Unified Brain MR-Ultrasound Synthesis Using Multi-modal Hierarchical Representations

Reuben Dorent[1(✉)], Nazim Haouchine[1], Fryderyk Kogl[1], Samuel Joutard[2], Parikshit Juvekar[1], Erickson Torio[1], Alexandra J. Golby[1], Sebastien Ourselin[2], Sarah Frisken[1], Tom Vercauteren[2], Tina Kapur[1], and William M. Wells III[1,3]

[1] Harvard Medical School, Brigham and Women's Hospital, Boston, MA, USA
rdorent@bwh.harvard.edu
[2] King's College London, London, UK
[3] Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract. We introduce MHVAE, a deep hierarchical variational auto-encoder (VAE) that synthesizes missing images from various modalities. Extending multi-modal VAEs with a hierarchical latent structure, we introduce a probabilistic formulation for fusing multi-modal images in a common latent representation while having the flexibility to handle incomplete image sets as input. Moreover, adversarial learning is employed to generate sharper images. Extensive experiments are performed on the challenging problem of joint intra-operative ultrasound (iUS) and Magnetic Resonance (MR) synthesis. Our model outperformed multi-modal VAEs, conditional GANs, and the current state-of-the-art unified method (ResViT) for synthesizing missing images, demonstrating the advantage of using a hierarchical latent representation and a principled probabilistic fusion operation. Our code is publicly available (https://github.com/ReubenDo/MHVAE).

Keywords: Variational Auto-Encoder · Ultrasound · Brain Resection · Image Synthesis

1 Introduction

Medical imaging is essential during diagnosis, surgical planning, surgical guidance, and follow-up for treating brain pathology. Images from multiple modalities are typically acquired to distinguish clinical targets from surrounding tissues. For example, intra-operative ultrasound (iUS) imaging and Magnetic Resonance Imaging (MRI) capture complementary characteristics of brain tissues that can be used to guide brain tumor resection. However, as noted in [30], multi-modal
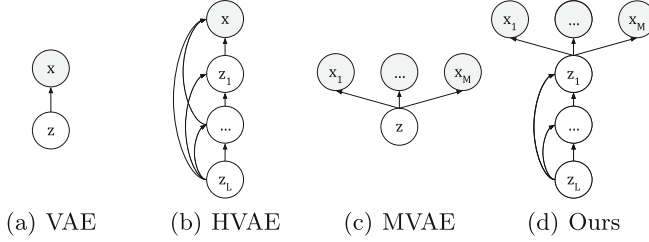
Fig. 1. Graphical models of: (a) variational auto-encoder (VAE); (b) hierarchical VAE (HVAE); (c) Multi-modal VAE (MVAE); (d) Multi-Modal Hiearchical VAE (Ours).

data is *expensive* and *sparse*, typically leading to incomplete sets of images. For example, the prohibitive cost of intra-operative MRI (iMRI) scanners often hampers the acquisition of iMRI during surgical procedures. Conversely, iUS is an affordable tool but has been perceived as difficult to read compared to iMRI [5]. Consequently, there is growing interest in synthesizing missing images from a subset of available images for enhanced visualization and clinical training.

Medical image synthesis aims to predict missing images given available images. Deep-learning based methods have reached the highest level of performance [29], including conditional generative adversarial (GAN) models [6,14,15,21] and conditional variational auto-encoders [3]. However, a key limitation of these techniques is that they must be trained for each subset of available images.

To tackle this challenge, unified approaches have been proposed. These approaches are designed to have the flexibility to handle incomplete image sets as input, improving practicality as only one network is used for generating missing images. To handle partial inputs, some studies proposed to fill missing images with arbitrary values [4,17,18,24]. Alternatively, other work aim at creating a common feature space that encodes shared information from different modalities. Feature representations are extracted independently for each modality. Then, arithmetic operations (e.g., mean [7,11,28], max [2] or a combination of sum, product and max [32]) are used to fuse these feature representations. However, these operations do not force the network to learn a shared latent representation of multi-modal data and lack theoretical foundations. In contrast, Multi-modal Variational Auto-Encoders (MVAEs) provide a principled probabilistic fusion operation to create a common representation space [8,30]. In MVAEs, the common representation space is low-dimensional (e.g., $\mathbb{R}^{256}$), which usually leads to blurry synthetic images. In contrast, hierarchical VAEs (HVAEs) [19,22,26,27] allow for learning complex latent representations by using a hierarchical latent structure, where the coarsest latent variable (z_L) represents global features, as in MVAEs, while the finer variables capture local characteristics. However, HVAEs have not yet been extended to multi-modal settings to synthesize missing images.

In this work, we introduce Multi-Modal Hierarchical Latent Representation VAE (MHVAE), the first multi-modal VAE approach with a hierarchical latent representation for unified medical image synthesis. Our contribution is four-fold.

First, we integrate a hierarchical latent representation into the multi-modal variational setting to improve the expressiveness of the model. Second, we propose a principled fusion operation derived from a probabilistic formulation to support missing modalities, thereby enabling image synthesis. Third, adversarial learning is employed to generate realistic image synthesis. Finally, experiments on the challenging problem of iUS and MR synthesis demonstrate the effectiveness of the proposed approach, enabling the synthesis of high-quality images while establishing a mathematically grounded formulation for unified image synthesis and outperforming non-unified GAN-based approaches and the state-of-the-art method for unified multi-modal medical image synthesis.

2   Background

Variational Auto-Encoders (VAEs). The goal of VAEs [16] is to train a generative model in the form of $p(x, z) = p(z)p(x|z)$ where $p(z)$ is a prior distribution (e.g. isotropic Normal distribution) over latent variables $z \in \mathbb{R}^H$ and where $p_\theta(x|z)$ is a decoder parameterized by θ that reconstructs data $x \in \mathbb{R}^N$ given z. The latent space dimension H is typically much lower than the image space dimension N, i.e. $H \ll N$. The training goal with respect to θ is to maximize the marginal likelihood of the data $p_\theta(x)$ (the "evidence"); however since the true posterior $p_\theta(z|x)$ is in general intractable, the variational evidence lower bound (ELBO) is instead optimized. The ELBO $\mathcal{L}_{\text{VAE}}(x; \theta, \phi)$ is defined by introducing an approximate posterior $q_\phi(z|x)$ with parameters ϕ:

$$\mathcal{L}_{\text{VAE}}(x; \theta, \phi) := \mathbb{E}_{q_\phi(z|x)}[\log(p_\theta(x|z))] - \text{KL}[q_\phi(z|x)||p(z)] , \qquad (1)$$

where $\text{KL}[q||p]$ is the Kullback-Leibler divergence between distributions q and p.

Multi-modal Variational Auto-Encoders (MVAE). Multi-modal VAEs [8, 25, 30] introduced a principled probabilistic formulation to support missing data at training and inference time. Multi-modal VAEs assume that M paired images $x = (x_1, ..., x_M) \in \mathbb{R}^{M \times N}$ are conditionally independent given a shared representation z as highlighted in Fig. 1, i.e. $p_\theta(x|z) = \prod_{i=1}^{M} p(x_i|z)$.

Instead of training one single variational network $q_\phi(z|x)$ that requires all images to be presented at all times, MVAEs factorize the approximate posterior as a combination of unimodal variational posteriors $(q_\phi(z|x_i))_{i=1}^{M}$. Given any subset of modalities $\pi \subseteq \{1, ..., M\}$, MVAEs have the flexibility to approximate the π-marginal posteriors $p(z|(x_i)_{i \in \pi})$ using the $|\pi|$ unimodal variational posteriors $(q_\phi(z|x_i))_{i \in \pi}$. MVAE [30] and U-HVED [8] factorize the π-marginal variational posterior as a product-of-experts (PoE), i.e.:

$$q_\phi^{\text{PoE}}(z|x_\pi) = p(z) \prod_{i \in \pi} q_\phi(z|x_i) . \qquad (2)$$

3   Methods

In this paper, we propose a deep multi-modal hierarchical VAE called MHVAE that synthesizes missing images from available images. MHVAE's design focuses on tackling three challenges: (i) improving expressiveness of VAEs and MVAEs using a hierarchical latent representation; (ii) parametrizing the variational posterior to handle missing modalities; (iii) synthesizing realistic images.

3.1   Hierarchical Latent representation

Let $x = (x_i)_{i=1}^M \in \mathbb{R}^{M \times N}$ be a complete set of paired (i.e. co-registered) images of different modalities where M is the total number of image modalities and N the number of pixels (e.g. $M = 2$ for T_2 MRI and iUS synthesis). The images x_i are assumed to be conditionally independent given a latent variable z. Then, the conditional distribution $p_\theta(x|z)$ parameterized by θ can be written as:

$$p_\theta(x|z) = \prod_{i=1}^M p_\theta(x_i|z) \ . \tag{3}$$

Given that VAEs and MVAEs typically produce blurry images, we propose to use a hierarchical representation of the latent variable z to increase the expressiveness the model as in HVAEs [19, 22, 26, 27]. Specifically, the latent variable z is partitioned into disjoint groups, as shown in Fig. 1 i.e. $z = \{z_1, ... z_L\}$, where L is the number of groups. The prior $p(z)$ is then represented by:

$$p_\theta(z) = p(z_L) \prod_{l=1}^{L-1} p_{\theta_l}(z_l|z_{>l}) \ , \tag{4}$$

where $p(z_L) = \mathcal{N}(z_L; 0, I)$ is an isotropic Normal prior distribution and the conditional prior distributions $p_{\theta_l}(z_l|z_{>l})$ are factorized Normal distributions with diagonal covariance parameterized using neural networks, i.e. $p_{\theta_l}(z_l|z_{>l}) = \mathcal{N}(z_l; \mu_{\theta_l}(z_{>l}), D_{\theta_l}(z_{>l}))$. Note that the dimension of the finest latent variable $z_1 \in \mathbb{R}^{H_1}$ is similar to number of pixels, i.e. $H_1 = \mathcal{O}(N)$ and the dimension of the latent representation exponentially decreases with the depth, i.e. $H_L \ll H_1$.

Reusing Eq. 1, the evidence $\log(p_\theta(x))$ is lower-bounded by the tractable variational ELBO $\mathcal{L}_{\mathrm{MHVAE}}^{\mathrm{ELBO}}(x; \theta, \phi)$:

$$\mathcal{L}_{\mathrm{MHVAE}}^{\mathrm{ELBO}}(x; \theta, \phi) = \sum_{i=1}^M \mathbb{E}_{q_\phi(z|x)}[\log(p_\theta(x_i|z))] - \mathrm{KL}\left[q_\phi(z_L|x)||p(z_L)\right]$$
$$- \sum_{l=1}^{L-1} \mathbb{E}_{q_\phi(z_{>l}|x)}\left[\mathrm{KL}[q_\phi(z_l|x, z_{>l})||p_\theta(z_l|z_{>l})]\right] \tag{5}$$

where $q_\phi(z|x) = \prod_{l=1}^L q_\phi(z_l|x, z_{>l})$ is a variational posterior that approximates the intractable true posterior $p_\theta(z|x)$.

3.2   Variational Posterior's Parametrization for Incomplete Inputs

To synthesize missing images, the variational posterior $(q_\phi(z_l|x, z_{>l}))_{l=1}^L$ should handle missing images. We propose to parameterize it as a combination of uni-modal variational posteriors. Similarly to MVAEs, for any set $\pi \subseteq \{1, ..., M\}$ of images, the conditional posterior distribution at the coarsest level L is expressed

$$q_{\phi_L}^{\text{PoE}}(z_L|x_\pi) = p(z_L) \prod_{i \in \pi} q_{\phi_L^i}(z|x_i) . \tag{6}$$

where $p(z_L) = \mathcal{N}(z_L; 0, I)$ is an isotropic Normal prior distribution and $q_{\phi_L}(z|x_i)$ is a Normal distribution with diagonal covariance parameterized using CNNs.

For the other levels $l \in \{1, .., L-1\}$, we similarly propose to express the conditional variational posterior distributions as a product-of-experts:

$$q_{\phi_l, \theta_l}^{\text{PoE}}(z_l|x_\pi, z_{>l}) = p_{\theta_l}(z_l|z_{>l}) \prod_{i \in \pi} q_{\phi_l^i}(z_l|x_i, z_{>l}) \tag{7}$$

where $q_{\phi_l^i}(z_l|x_i, z_{>l})$ is a Normal distribution with diagonal covariance parameterized using CNNs, i.e. $q_{\phi_l^i}(z_l|x_i, z_{>l}) = \mathcal{N}(z_l; \mu_{\phi_l^i}(x_i, z_{>l}); D_{\phi_l^i}(x_i, z_{>l}))$.

This formulation allows for a principled operation to fuse content information from available images while having the flexibility to handle missing ones. Indeed, at each level $l \in \{1, ..., L\}$, the conditional variational distributions $q_{\phi_l, \theta_l}^{\text{PoE}}(z_l|x_\pi, z_{>l})$ are Normal distributions with mean $\mu_{\phi_l, \theta_l}(x_\pi, z_{>l})$ and diagonal covariance $D_{\phi_l, \theta_l}(x_\pi, z_{>l})$ expressed in closed-form solution [12] as:

$$\begin{cases} D_{\phi_l, \theta_l}(x_\pi, z_{>l}) = \left(D_{\theta_l}^{-1}(z_{>l}) + \sum_{i \in \pi} D_{\phi_l^i}^{-1}(x_i, z_{>l}) \right)^{-1} \\ \mu_{\phi_l, \theta_l}(x_\pi, z_{>l}) = D_{\phi_l, \theta_l}^{-1}(x_\pi, z_{>l}) \left(D_{\theta_l}^{-1}(z_{>l})\mu_{\theta_l}(z_{>l}) + \sum_{i \in \pi} D_{\phi_l^i}^{-1}(x_i, z_{>l})\mu_{\phi_l^i}(x_i, z_{>l}) \right) \end{cases}$$

with $D_{\theta_L}(z_{>L}) = I$ and $\mu_{\theta_L}(z_{>L}) = 0$.

3.3   Optimization Strategy for Image Synthesis

The joint reconstruction and synthesis optimization goal is to maximize the expected evidence $\mathbb{E}_{x \sim p_{\text{data}}}[\log(p(x))]$. As the ELBO defined in Eq. 5 is valid for any approximate distribution q, the evidence, $\log(p_\theta(x))$, is in particular lower-bounded by the following subset-specific ELBO for any subset of images π:

$$\mathcal{L}_{\text{MAVAE}}^{\text{ELBO}}(x_\pi; \theta, \phi) = \underbrace{\sum_{i=1}^M \mathbb{E}_{q_\phi(z|x_\pi)}[\log(p_\theta(x_i|z_1))]}_{\text{reconstruction}} - \text{KL}\left[q_{\phi_L}(z_L|x_\pi)||p(z_L)\right]$$

$$- \sum_{l=1}^{L-1} \mathbb{E}_{q_{\phi_l, \theta_l}(z_{>l}|x_\pi)}\left[\text{KL}[q_{\phi_l, \theta_l}(z_l|x_\pi, z_{>l})||p_{\theta_l}(z_l|z_{>l})]\right] . \tag{8}$$

Hence, the expected evidence $\mathbb{E}_{x \sim p_{\text{data}}}[\log(p(x))]$ is lower-bounded by the average of the subset-specific ELBO, i.e.:

$$\mathcal{L}_{\text{MHVAE}} \coloneqq \frac{1}{|\mathcal{P}|} \sum_{\pi \in \mathcal{P}} \mathcal{L}_{\text{MAVAE}}^{\text{ELBO}}(x_\pi; \theta, \phi) \ . \tag{9}$$

Consequently, we propose to average all the subset-specific losses at each training iteration. The image decoding distributions are modelled as Normal with variance σ, i.e. $p_\theta(x_i|z_1) = \mathcal{N}(x_i; \mu_i(z_1), \sigma I)$, leading to reconstruction losses $-\log(p_\theta(x_i|z_1))$, which are proportional to $||x_i - \mu_i(z_1)||^2$. To generate sharper images, the L_2 loss is replaced by a combination of L_1 loss and GAN loss via a PatchGAN discriminator [14]. Moreover, the expected KL divergences are estimated with one sample as in [19]. Finally, the loss associated with the subset-specific ELBOs Eq. (9) is:

$$L = \sum_{i=1}^{N} \left(\lambda_{L_1} L_1(\mu_i, x_i) + \lambda_{\text{GAN}} L_{\text{GAN}}(\mu_i) \right) + \text{KL} \ .$$

Following standard practices [4,14], images are normalized in $[-1, 1]$ and the weights of the L_1 and GAN losses are set to $\lambda_{L_1} = 100$ and $\lambda_{\text{GAN}} = 1$.

4   Experiments

In this section, we report experiments conducted on the challenging problem of MR and iUS image synthesis.

Data. We evaluated our method on a dataset of 66 consecutive adult patients with brain gliomas who were surgically treated at the Brigham and Women's hospital, Boston USA, where both pre-operative 3D T2-SPACE and pre-dural opening intraoperative US (iUS) reconstructed from a tracked handheld 2D probe were acquired. The data will be released on TCIA in 2023. 3D T2-SPACE scans were affinely registered with the pre-dura iUS using NiftyReg [20] following the pipeline described in [10]. Three neurological experts manually checked registration outputs. The dataset was randomly split into a training set (N = 56) and a testing set (N = 10). Images were resampled to an isotropic 0.5 mm resolution, padded for an in-plane matrix of $(192, 192)$, and normalized in $[-1, 1]$.

Implementation Details. Since raw brain ultrasound images are typically 2D, we employed a 2D U-Net-based architecture. The spatial resolution and the feature dimension of the coarsest latent variable (z_L) were set to 1×1 and 256. The spatial and feature dimensions are respectively doubled and halved after each level to reach a feature representation of dimension 8 for each pixel, i.e. $z_1 \in \mathbb{R}^{196 \times 196 \times 8}$ and $z_L \in \mathbb{R}^{1 \times 1 \times 256}$. This leads to 7 latent variable levels, i.e. $L = 7$. Following state-of-the-art NVAE architecture [27], residual cells
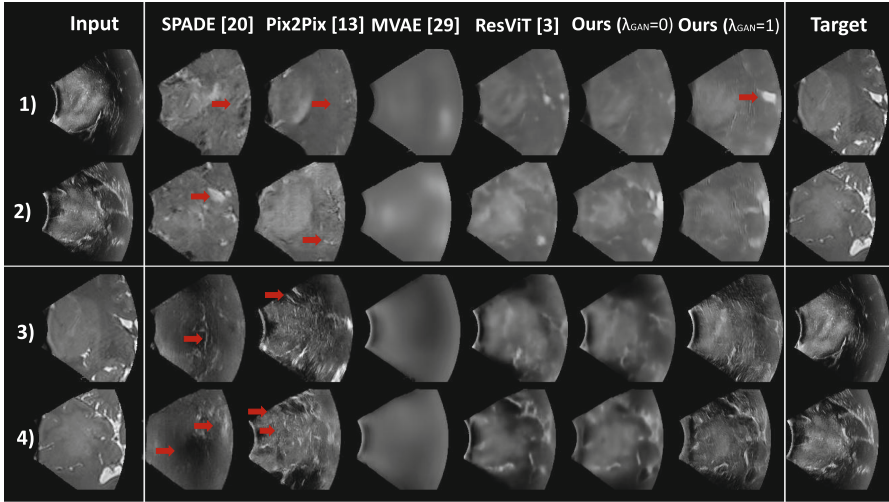
Fig. 2. Examples of image synthesis (rows 1 and 2: iUS → T$_2$; rows 3 and 4: T$_2$ → iUS) using SPADE [21], Pix2Pix [14], MVAE [30], ResViT [4] and MHVAE (ours) without and with GAN loss. As highlighted by the arrows, our approach better preserves anatomy compared to GAN-based approach and produces more realistic approach than the transformer-based approach (ResViT).

for the encoder and decoder from MobileNetV2 [23] are used with Squeeze and Excitation [13] and Swish activation. The image decoders $(\mu_i)_{i=1}^M$ correspond to 5 ResNet blocks. Following state-of-the-art bidirectional inference architectures [19,27], the representations extracted in the contracting path (from x_i to $(z_l)_l$) and the expansive path (from z_L to x_i and $(z_l)_{l<L}$) are partially shared. Models are trained for 1000 epochs with a batch size of 16. To improve convergence, λ_{GAN} is set to 0 for the first 800 epochs. Network architecture is presented in Appendix, and the code is available at https://github.com/ReubenDo/MHVAE.

Evaluation. Since paired data was available for evaluation, standard supervised evaluation metrics are employed: PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity), and LPIPS [31] (Learned Perceptual Image Patch Similarity). Quantitative results are presented in Table 1, and qualitative results are shown in Fig. 2. Wilcoxon signed rank tests ($p < 0.01$) were performed.

Ablation Study. To quantify the importance of each component of our approach, we conducted an ablation study. First, our model (MHVAE) was compared with MVAE, the non-hierarchical multi-modal VAE described in [30]. It can be observed in Table 1 that MHVAE (ours) significantly outperformed MVAE. This highlights the benefits of introducing a hierarchy in the latent representation. As shown in Fig. 2, MVAE generated blurry images, while our approach produced

sharp and detailed synthetic images. Second, the impact of the GAN loss was evaluated by comparing our model with ($\lambda_{GAN} = 0$) and without ($\lambda_{GAN} = 1$) the adversarial loss. Both models performed similarly in terms of evaluation metrics. However, as highlighted in Fig. 2, adding the GAN loss led to more realistic textures with characteristic iUS speckles on synthetic iUS. Finally, the image similarity between the target and reconstructed images (i.e., target image used as input) was excellent, as highlighted in Table 1. This shows that the learned latent representations preserved the content information from input modalities.

Table 1. Comparison against the state-of-the-art conditional GAN models for image synthesis. Available modalities are denoted by ●, the missing ones by ○. Mean and standard deviation values are presented. * denotes significant improvement provided by a Wilcoxon test ($p < 0.01$). Arrows indicate favorable direction of each metric.

	Input		iUS			T_2		
	iUS	T_2	PSNR(dB)↑	SSIM(%)↑	LPIPS(%)↓	PSNR(dB)↑	SSIM(%)↑	LPIPS(%)↓
MHVAE ($\lambda_{GAN} = 0$)	●	●	33.15 (2.48)	91.3 (3.5)	6.3 (2.3)	36.38 (2.40)	95.3 (1.9)	2.2 (0.8)
MHVAE ($\lambda_{GAN} = 1$)	●	●	31.54 (2.62)	89.1 (4.3)	7.1 (2.6)	34.35 (2.67)	93.6 (2.7)	2.8 (1.2)
Pix2Pix [14] $T_2 \rightarrow$ iUS	○	●	20.31 (3.78)	70.2 (12.0)	19.8 (5.7)	×	×	×
SPADE [21] $T_2 \rightarrow$ iUS	○	●	20.30 (3.62)	70.1 (12.1)	21.5 (6.9)	×	×	×
MVAE [30]	○	●	21.21 (4.20)	73.5 (10.9)	26.9 (10.5)	23.23 (4.55)	83.4 (8.1)	21.4 (9.0)
ResViT [4]	○	●	21.22 (3.10)	**75.2* (9.7)**	24.0 (7.5)	37.14 (5.94)	99.1 (0.9)	1.0 (0.5)
MHVAE ($\lambda_{GAN} = 0$)	○	●	**21.87* (4.06)**	74.9 (10.4)	24.2 (9.1)	36.41 (2.13)	95.5 (1.8)	7.2 (3.0)
MHVAE ($\lambda_{GAN} = 1$)	○	●	21.26 (3.93)	71.9 (11.4)	**19.0* (7.6)**	34.94 (2.27)	94.4 (2.3)	7.6 (3.2)
Pix2Pix [14] iUS $\rightarrow T_2$	●	○	×	×	×	21.01 (3.70)	77.9 (9.2)	17.4 (4.7)
SPADE [21] iUS $\rightarrow T_2$	●	○	×	×	×	20.12 (3.20)	74.3 (8.5)	18.6 (3.8)
MVAE [30]	●	○	23.02 (4.12)	75.3 (10.4)	25.5 (9.9)	21.70 (4.60)	82.6 (8.2)	21.7 (9.1)
ResViT [4]	●	○	35.09 (3.96)	97.6 (1.0)	3.5 (1.2)	21.70 (3.40)	**82.8* (7.6)**	18.9 (6.8)
MHVAE ($\lambda_{GAN} = 0$)	●	○	33.07 (2.34)	91.3 (3.4)	13.2 (4.8)	**22.16* (4.13)**	**82.8* (8.0)**	18.3 (7.6)
MHVAE ($\lambda_{GAN} = 1$)	●	○	31.58 (2.26)	90.8 (3.6)	12.0 (4.4)	22.12 (4.28)	81.7 (8.2)	**17.4* (7.3))**

State-of-the-Art Comparison. To evaluate the performance of our model (MHVAE) against existing image synthesis frameworks, we compared it to two state-of-the-art GAN-based conditional image synthesis methods: Pix2Pix [14] and SPADE [21]. These models have especially been used as synthesis backbones in previous MR/iUS synthesis studies [6,15]. Results in Table 1 show that our approach statistically outperformed these GAN methods with and without adversarial learning. As shown in Fig. 2, these conditional GANs produced realistic images but did not preserve the brain anatomy. Given that these models are not unified, Pix2Pix and SPADE must be trained for each synthesis direction ($T_2 \rightarrow$ iUS and iUS $\rightarrow T_2$). In contrast, MHVAE is a unified approach where one model is trained for both synthesis directions, improving inference practicality without a drop in performance. Finally, we compared our approach with ResViT [4], a transformer-based method that is the current state-of-the-art for unified multi-modal medical image synthesis. Our approach outperformed or reached similar performance depending on the metric. In particular, as shown in Fig. 2 and in Table 1 for the perceptual LPIPS metric, our GAN model synthesizes images that are visually more similar to the target images. Finally, our

approach demonstrates significantly lighter computational demands when compared to the current SOTA unified image synthesis framework (ResViT), both in terms of time complexity (8G MACs vs. 487G MACs) and model size (10M vs. 293M parameters). Compared to MVAEs, our hierarchical multi-modal approach only incurs a marginal increase in time complexity (19%) and model size (4%). Overall, this set of experiments demonstrates that variational auto-encoders with hierarchical latent representations, which offer a principled formulation for fusing multi-modal images in a shared latent representation, are effective for image synthesis.

5   Discussion and Conclusion

Other Potential Applications. The current framework enables the generation of iUS data using T_2 MRI data. Since image delineation is much more efficient on MRI than on US, annotations performed on MRI could be used to train a segmentation network on pseudo-iUS data, as performed by the top-performing teams in the crossMoDA challenge [9]. For example, synthetic ultrasound images could be generated from the BraTS dataset [1], the largest collection of annotated brain tumor MR scans. Qualitative results shown in Appendix demonstrate the ability of our approach to generalize well to T_2 imaging from BraTS. Finally, the synthetic images could be used for improved iUS and T_2 image registration.

Conclusion and Future Work. We introduced a multi-modal hierarchical variational auto-encoder to perform unified MR/iUS synthesis. By approximating the true posterior using a combination of unimodal approximates and optimizing the ELBO with multi-modal and uni-modal examples, MHVAE demonstrated state-of-the-art performance on the challenging problem of iUS and MR synthesis. Future work will investigate synthesizing additional imaging modalities such as CT and other MR sequences.

References

1. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge (2019)
2. Chartsias, A., Joyce, T., Giuffrida, M.V., Tsaftaris, S.A.: Multimodal mr synthesis via modality-invariant latent representation. IEEE Trans. Med. Imaging **37**(3), 803–814 (2017)

3. Chartsias, A., et al.: Disentangled representation learning in cardiac image analysis. Med. Image Anal. **58**, 101535 (2019)
4. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: residual vision transformers for multimodal medical image synthesis. IEEE Trans. Med. Imaging **41**(10), 2598–2614 (2022). https://doi.org/10.1109/TMI.2022.3167808
5. Dixon, L., Lim, A., Grech-Sollars, M., Nandi, D., Camp, S.: Intraoperative ultrasound in brain tumor surgery: a review and implementation guide. Neurosurg. Rev. **45**(4), 2503–2515 (2022)
6. Donnez, M., Carton, F.X., Le Lann, F., De Schlichting, E., Chabanas, M.: Realistic synthesis of brain tumor resection ultrasound images with a generative adversarial network. In: Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 11598, pp. 637–642. SPIE (2021)
7. Dorent, R., et al.: Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets. Med. Image Anal. **67**, 101862 (2021)
8. Dorent, R., Joutard, S., Modat, M., Ourselin, S., Vercauteren, T.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 74–82. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_9
9. Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., et al.: Crossmoda 2021 challenge: benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. Med. Image Anal. **83**, 102628 (2023)
10. Drobny, D., Vercauteren, T., Ourselin, S., Modat, M.: Registration of MRI and iUS data to compensate brain shift using a symmetric block-matching based approach. In: CuRIOUS (2018)
11. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: HeMIS: hetero-modal image segmentation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 469–477. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_54
12. Hernández-Lobato, J.M., et al.: Balancing flexibility and robustness in machine learning: semi-parametric methods and sparse linear models. Appendix C.2 (2010)
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-To-Image translation with conditional adversarial networks. In: CVPR (2017)
15. Jiao, J., Namburete, A.I.L., Papageorghiou, A.T., Noble, J.A.: Self-supervised ultrasound to mri fetal brain image synthesis. IEEE Trans. Med. Imaging **39**(12), 4413–4424 (2020). https://doi.org/10.1109/TMI.2020.3018560
16. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: ICLR (2014)
17. Lee, D., Kim, J., Moon, W.J., Ye, J.C.: Collagan: collaborative gan for missing image data imputation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2487–2496 (2019)
18. Li, H., et al.: DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 795–803. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_87
19. Maaløe, L., Fraccaro, M., Liévin, V., Winther, O.: BIVA: a very deep hierarchy of latent variables for generative modeling. In: NeurIPS (2019)
20. Modat, M., et al.: Fast free-form deformation using graphics processing units. Comput. Methods Prog. Biomed. **98**, 278–224 (2010)
21. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019)

22. Ranganath, R., Tran, D., Blei, D.: Hierarchical variational models. In: ICML (2016)
23. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: CVPR (2018)
24. Sharma, A., Hamarneh, G.: Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. IEEE Trans. Med. Imaging **39**(4), 1170–1183 (2020)
25. Shi, Y., Paige, B., Torr, P., et al.: Variational mixture-of-experts autoencoders for multi-modal deep generative models. In: NeurIPS, vol. 32 (2019)
26. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O.: Ladder variational autoencoders. In: NeurIPS (2016)
27. Vahdat, A., Kautz, J.: NVAE: a deep hierarchical variational autoencoder. In: NeurIPS, vol. 33 (2020)
28. Varsavsky, T., Eaton-Rosen, Z., Sudre, C.H., Nachev, P., Cardoso, M.J.: PIMMS: permutation invariant multi-modal segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 201–209. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_23
29. Wang, T., et al.: A review on medical imaging synthesis using deep learning and its clinical applications. J. Appl. Clin. Med. Phys. **22**(1), 11–36 (2021)
30. Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. In: NeurIPS, vol. 31 (2018)
31. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
32. Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L.: Hi-net: hybrid-fusion network for multi-modal mr image synthesis. IEEE Trans. Med. Imaging **39**(9), 2772–2781 (2020)