A Probabilistic U-Net for Segmentation of Ambiguous Images

Simon A. A. Kohl et al., DeepMind, London, UK

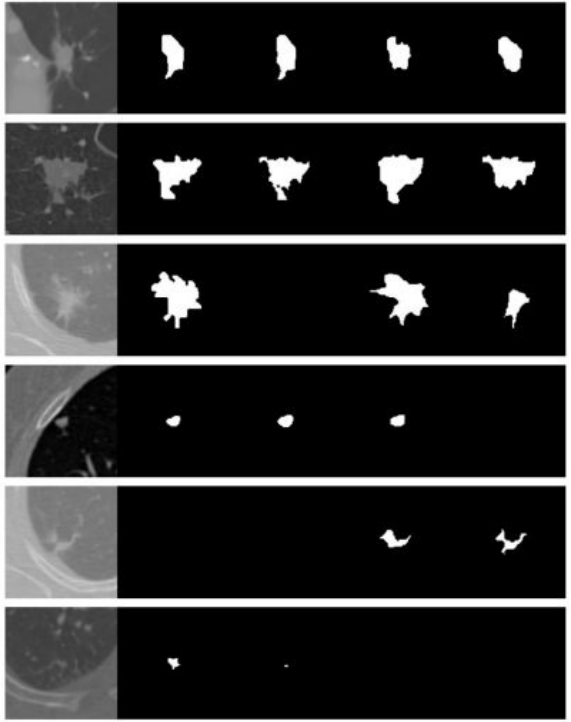
In NeurIPS 2018

2020-09-28
MI2RL
Kyuri Kim

Introduction

- Image ambiguity

- There exists an important class of images where even the full image context is not sufficient to resolve all ambiguities.
- A lesion might be clearly visible, but the information about whether it is cancer tissue or not might not be available from this image alone.
- Especially in medical applications where a subsequent diagnosis or a treatment depends on the segmentation map, an algorithm that only provides the most likely hypothesis might lead to misdiagnoses and sub-optimal treatment.
- If multiple consistent hypotheses are provided, these can be directly propagated into the next step in a diagnosis pipeline, they can be used to suggest further diagnostic tests to resolve the ambiguities



Potential Cancer Expert Graders

Introduction

- Contributions

(1) Our framework provides consistent segmentation maps instead of pixel-wise probabilities and can therefore give a joint likelihood of modes.

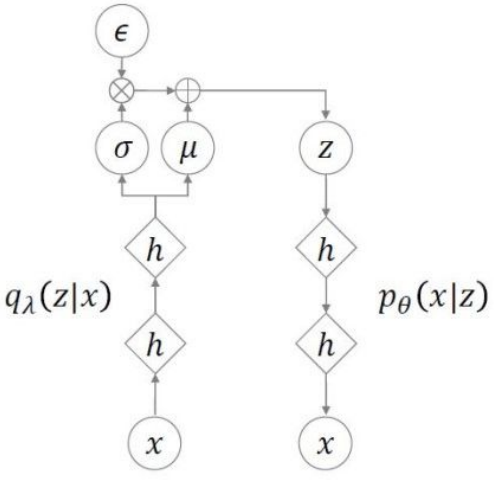
(2) Our model can induce arbitrarily complex output distributions including the occurrence of very rare modes, and is able to learn calibrated probabilities of segmentation modes.

(3) Sampling from our model is computationally cheap.

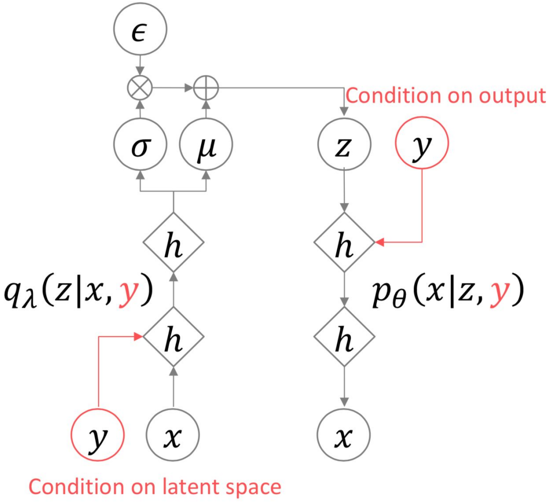
(4) In contrast to many existing applications of deep generative models that can only be qualitatively evaluated, our application and datasets allow quantitative performance evaluation including penalization of missing modes.

Related Work

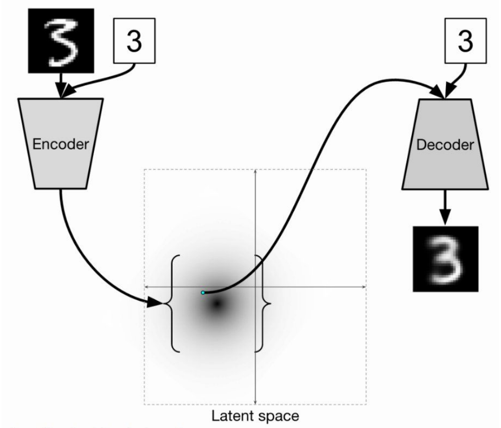
- Conditional VAE



Vanilla VAE



CVAE

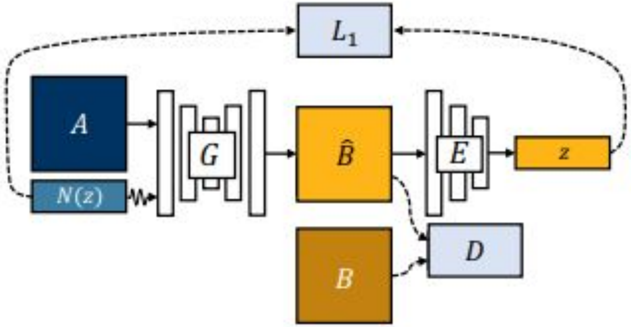


https://ratsgo.github.io/generative%20model/2018/01/28/VAEs/, https://ijdykeman.github.io/ml/2016/12/21/cvae.html

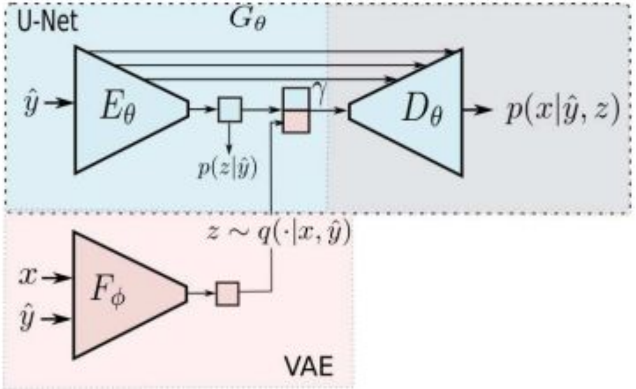
Related Work

- Image-to-image translation

- Bicycle GAN
    - Attempt to solve the mode-collapse problem.
    - their model encompasses a fixed prior distribution and during training their posterior distribution is only conditioned on the output image.



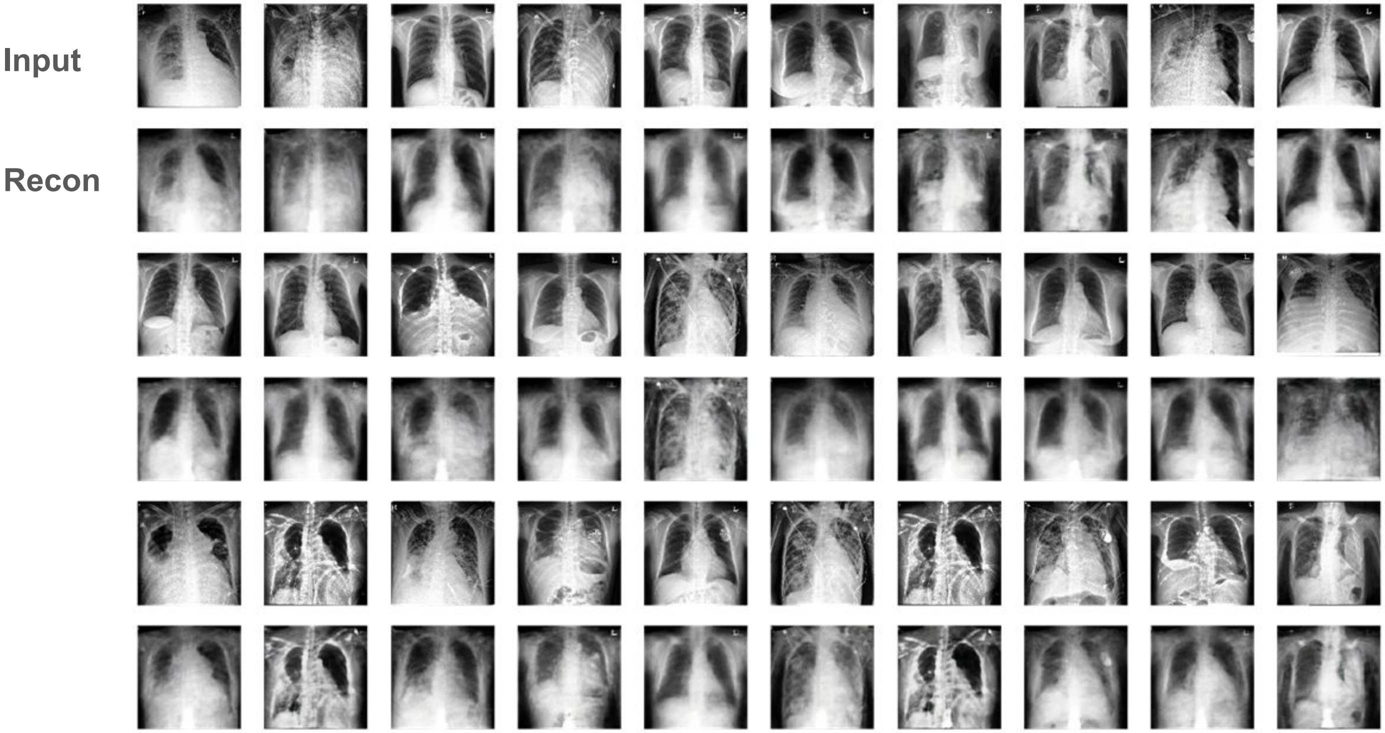
- Unet with VAE
    - additional pretrained VGG-net that is employed as a reconstruction loss.



Toward Multimodal Image-to-Image Translation https://arxiv.org/pdf/1711.11586.pdf
A Variational U-Net for Conditional Appearance and Shape Generation, https://arxiv.org/abs/1804.04694

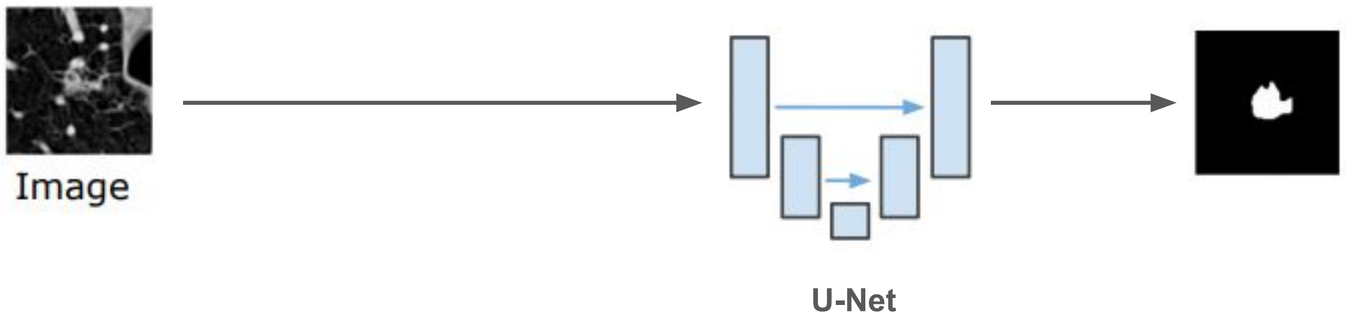
Related Work

- Examples of CXR using ConvVAE



Method

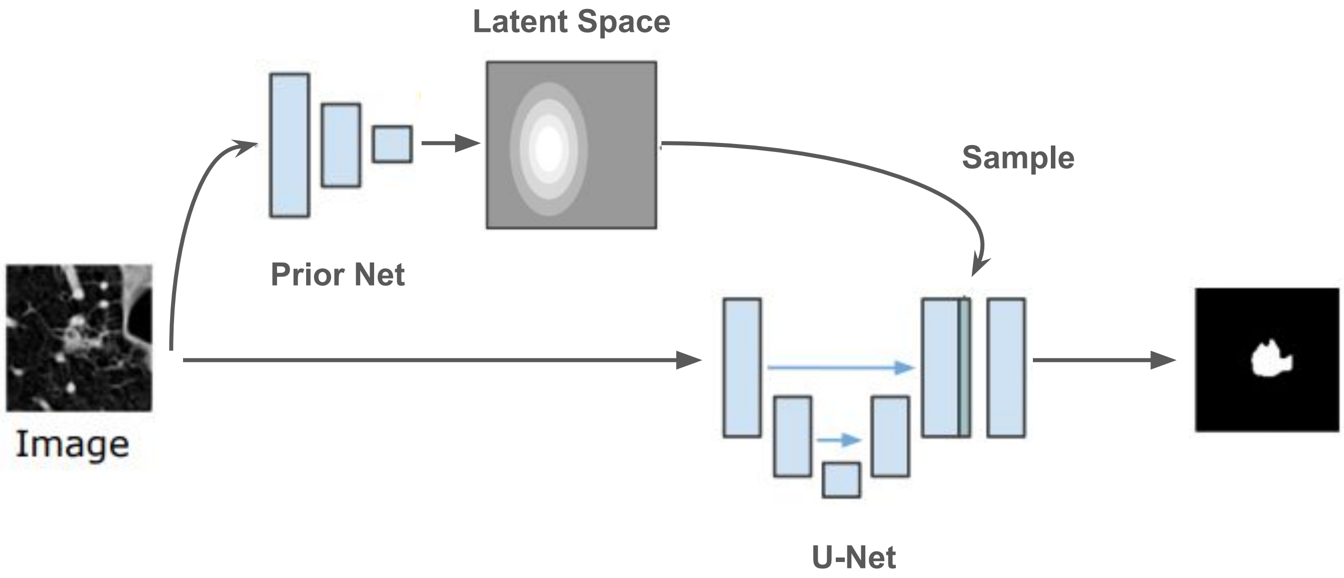
- Deterministic U-Net



https://nips.cc/media/Slides/nips/2018/220e(05-09-45)-05-10-35-12641-A_Probabilistic.pdf

Method

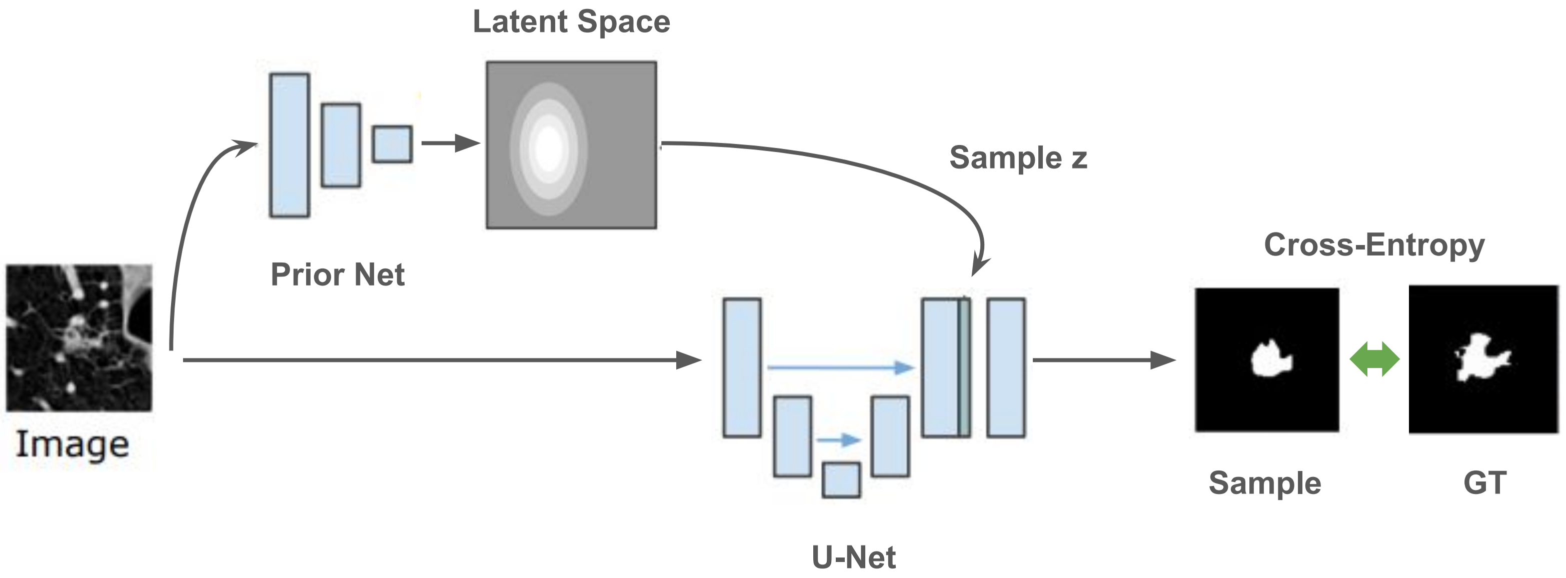
- Probabilistic U-Net: Sampling



https://nips.cc/media/Slides/nips/2018/220e(05-09-45)-05-10-35-12641-A_Probabilistic.pdf

Method

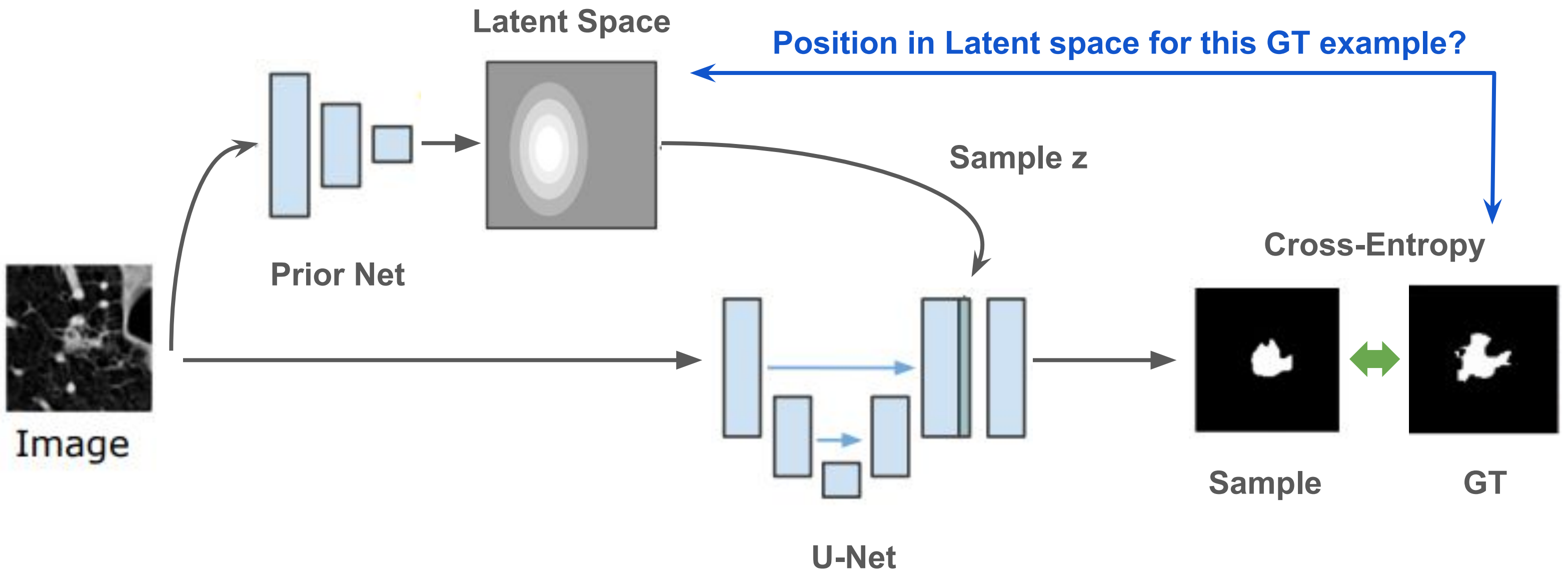
- Training



https://nips.cc/media/Slides/nips/2018/220e(05-09-45)-05-10-35-12641-A_Probabilistic.pdf

Method

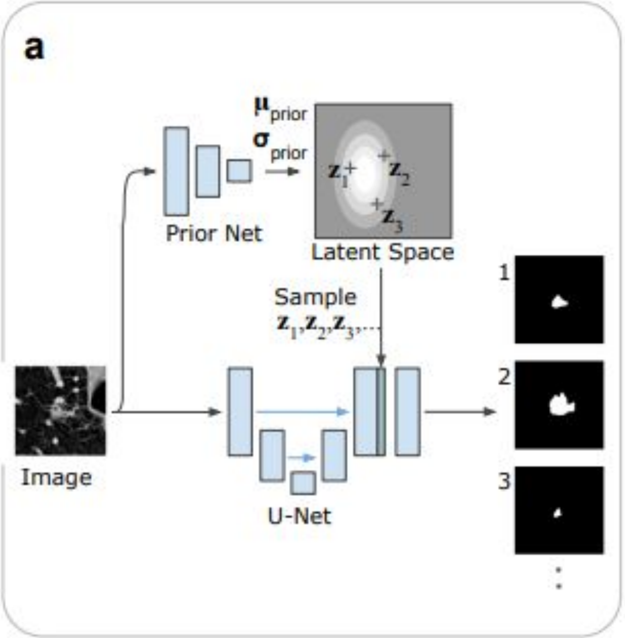
- Training



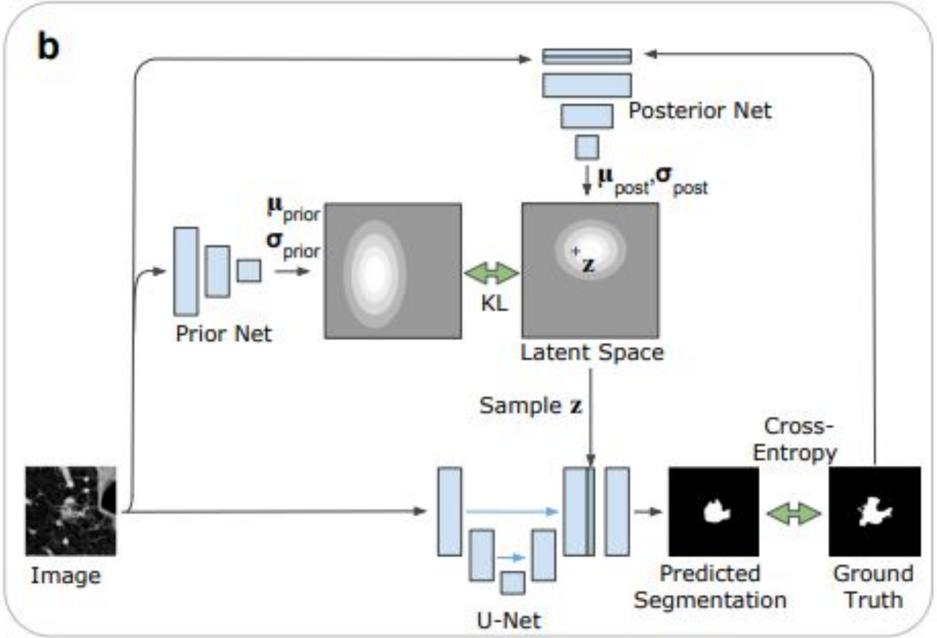
https://nips.cc/media/Slides/nips/2018/220e(05-09-45)-05-10-35-12641-A_Probabilistic.pdf

Method

- Network Architecture



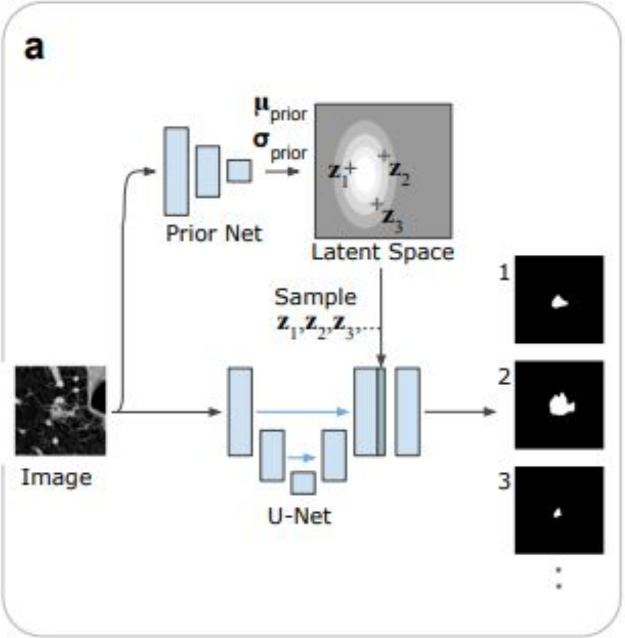
a. Sampling process



b. Training process

Method

- Network Architecture



a. Sampling process

$$\mathbf{z}_i \sim P(\cdot|X) = \mathcal{N}\left(\boldsymbol{\mu}_{\text{prior}}(X;\omega), \operatorname{diag}(\boldsymbol{\sigma}_{\text{prior}}(X;\omega))\right) \quad (1)$$

zi : random sample
P(·|X): priorm axis-aligned gaussian.
μprior(X; ω) ∈ RN: mean, σprior(X; ω) ∈ RN: variance
ω: parametrized by weights
X: input image

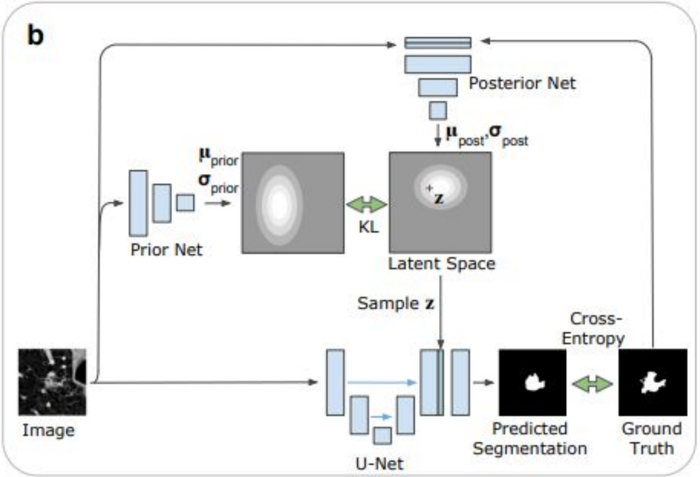
$$S_i = f_{\text{comb.}}\left(f_{\text{U-Net}}(X;\theta), \mathbf{z}_i; \psi\right) \quad (2)$$

Si : segmentation map
fcom: three subsequent 1 × 1 convolutions
Ψ: weights

Only the function fcomb. needs to be re-evaluated m times.

Method

- Network Architecture



b. Training process

$$\mathbf{z} \sim Q(\cdot|X,Y) = \mathcal{N}\left(\boldsymbol{\mu}_{\text{post}}(X,Y;\nu), \text{diag}(\boldsymbol{\sigma}_{\text{post}}(X,Y;\nu))\right) \quad (3)$$

z : random sample from posterior Q(·|X)
μpost(X, Y ; ν): posterior mean, σpost(X, Y ; ν): posterior variance
ν: posterior net weights
Y: segmentation mask

$$\mathcal{L}(Y,X) = \overbrace{\boxed{\mathbb{E}_{z\sim Q(\cdot|Y,X)}\left[-\log P_c(Y|S(X,z))\right]} + \beta \cdot \boxed{D_{\text{KL}}\left(Q(z|Y,X)||P(z|X)\right)}}^{\text{-ELBO}}$$

Cross-entropy loss (S, Y)

Kullback-Leibler divergence (posterior Q, prior P)

Method

- Performance Measure

Not only want to compare a deterministic prediction with a unique ground truth, but rather we are interested in comparing distributions of segmentations.

Use the **generalized energy distance**, which leverages distances between observations:

Disagreement between
gt and pred sample

Disagreement between
a pair of gt masks

$$D^2_{\mathrm{GED}}(P_{\mathrm{gt}}, P_{\mathrm{out}}) = 2\boxed{\mathbb{E}\left[d(S,Y)\right]} - \boxed{\mathbb{E}\left[d(S,S')\right]} - \boxed{\mathbb{E}\left[d(Y,Y')\right]},$$

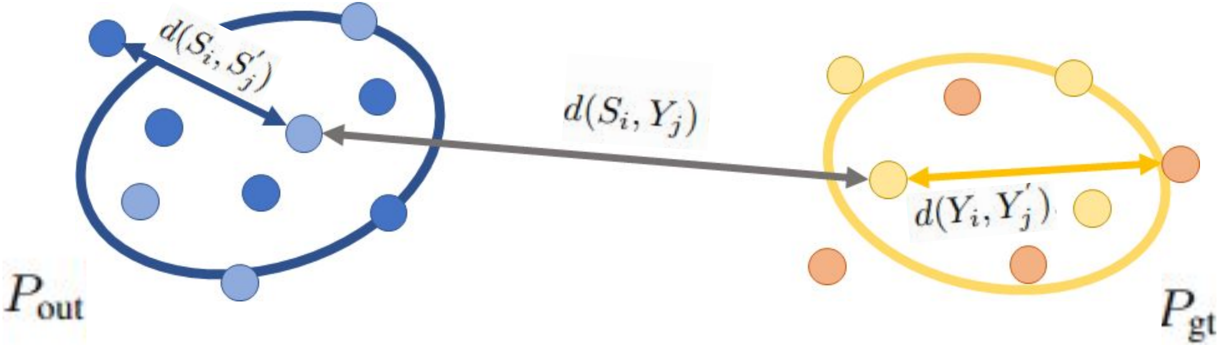
Disagreement between
a pair of pred masks

S, S 0: Independent samples from the predicted distribution Pout
Y, Y 0: Independent samples from the ground truth distribution Pgt
d(x, y) = 1 − IoU(x, y)

A class of statistics based on distances.

Method

- Performance Measure: generalized energy distance (MMD)



$$\hat{D}^2_{\text{GED}}(P_{\text{gt}}, P_{\text{out}}) = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} d(S_i, Y_j) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d(S_i, S_j') - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} d(Y_i, Y_j').$$

(m: ground truth samples, n: samples from the model)

Method

- Baseline Methods

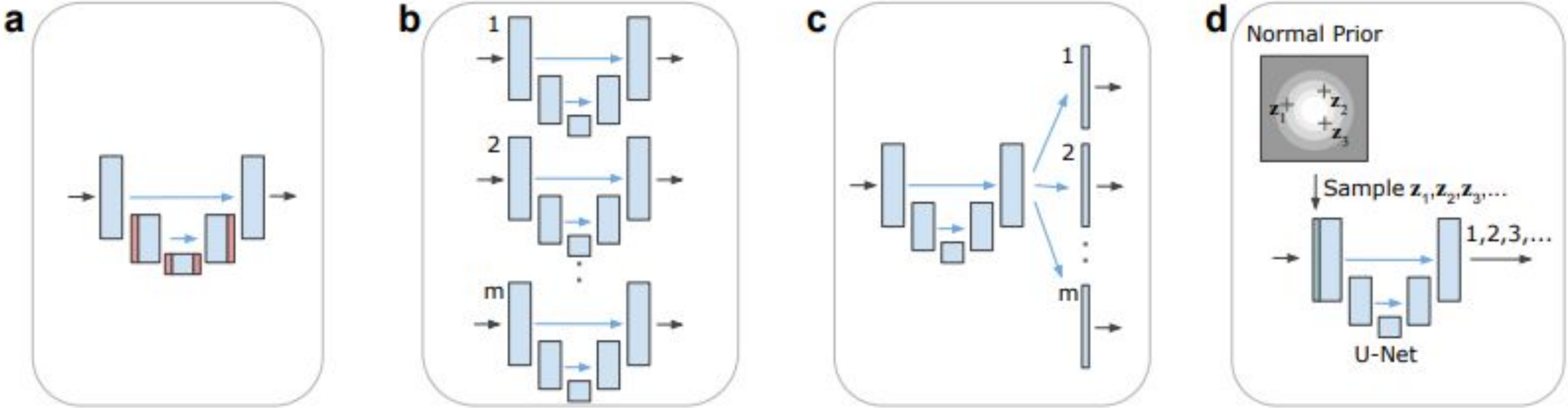


Figure 2: Baseline architectures. Arrows: flow of operations; blue blocks: feature maps; red blocks: feature maps with dropout; green block broadcasted latents. Note that the number of feature map blocks shown is reduced for clarity of presentation. (a) Dropout U-Net. (b) U-Net Ensemble. (c) M-Heads. (d) Image2Image VAE.

Experiments

- Experiment Dataset

- Lung abnormalities segmentation (LIDC)
  - 1018 lung CT scans from 1010 lung patients
  - 4 radiologists (from a total of 12) provided annotation masks for lesions
  - CT scans to 0.5 mm × 0.5 mm in-plane resolution
  - cropped 2D images (180 × 180 pixels) centered at the lesion positions

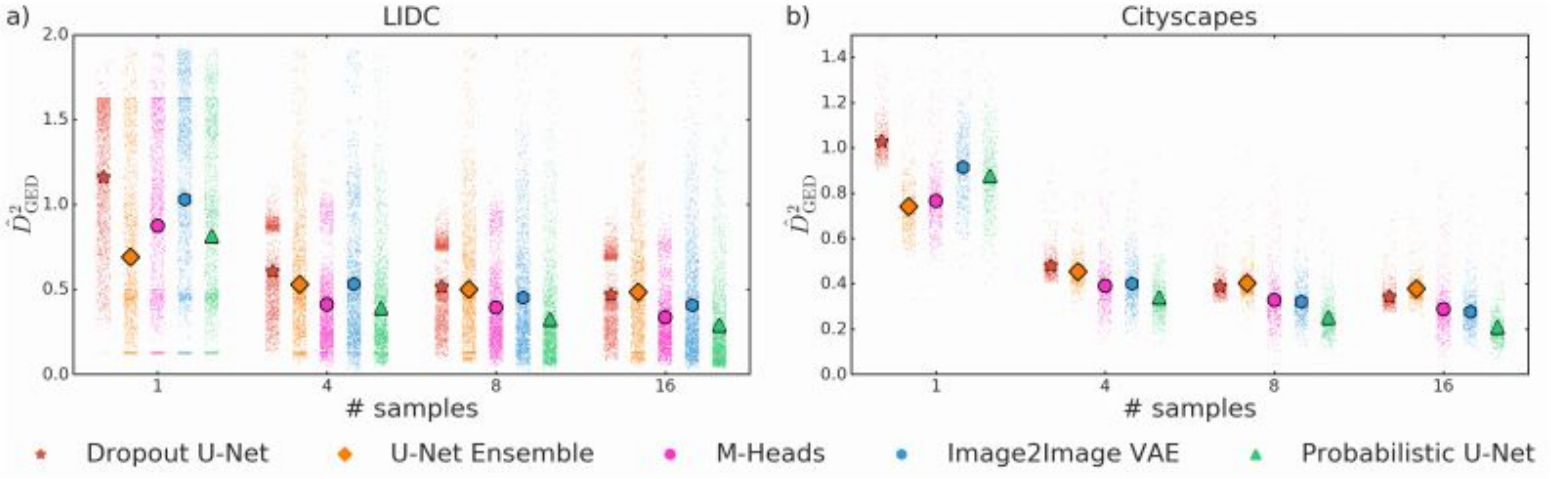
	Train	Valid	Test
Patient	722	144	144
Slice	8882	1996	1992

- Cityscapes
  - Create ambiguities by artificial random flips of five classes to newly introduced classes.
    - 'sidewalk' to 'sidewalk 2' with a probability of 8/17
    - 'person' to 'person 2' with a probability of 7/17
    - 'car' to 'car 2' with 6/17
    - 'vegetation' to 'vegetation 2' with 5/17
    - 'road' to 'road 2' with probability 4/17.
  - 2^5 = 32 discrete modes with probabilities ranging from 10.9% (all unflipped) down to 0.5% (all flipped)

	Train	Valid	Test
Slice	2957	274	500

Results

- Comparison of approaches using the squared energy distance

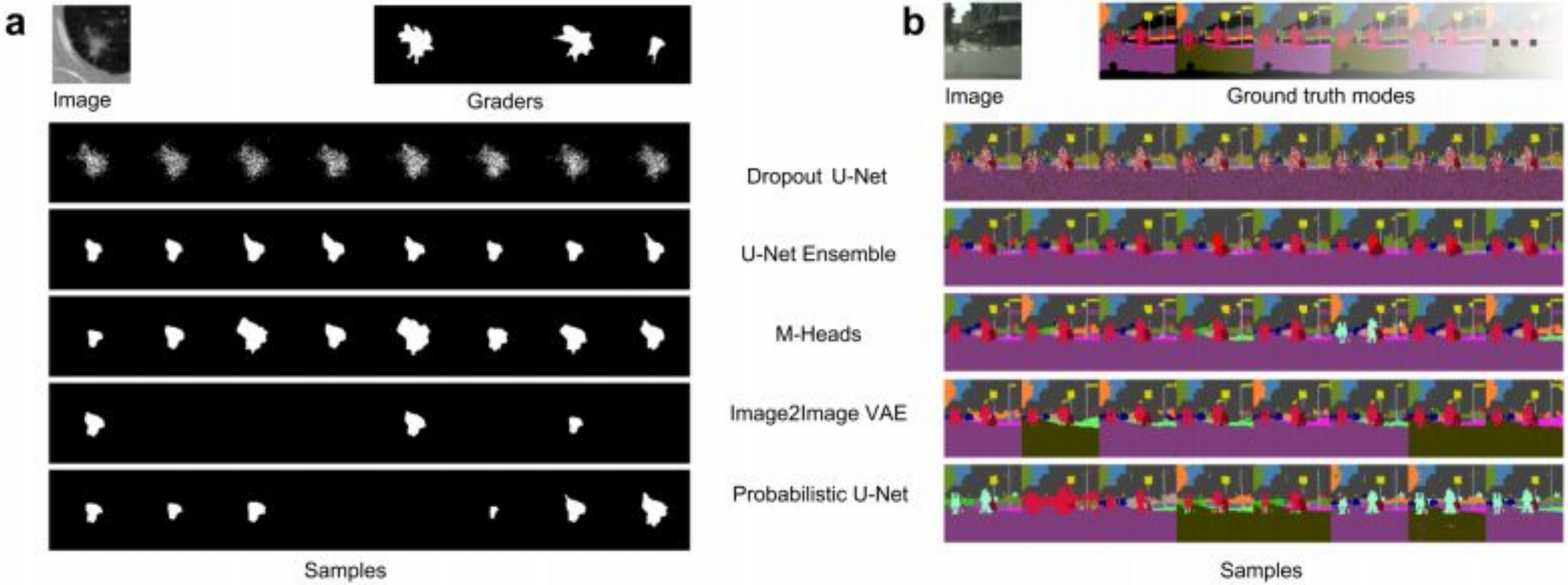


(m = 4, n samples = 1,2,8,16)

Lower energy distances correspond to better agreement between predicted distributions and ground truth distribution of segmentations. The symbols that overlay the distributions of data points mark the mean performance.

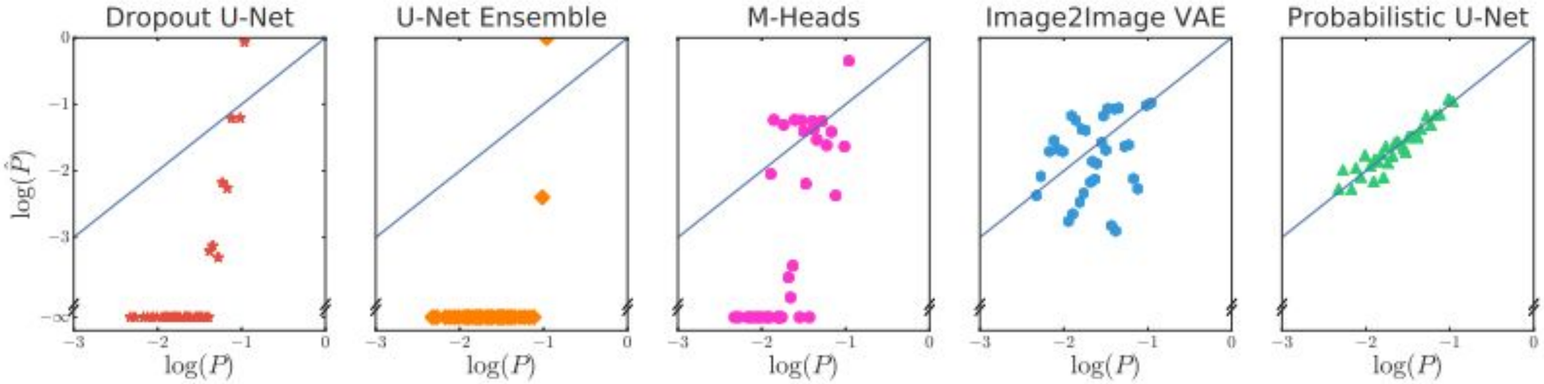
Results

- Qualitative results



Results

- Reproduction of the probabilities of the segmentation modes on the Cityscapes task.



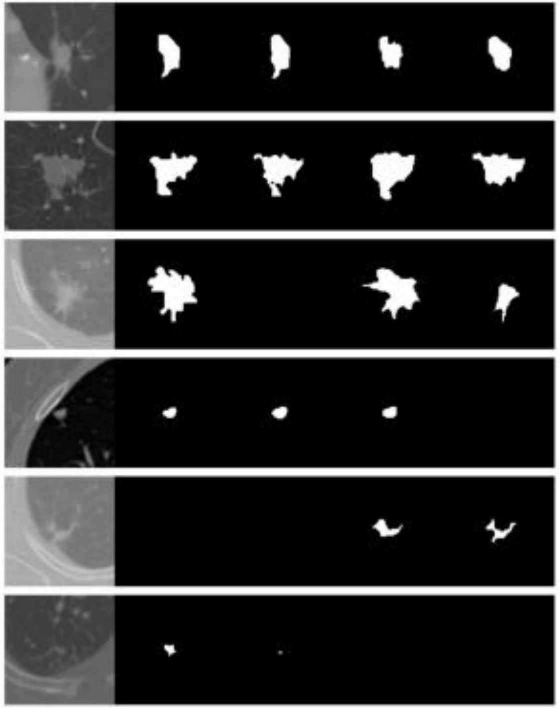
The artificial flipping of 5 classes results in 32 modes with different ground truth probability (x-axis).
The y-axis shows the frequency of how often the model predicted this variant in the whole test set.
Agreement with the bisector line indicates calibration quality.

Conclusions

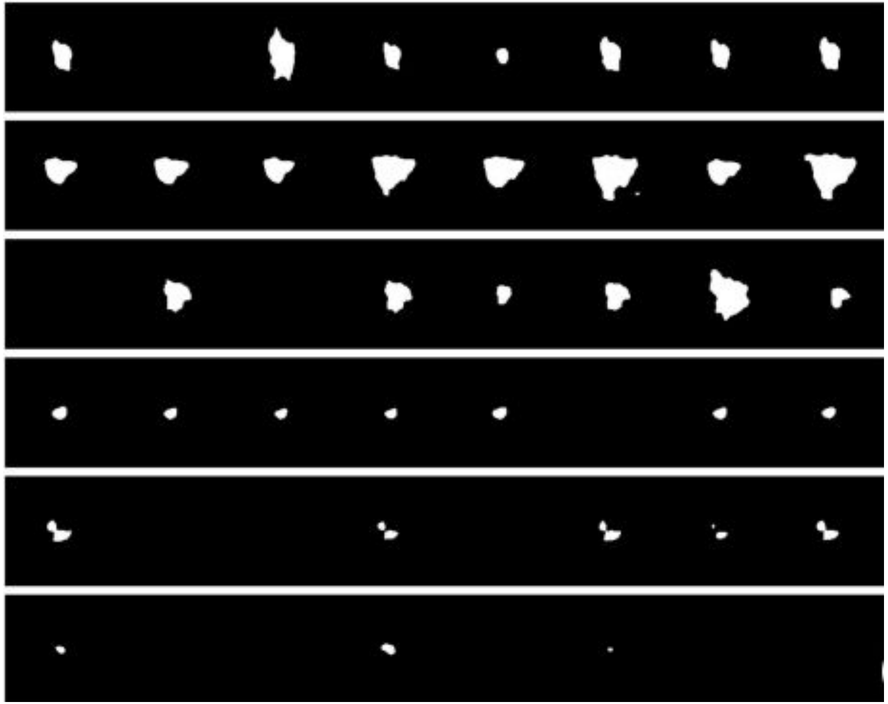
- Our proposed architecture provides consistent segmentation maps that closely match the multi-modal ground-truth distributions
- model scales to complex output distributions including the occurrence of very rare modes.
- Could prove useful beyond explicitly multi-modal tasks
- Especially in the medical domain, with its often ambiguous images and highly critical decisions that depend on the correct interpretation of the image, our model's segmentation hypotheses and their likelihoods could 1) inform diagnosis/classification probabilities or 2) guide steps to resolve ambiguities.

Appendix

- Images are often Ambiguous



Potential Cancer Expert Graders

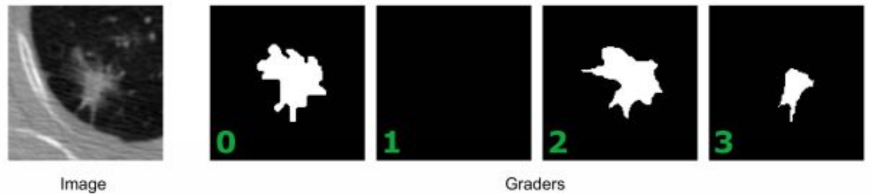


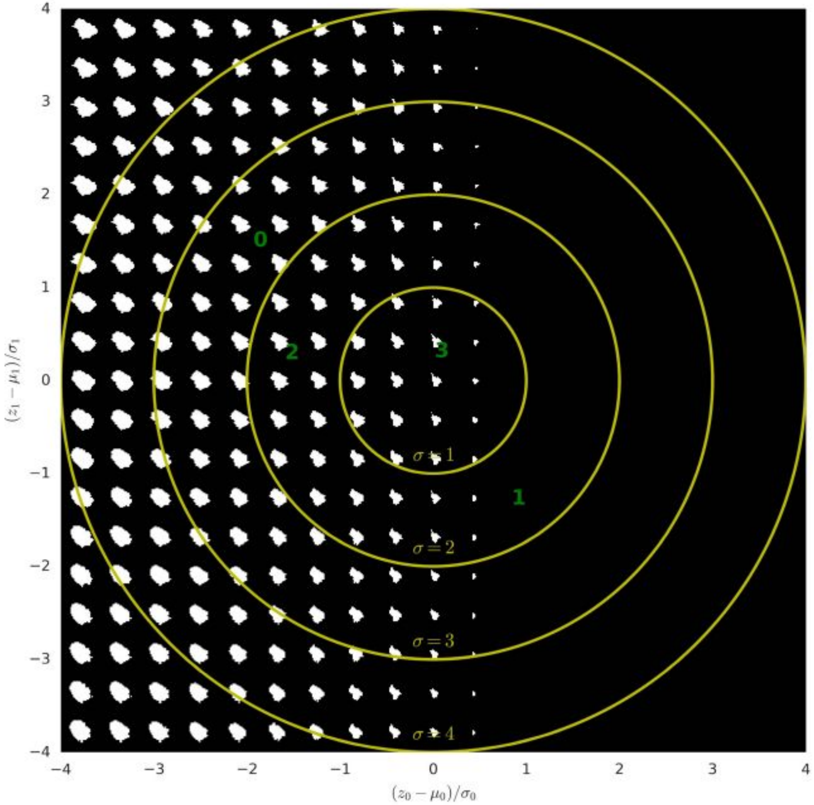
U-net + CVAE segmentation results

Appendix

- Visualization of the latent space for the lung abnormalities segmentation

- The latent space is re-scaled so that the prior likelihood is a spherical unit-Gaussian. The isoprobable yellow circles denote deviations from the mean in sigma.
- The ground-truth grader masks' posterior position in this latent space is indicated by green numbers.
- The input image is shown in the lower left, to the right of it, the 4 grader masks are shown.

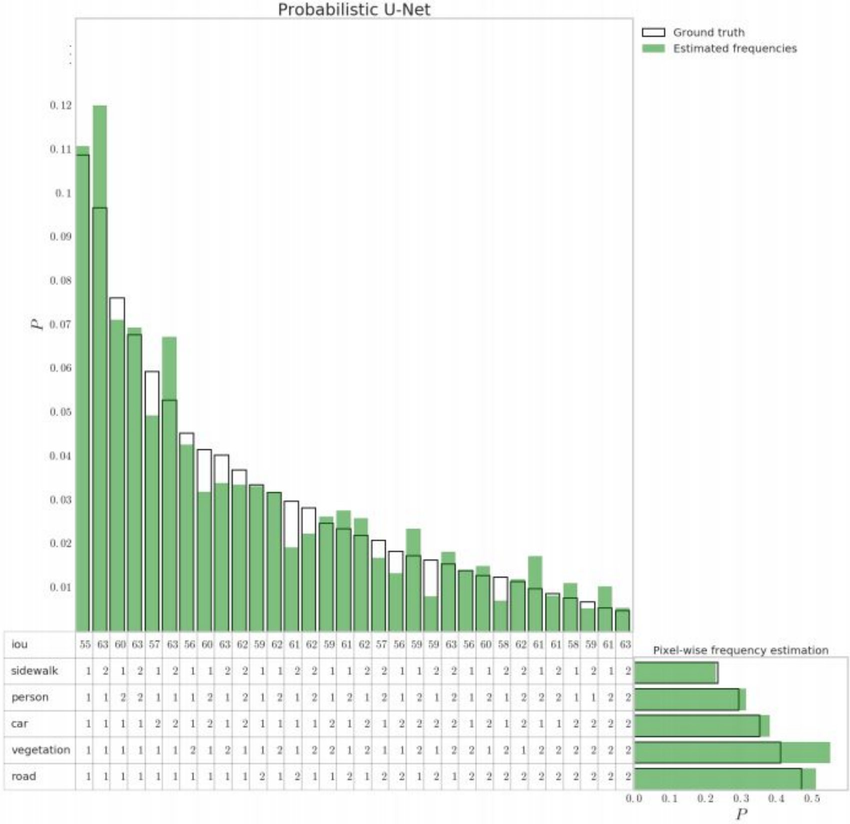




Appendix

- Reproduction of probabilities by our Probabilistic U-Net.

The vertical histogram shows the mode-wise occurrence frequencies of samples in comparison to the ground-truth probability of the modes, and the horizontal histogram reports the pixel-wise marginal frequencies, i.e. the sampled pixel-fractions for each new stochastic class (e.g. sidewalk 2) with respect to the corresponding existing one (sidewalk).

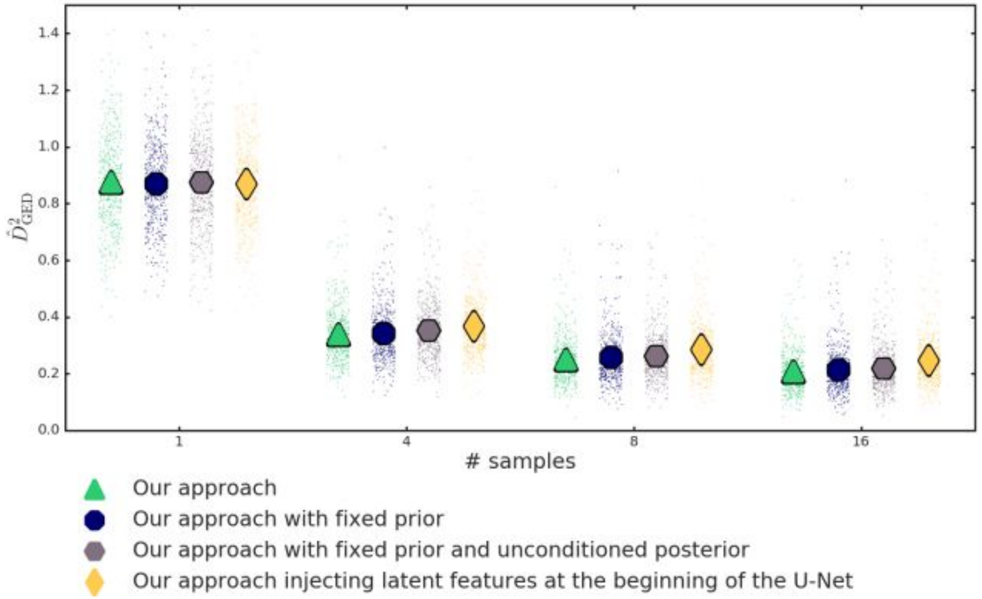


iou	55	63	60	63	57	63	56	60	63	62	59	62	61	62	59	61	62	57	56	59	59	63	56	60	58	62	61	61	58	59	61	63
sidewalk	1	2	1	2	1	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	2	1	2	1	2
person	1	1	2	2	1	1	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	2	2	1	1	2	2
car	1	1	1	1	2	2	1	2	1	2	1	1	1	1	2	1	2	1	2	2	2	2	1	2	1	2	1	1	2	2	2	2
vegetation	1	1	1	1	1	1	2	1	2	1	1	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1	2	2	2	2	2	2
road	1	1	1	1	1	1	1	1	1	1	2	1	2	1	1	2	1	2	2	1	2	1	2	2	2	2	2	2	2	2	2	2

Appendix

- Ablation Analysis

- **Fixing the prior:** Instead of making the prior a function of the context, fix to be a Gaussian distribution.
- **Fixing the prior, and not using the context (input image) in the posterior:** In addition to fixing the prior to be Gaussian, also make the posterior a function of the ground truth mask only, ignoring the context.
- **Injecting the latent features at the beginning of the U-Net:** change the position in which the latent variables are used.


$\hat{D}^2_{\mathrm{GED}}$
# samples
▲ Our approach
● Our approach with fixed prior
⬢ Our approach with fixed prior and unconditioned posterior
◆ Our approach injecting latent features at the beginning of the U-Net