TECHNICAL UNIVERSITY OF DENMARK

Generative Medical Image Segmentation

Date of submission: May 29, 2025

Submitted By:

Kasper Rønberg, s214786

Mads-Emil Leth Honoré, s214808

Lasse Bisp, s214807

Holger Lyng, s214776

Table of Contents

List of Figures

List of Tables

1 Introduction

The application of machine learning (ML) in the medical field has become increasingly prominent, particularly in the domain of medical imaging. ML techniques are now frequently employed to detect diseases or injuries that may be difficult for clinicians to identify visually, or to assist in accelerating the diagnostic review process. This report investigates the use of generative probabilistic ML models such as variational autoencoders (VAEs) for image segmentation of magnetic resonance imaging (MRI) scans.

2 Research Question

Given the motivation to explore how generative probabilistic models can enhance medical image analysis and support clinical decision-making, particularly in reducing the time and effort required to assess radiological scans, the following research question is proposed:

> *How effectively can different variational autoencoders perform brain tumor segmentation on MRI scans, and how do they compare to other approaches in terms of accuracy and practical applicability?*

3 Data

The dataset used in this project is the "LGG Segmentation Dataset". It contains brain MRIs with manual fluid-attenuated inversion recovery (FLAIR) abnormality segmentation masks, obtained from The Cancer Imaging Archive. These correspond to 110 patients included in The Cancer Genome Atlas (TCGA) lower-grade glioma collection, each with at least a FLAIR sequence and available genomic cluster data. The data is downsampled and split into training and test images. The images are also normalized from [0, 255] range to [0, 1]. Pre-processing is important to improve results and minimize computational cost while maximizing efficiency.

4 Theory

4.1 Variational Autoencoder

A Variational Autoencoder (VAE) is a generative model that learns a low-dimensional latent representation of data by encoding inputs into a distribution over latent variables rather than fixed points. This enables both reconstruction and generation of new, coherent samples.

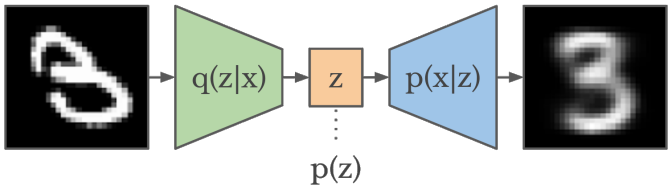


Figure 1: *Variational Autoencoder Illustration.*

The encoder maps an input $\mathbf{x}$ to the parameters of a Gaussian distribution $q_\phi(\mathbf{z}|\mathbf{x})$. A latent vector $\mathbf{z}$ is then sampled using the reparameterization trick to allow for gradient-based optimization. The decoder reconstructs $\mathbf{x}$ from $\mathbf{z}$ via the likelihood $p_\theta(\mathbf{x}|\mathbf{z})$. Training maximizes the evidence lower bound (ELBO), which balances reconstruction accuracy and regularization via Kullback-Leibler (KL) divergence [1]:

$$\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))$$

This encourages meaningful latent spaces aligned with a prior $p(\mathbf{z})$, typically a standard normal, making the VAE suitable for generative tasks. The advantage of using probabilistic generative models such as the VAE and its variations lies in their flexibility. Unlike traditional segmentation models such as U-Net, which produce a single deterministic solution, VAEs can generate multiple plausible segmentations, offering a more expressive representation of uncertainty in the segmentation task.

4.2 Hierarchical VAE

The hierarchical variational auto encoder (HVAE) extends the standard VAE framework by incorporating multiple stochastic latent layers. Instead of a single latent variable, a HVAE employs a sequence of latent variables, often denoted with layer indicators, such as $\mathbf{z}_l$, where $\mathbf{z}_{>l}$ refers to layers higher in the hierarchy. In a HVAE, training typically involves maximizing the ELBO. The ELBO objective includes

a KL penalty term for each latent layer. The KL is a measure of how much more information the posterior distribution carries compared to the prior, a quantity that is aimed to be minimized [1].

The generative model in a HVAE describes how data is produced from these layers, potentially modeling dependencies between adjacent layers, such as $p_\theta(\mathbf{z}_l^i|\mathbf{z}_{>l})$. Correspondingly, the inference model $q_\phi(\mathbf{z}_l^i|\mathbf{x}, \mathbf{z}_{>l})$ approximates the posterior distribution over the latent variables given the observed data x and potentially higher latent layers. For training, it involves maximizing the ELBO, and this includes a KL penalty term for each latent layer, as shown in the form [1]:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{z})\right] - D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}_L|\mathbf{x}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{z}_L)\right) \\ - \sum_{l=L-1}^{1} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_{>l}|\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}_l|\mathbf{x}, \mathbf{z}_{>l}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{z}_l|\mathbf{z}_{>l})\right)\right]\end{aligned}$$

4.3 Conditional-VAE

Conditional Variational Autoencoders (CVAEs) are a generative modeling framework designed to learn the conditional distribution $p(\mathbf{y}|\mathbf{x})$. Unlike standard VAEs, which model the unconditional data distribution, CVAEs incorporate conditioning information directly into both the encoder and decoder networks. In a CVAE, the generative process includes a prior network $p(\mathbf{z}|\mathbf{x})$, a decoder $p(\mathbf{y}|\mathbf{x}, \mathbf{z})$, and a recognition network $q(\mathbf{z}|\mathbf{x}, \mathbf{y})$ that approximates the posterior. The model is trained by maximizing a conditional ELBO:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})}[\log p(\mathbf{y}|\mathbf{x}, \mathbf{z})] - \beta\, \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}, \mathbf{y})\|p(\mathbf{z}|\mathbf{x})),$$

where β is a hyperparameter that controls the regularization strength, as further discussed in Section 4.4.2. Applications of CVAEs in medical imaging include the Probabilistic U-Net (PU-Net), which combines a U-Net with a CVAE to generate multiple plausible segmentation maps [2].

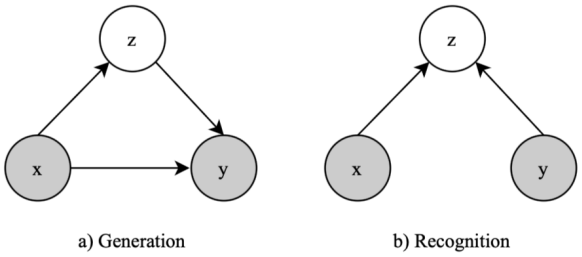


Figure 2: *CVAE PGM. (a) Generation network used for prediction. (b) Recognition network used during training.*

4.4 Avoiding Posterior Collapse

A significant challenge in VAE is posterior collapse. This phenomenon can appear as variational pruning, where the model disregards latent layers. Careful consideration is thus necessary to effectively train VAEs and ensure meaningful information flows through the entire model. The following subsections present approaches for reducing posterior collapse [1].

4.4.1 Skip Connections

A well-known challenge in training deep neural networks is the degradation problem, where increasing network depth leads to poorer performance. To address this, skip connections have been introduced to mitigate the issue by reintroducing $\mathbf{x}$. Similarly, a major issue in training VAEs is latent variable collapse, where the decoder is prone to ignoring $\mathbf{z}$. This problem is particularly pronounced when the latent variables are only weakly coupled to the data [1].

To counteract this, skip connections can be introduced between the latent space and the decoder. These connections allow direct information flow from the latent variables to multiple layers of the decoder, enhancing both expressivity and gradient propagation. This architecture, referred to as a skip-VAE, encourages stronger interaction between $\mathbf{z}$ and $\mathbf{y}$, thereby reducing the likelihood of latent variables being ignored and improving the ability of the model to capture complex output distributions [1].

4.4.2 KL Annealing

KL annealing is a common technique used to address latent variable collapse in VAEs. It gradually increases the weight β on the KL divergence term in the ELBO, starting from zero (equivalent to an autoencoder) and moving towards one (standard VAE training). This progressive schedule encourages the model to first focus on reconstruction before enforcing prior regularization [1].

5 Methods

To predict segmentation masks from the MRI-derived input data, several variational autoencoders are employed, namely traditional VAE, Skip-VAE, HVAE, Probabilistic U-Net (PU-Net), and Hierarchical

Probabilistic U-Net (HPU-Net). These are then compared to a U-Net, a standard image segmentation model. KL annealing is employed in every VAE tested in this report, where β is increased gradually during a set warm-up. Furthermore, to ensure a fair comparison between models, the encoder and decoder architectures are kept as similar as possible.

5.1 VAE

A traditional VAE, as described in Section 4.1, is used as a baseline model for comparison with more advanced architectural variants. For reconstruction of the segmentation masks, the ELBO becomes:

$$\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{y}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))$$

5.2 Skip-VAE

The Skip-VAE implementation includes skip connections from the encoder to the decoder, as well as from the latent variable to every layer of the decoder. This helps prevent posterior collapse and degradation [1]. Since the core VAE architecture remains unchanged, the ELBO formulation is identical to that of the standard VAE.

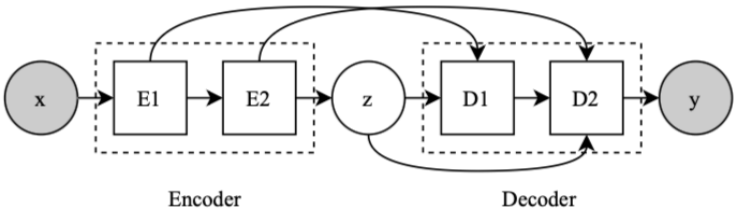


Figure 3: *Skip-VAE PGM Representation.*

5.3 Hierarchical VAE

The HVAE is implemented with three latent layers, as shown in Figure 4. Latents are sampled in a top-down manner, enabling the model to capture multi-scale structure. Residual networks (diamonds) help maintain information flow across layers [1]. The training objective corresponds to the ELBO defined in Section 4.2.

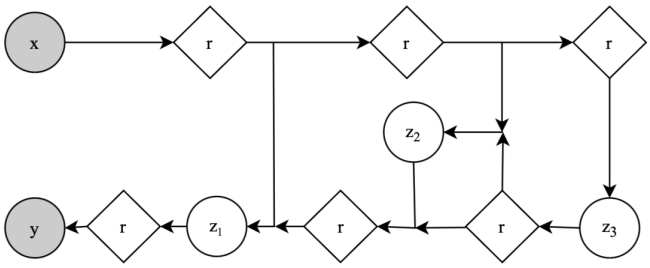


Figure 4: *Hierarchical VAE PGM. Diamonds represent a residual network.*

5.4 Probabilistic U-Net

The PU-Net extends the classic encoder–decoder U-Net with a conditional VAE. The encoder extracts multiscale features, while a parallel recognition network estimates $q(\mathbf{z}|\mathbf{x},\mathbf{y})$ during training. At inference, $\mathbf{z}$ is drawn from $p(\mathbf{z}|\mathbf{x})$, concatenated with the last layer of the U-Net, and passed through a decoder [3]. As this model is a variation of a conditional VAE, the ELBO follows the formulation given in Section 4.3.

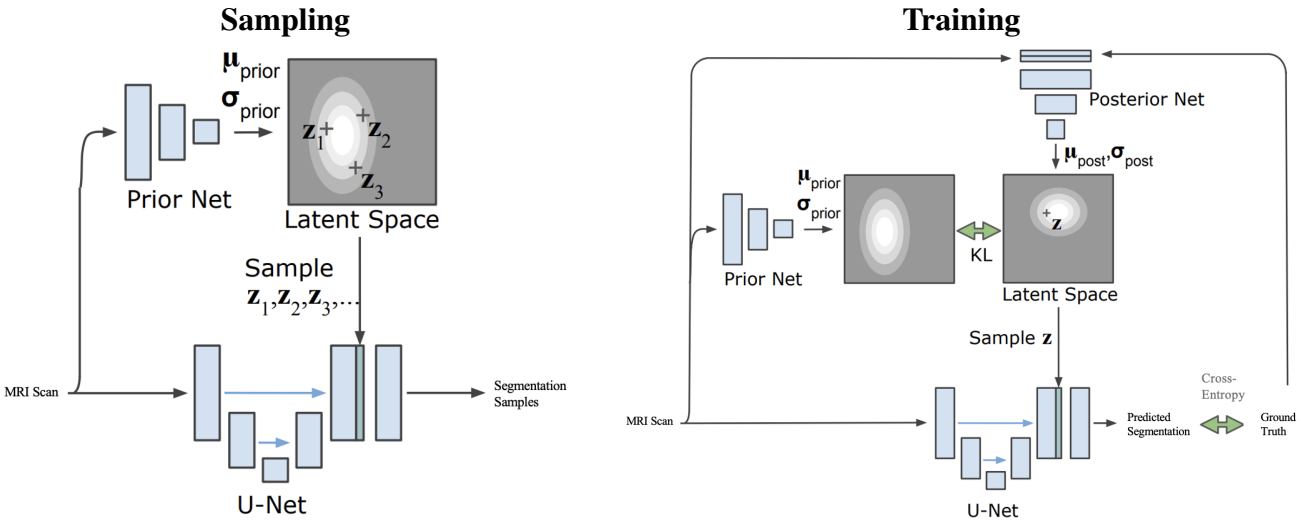


Figure 5: *Overview of the Probabilistic U-Net. (a) Sampling process during inference. (b) Training process for a single example.*

5.5 Hierarchical Probabilistic U-Net

The implementation of the HPU-Net replaces the single latent of the PU-Net with a hierarchical latent structure that mirrors the up-sampling from the decoder (Figure 6) [4]. A total of three latent layers were

used. Since this model is a mixture of conditional and hierarchical VAE the ELBO becomes:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[\log p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}, \mathbf{z})\right] - \beta(D_{\mathrm{KL}} \left(q_{\boldsymbol{\phi}}(\mathbf{z}_L|\mathbf{x}, \mathbf{y}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{z}_L)\right) + \sum_{l=L-1}^{1} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_{>l}|\mathbf{x})} \left[D_{\mathrm{KL}} \left(q_{\boldsymbol{\phi}}(\mathbf{z}_l|\mathbf{x}, \mathbf{y}, \mathbf{z}_{>l}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{z}_l|\mathbf{z}_{>l}, \mathbf{x})\right)\right])$$

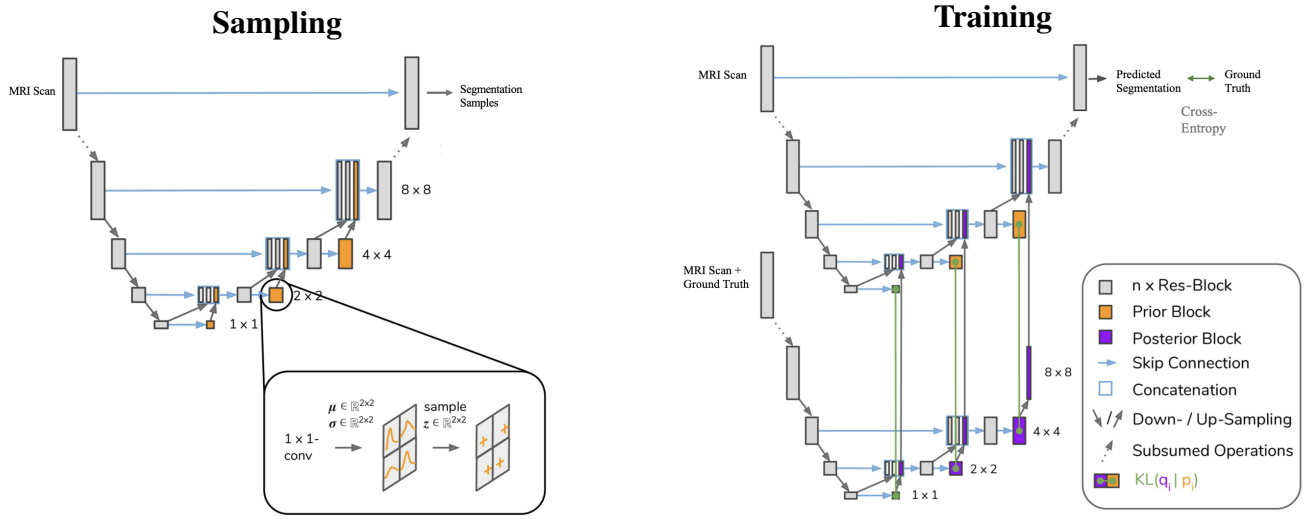
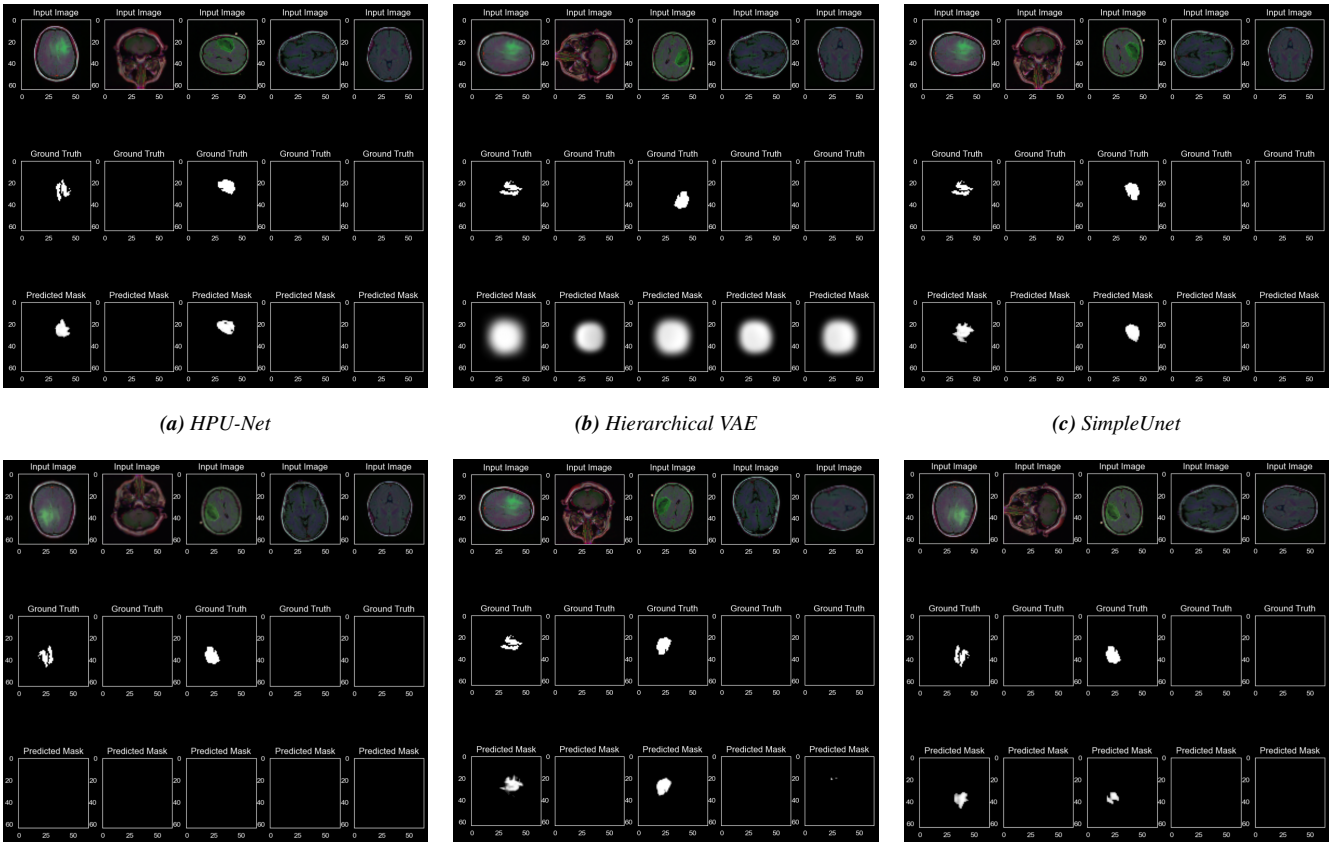


Figure 6: *Overview of the Hierarchical Probabilistic U-Net. (a) Sampling process during inference. (b) Training process for a single example.*

6 Results

By running the code provided in the notebook, results were obtained after training each model for 300 epochs using 64×64 images, in order to keep computational complexity low. All models were trained under identical conditions to ensure a fair comparison. Figure 7 presents visual outputs generated by the models, offering qualitative insights into their ability to reconstruct or segment relevant structures from the input data.



(a) *HPU-Net* ***(b)*** *Hierarchical VAE* ***(c)*** *SimpleUnet*

(d) *Traditional VAE* ***(e)*** *Skip-VAE* ***(f)*** *PU-Net*

Figure 7: *Comparison of segmentation outputs from HPU-Net, HVAE, U-Net, PU-Net, traditional VAE, and Skip-VAE models.*

Based on the visual comparison of predicted segmentation masks against the ground truth, the HPU-Net (Figure 7a), Skip-VAE (Figure 7e), and SimpleUNet (Figure 7c) demonstrate the strongest performance, producing segmentations that closely match the annotated tumor regions. Conversely, the HVAE (Figure 7b) struggles with precision, producing overly smooth and diffuse masks, likely as a consequence of high ELBO loss, possibly due to posterior collapse.

The quantitative evaluation of the model performances is summarized in Table 1. The reported metrics include the Binary Cross-Entropy (BCE) loss, which measures the mask reconstruction error; the average Dice loss, which quantifies the similarity between the predicted mask and the ground truth; and the ELBO loss. These metrics reflect reconstruction accuracy, segmentation quality, and generative performance, respectively.

Model	BCE Loss	Average DICE Score	ELBO Loss
Traditional VAE	1.03	0.65	608.08
Skip-VAE	0.01	0.9	24.11
PU-Net	0.02	0.78	26.93
HPU-Net	0.04	0.84	19.06
Hierarchical VAE	0.04	0.65	139.8
SimpleUnet	0.01	0.9	-

Table 1: *Performance metrics for various VAE architectures, 300 epochs.*

7 Discussion and Conclusion

The traditional VAE performs significantly worse than the other models, potentially due to posterior collapse or degradation caused by an excessively deep neural network. This depth may result in the loss of critical information in the latent layer of the network. In contrast, the Skip-VAE incorporates skip connections both from the encoder to the decoder and from the latent space to the different layers of the decoder. These architectural enhancements lead to notable improvements in performance as seen for the Skip-VAE. In the HVAE, skip connections are only introduced between the hierarchical latent variables. While this also improves performance in terms of BCE and ELBO losses, the HVAE still underperforms relative to other models. Future work could explore the implementation of skip connections between the encoder and decoder, as well as between the latent spaces and various decoder layers, to mitigate potential issues with posterior collapse and performance degradation.

Both the PU-Net and HPU-Net exhibit significant improvements over the HVAE and the traditional VAE. However, only the Skip-VAE achieves performance comparable to the standard U-Net. The performance gap may again be attributed to posterior collapse, wherein valuable information encoded in the latent representations is lost during training.

To better understand this phenomenon, future work could investigate the evolution of reconstruction and KL divergence losses throughout the training process. Moreover, as discussed in [4], the standard

ELBO loss for hierarchical models may lead to suboptimal convergence of the priors. As an alternative, the authors propose the GECO objective, which introduces a constraint on the reconstruction term and dynamically balances it with the KL divergence. This method has been shown to produce significantly better results.

Conclusively, it has been shown that while traditional VAEs are limited in their ability to perform accurate brain tumor segmentation on MRI scans, architectural improvements in models such as the Skip-VAE and HPU-Net lead to significant gains in both accuracy and practical applicability. Among the models evaluated, the Skip-VAE stands out by achieving performance comparable to the standard U-Net, highlighting its effectiveness relative to other approaches.

8 Bibliography

[1] Kevin P. Murphy. *Probabilistic Machine Learning Advanced Topics*. The MIT Press Cambridge, Massachusetts London, England, 2023.

[2] Honglak Lee Kihyuk Sohn Xinchen Yan. "Learning Structured Output Representation using Deep Conditional Generative Models". In: *Advances in Neural Information Processing Systems 28* (2015). URL: https://papers.nips.cc/paper_files/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html.

[3] Bernardino Romera-Paredes Simon A. A. Kohl. "A Probabilistic U-Net for Segmentation of Ambiguous Images". In: (2018). DOI: https://doi.org/10.48550/arXiv.1806.05034.

[4] Simon AA Kohl et al. "A Hierarchical Probabilistic U-Net for Modeling Multi-Scale Ambiguities". In: *arXiv preprint arXiv:1905.13077* (2019).