. We model anatomical features with a fully convolutional network and target-domain features as a disentangled manifold. We embed two manifold clusters onto the manifold using two style code manifold networks, which provide style codes for multi-modal (segmentation to MR and CT) medical image translation. These manifold clusters are found to determine the output modality and the target-domain features because our model is explorable. This valuable property could facilitate the detailed manifold learning of human diseases investigated with radiological techniques such as MR imaging.
Fig. 7. Diverse MR generated using MDGAN, MUNIT, and StarGAN v2 based on the same source CT. The pixel-wise standard deviation maps indicate MDGAN produced more diverse results than StarGAN v2 and MUNIT. Red highlights unrealistic-looking soft tissues produced by MUNIT.

Fig. 8. Left: 2D UMAP manifold mapping from 10,000 MR (orange) and 10,000 CT (blue) style codes. The manifold clusters are naturally separated in 2D. The geodesic path chosen to explore the MR manifold is indicated in green. Right: images generated using the 1st, 9th, 16th, 27th and 36th points from the manifold path.

Fig. 9. Change in tissue structures before (n=1) and after (n=36) transition.

REFERENCES

[1] J. A. Dowling, J. Sun, P. Pichler, D. Rivest-Hénault, S. Ghose, H. Richardson, C. Wratten, J. Martin, J. Arm, L. Best, S. S. Chandra, J. Fripp, F. W. Menk, and P. B. Greer, “Automatic substitute CT generation and contouring for MRI-alone external beam radiation therapy from standard MRI sequences,” International Journal of Radiation Oncology Biology Physics, vol. 93, no. 5, pp. 1144–1153, 2015.

[2] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, “Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks,” Scientific Reports, vol. 9, no. 1, p. 16884, Nov 2019. [Online]. Available: https://doi.org/10.1038/s41598-019-52737-x
