Probabilistic U-Net for Segmentation of Ambiguous Images

Original Paper: Kohl, Simon A. A., Romera-Paredes, et al. A Probabilistic U-Net for Segmentation of Ambiguous Images. (2019). arxiv:1806.05034 Advances in Neural Information Processing Systems

Presentation author: Krisztina Sinkovics

Venue: nPlan ML Paper Club (virtual)

Date: 30th of August 2020

Task

- Learn a distribution of segmentations given an input
- On a class of images where the image context alone is not enough to resolve ambiguities

- e.g. on medical images:

lesion (anomalous region) ≠ cancer

Problem Setup

- Pixel-wise probabilities vs
- Covariance between the pixels

- The most likely hypothesis vs
- Multiple hypothesis

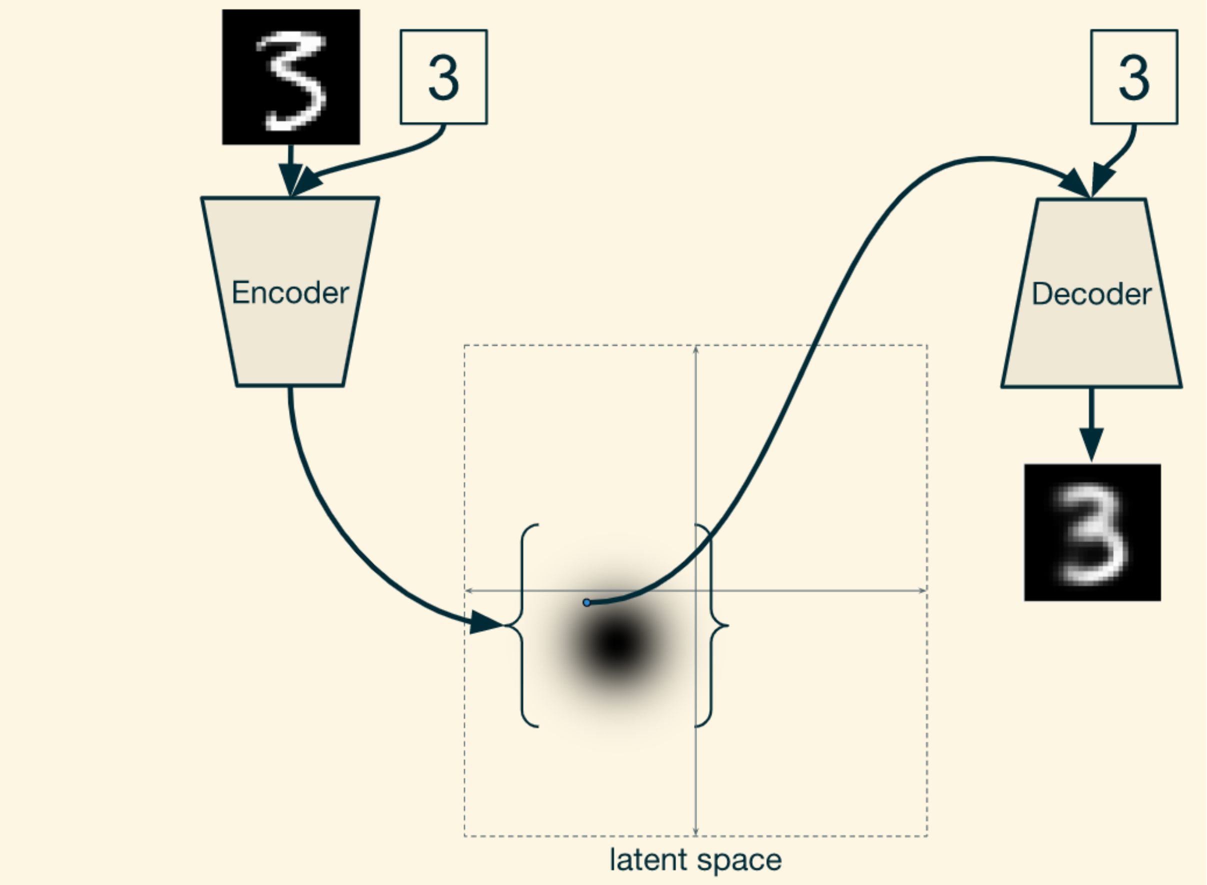
Misdiagnosis, sub-optimal treatment

Subsequent tests to resolve multiple ambiguities

CVAE

Objective:

$$\mathcal{L}(Y, X) =$$
$$\mathrm{E}\left[-\log P_c(X|z, c)\right] - D_{KL}(Q(z|X, c)||P(z|c))$$



A conditional variational autoencoder

How they do it (U-Net + cVAE)

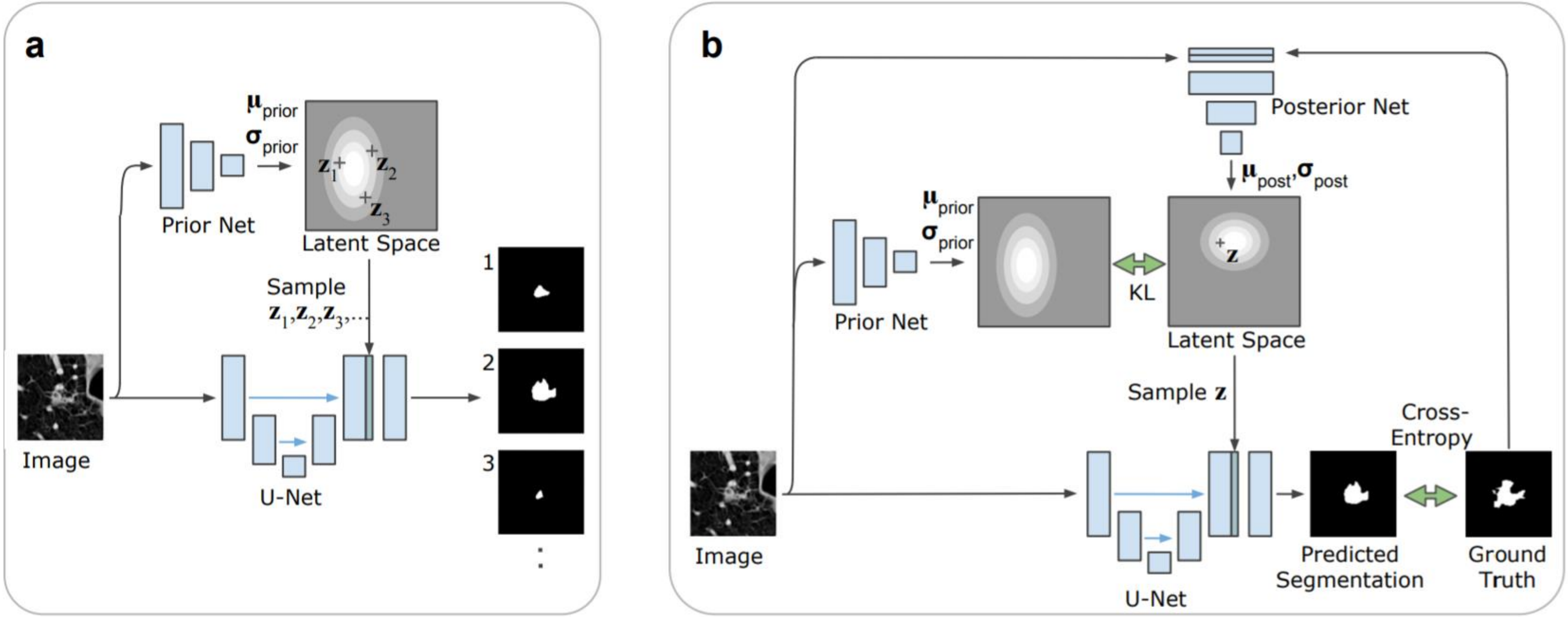


Figure 1: The Probabilistic U-Net.

(a) Sampling process for inference. Arrows: flow of operations; blue blocks: feature maps. The heatmap represents the probability distribution in the low-dimensional latent space RN (e.g., N = 6 in our experiments). For each execution of the network, one sample z ∈ RN is drawn to predict one segmentation mask. Green block: N-channel feature map from broadcasting sample z. The number of feature map blocks shown is reduced for clarity of presentation.

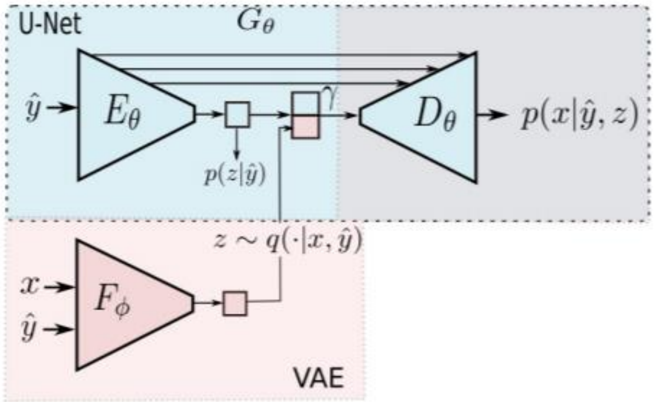
(b) Training process illustrated for one training example. Green arrows: loss functions.

How others do it

- Dropout U-Net
  + Probability distribution using dropout over spatial features → quantified pixel-wise uncertainty
  - Inconsistent outputs
- Ensemble of U-Nets trained separately
  + Consistent outputs
  - Outputs not diverse
  - Not able to learn rare variants, only most likely hypotheses
  - Does not scale well to large # of hypotheses
  - Need to fix # of hypotheses at training
- M-Heads
  + Captures diverse set of variants
  - But not occurrence of individual variants
  - Does not scale well to large # of hypotheses
  - Need to fix # of hypotheses at training
- Graphical models (Junction Chains, Markov Random Fields)
  + Captures diverse set of variants
  - Confined to structured problems ← tractable graphical models

Related work from Image2Image translation

- GANs
  - Suffer from mode collapse
  - ➢ Solved in "bicycleGAN"
    - CVAE - GAN + Conditional latent regressor GAN
    - Fixed prior distribution
    - Posterior only conditioned on output
- VAE + U-Net for generating appearances given a shape encoding
  - Involves pre-trained VGG19 that measures perceptual similarity and feeds into reconstruction loss



- Probabilistic model for Structured outputs
  - Optimizing dissimilarity coefficient between ground truth and predicted distributions
  - Assessed on hand pose estimation -> predict position on 14 joints

Contributions

1) Consistent segmentation maps of pixel-wise probabilities → joint likelihood of model
2) Arbitrarily complex output distributions
   - Rare modes
   - Calibrated probabilities of segmentation modes
3) Computationally cheap sampling
4) Assessing performance quantitatively

Network Architecture: Sampling for inference

(formal explanation to Figure 1. a))

$$z_i \sim \mathrm{P}(\cdot\,|X) = N\left(\mu_{prior}(X,\omega), diag\left(\sigma_{prior}(X,\omega)\right)\right) \quad (1)$$

z_i - a random sample

$\mathrm{P}(\cdot\,|X)$ - prior, axis-aligned Gaussian. Conditioned on image, enabling it to capture variant frequencies by allocating corresponding probability mass to the respective latent space regions.

$\mu_{prior}(X,\omega) \in \mathbb{R}^N$ − mean

$\sigma_{prior}(X,\omega) \in \mathbb{R}^N$ − variance

ω – prior net weights

X – input image

$$S_i = f_{comb.}(f_{U-Net}(X,\theta), z_{i,} \psi) \quad (2)$$

S_i - segmentation map

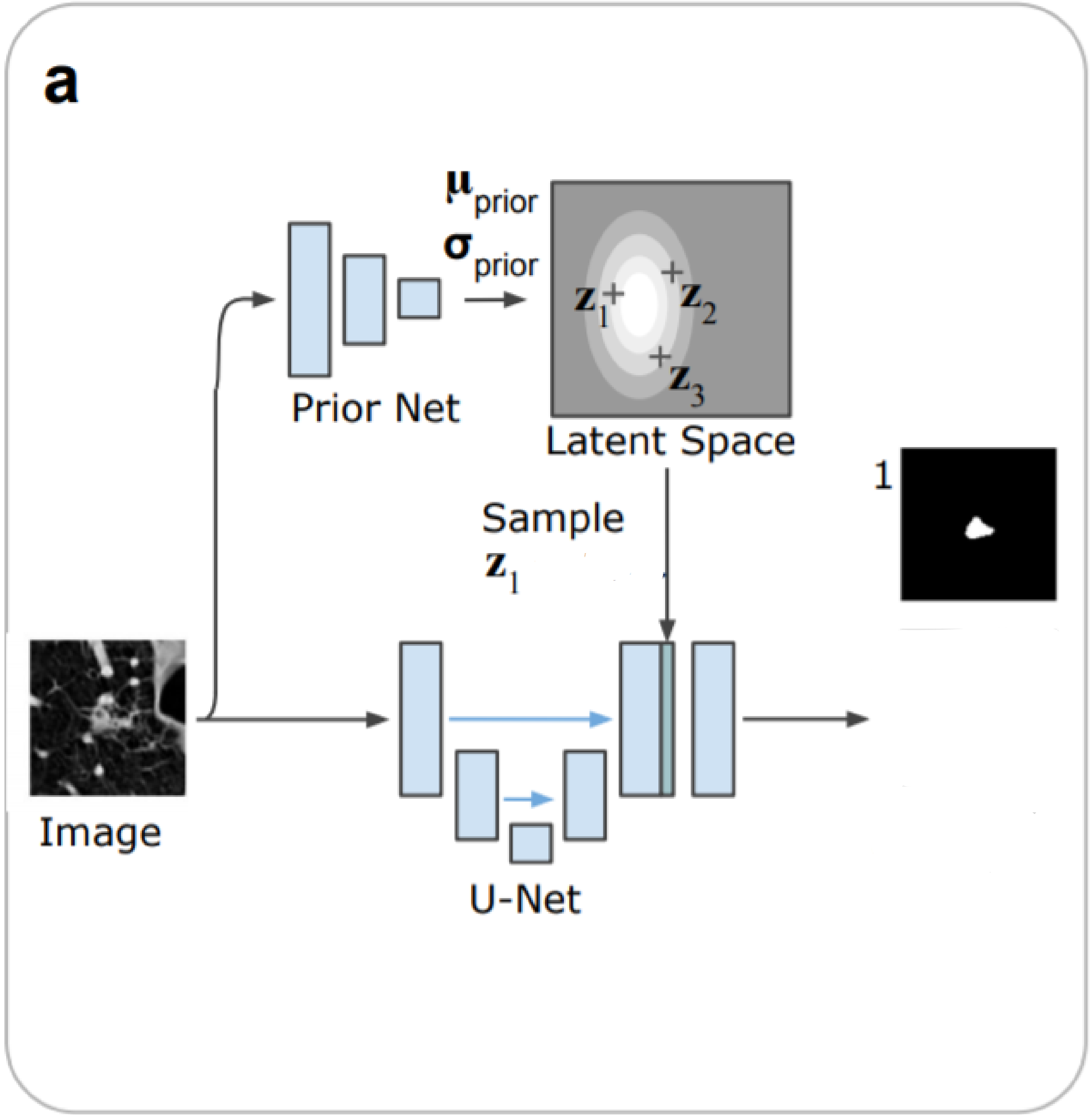
$f_{comb.}$ - three consequent 1X1 Convolutions

ψ – their weights

Sampling (for inference)

Figure 1. a)

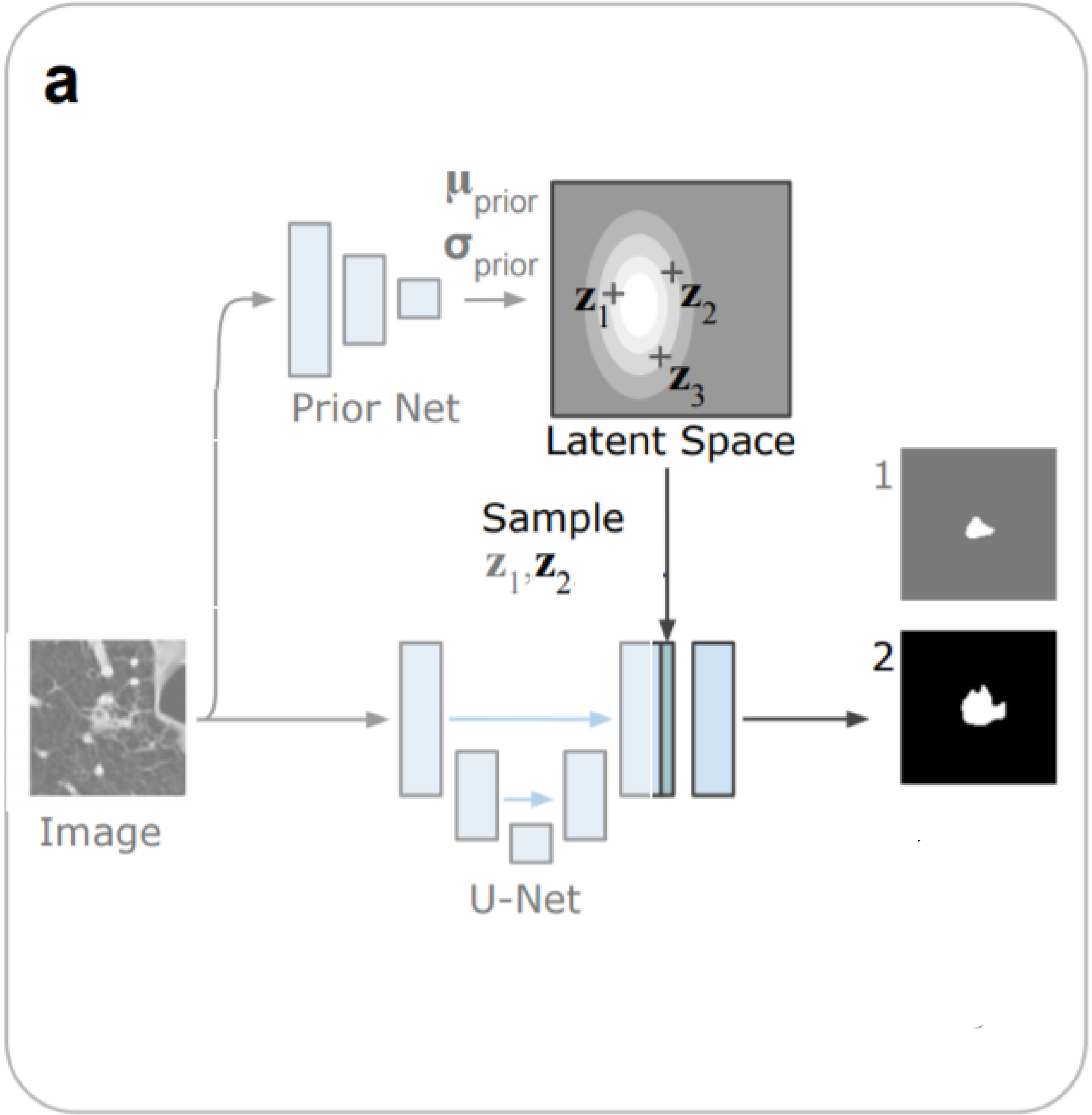
- Repeat m times



Sampling (for inference)

Figure 1. a)

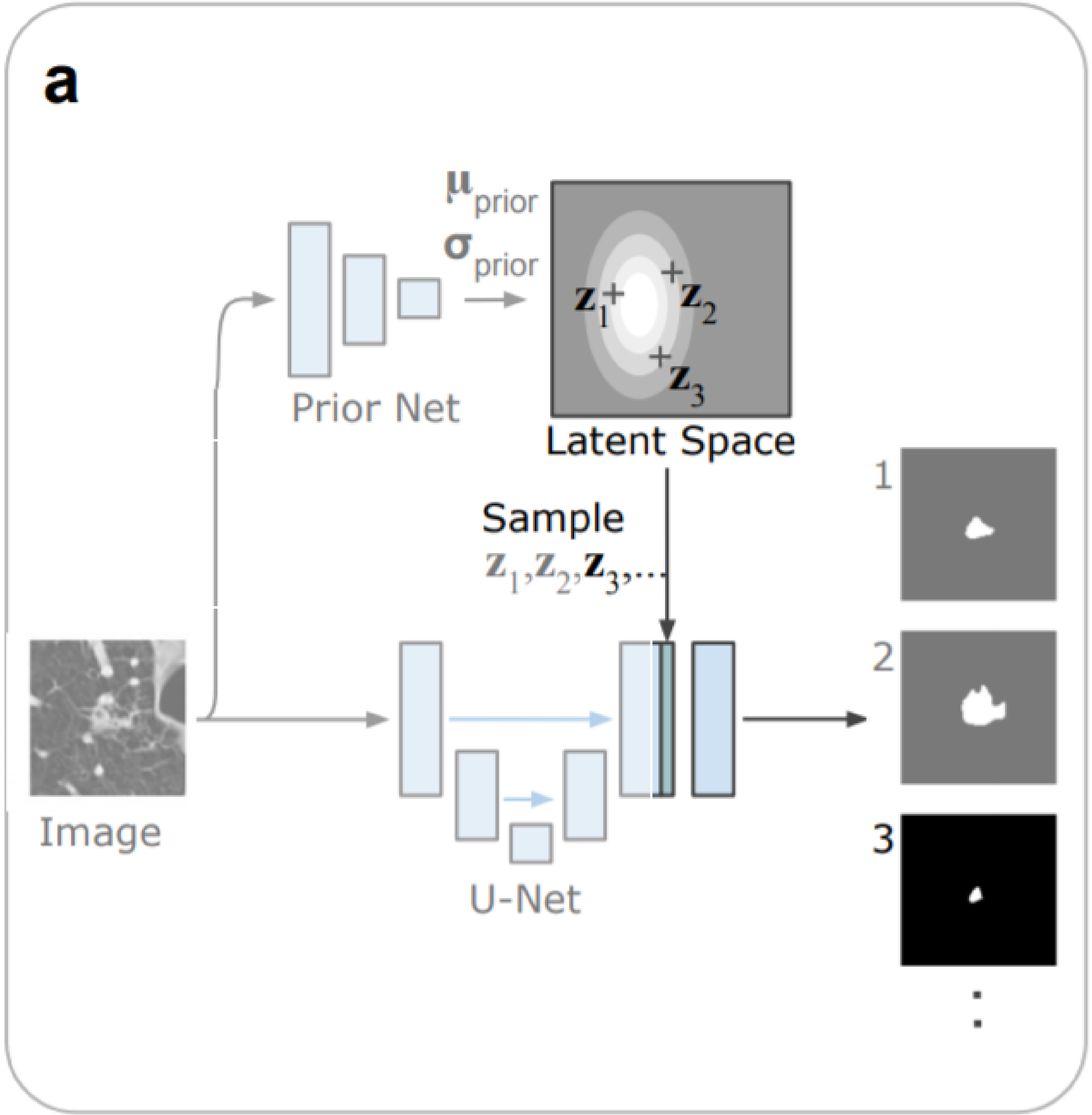
- Repeat m times



Sampling (for inference)

Figure 1. a)

- Repeat m times



Network Architecture: Training

(formal explanation to Figure 1. b))

$$z \sim Q(\cdot|Y,X) = N\left(\mu_{post}(X,Y;\ \upsilon),\ diag\ \sigma_{post}(X,Y;\ \nu)\right) \qquad (3)$$

z - a random sample from posterior $Q(\cdot|X)$

$\mu_{post}(X,Y;\ \nu)\ \in\ \mathbb{R}^N$ - posterior mean

$\sigma_{post}(X,Y;\ \nu)\ \in\ \mathbb{R}^N$ - posterior variance

ν – posterior net weights

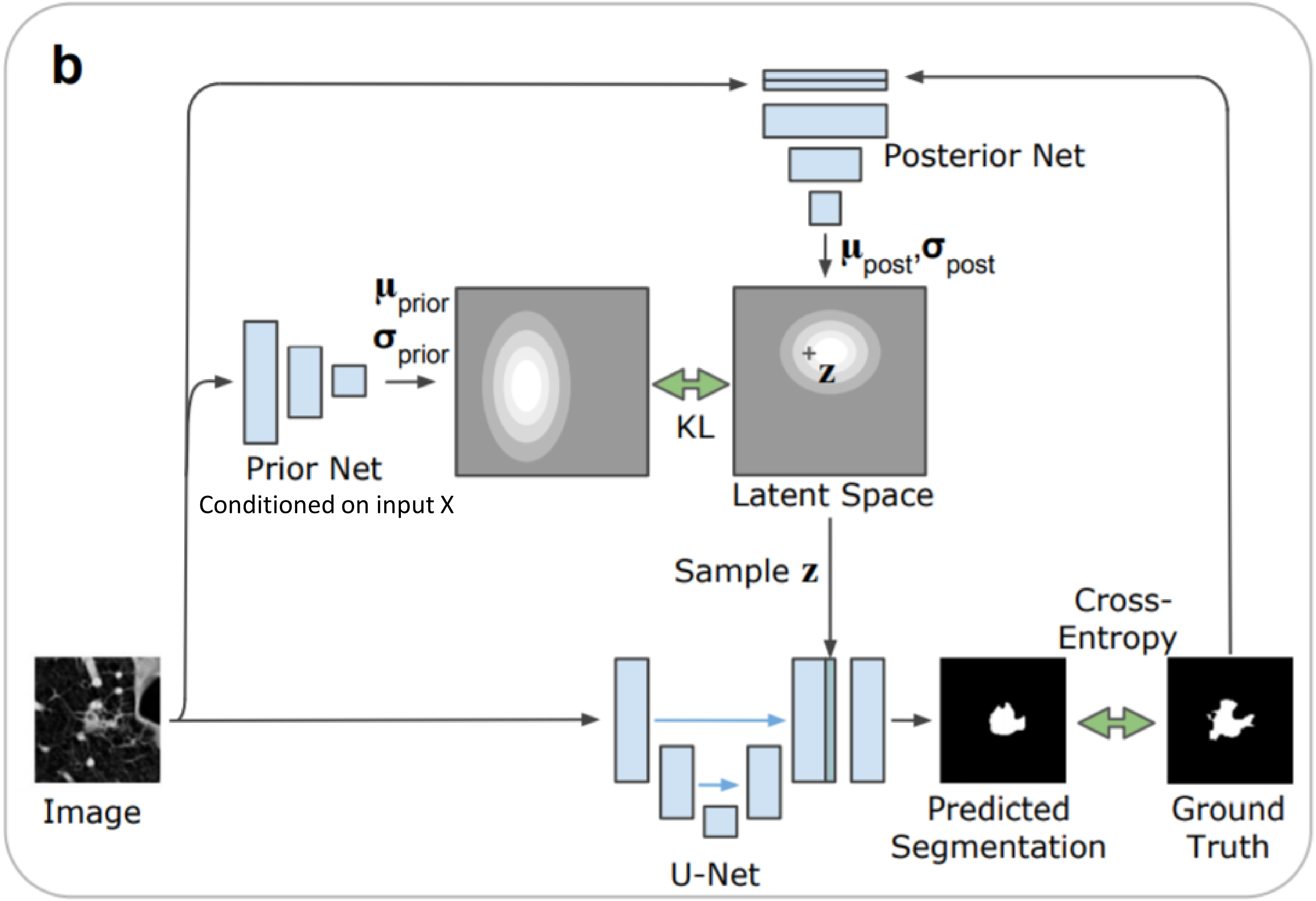
Y – segmentation mask

(1) + (3) → S - predicted segmentation ideally identical to Y.

Loss (4)

$$\mathcal{L}(Y,X) = \overbrace{\underbrace{\mathrm{E}_{Z\sim Q(\cdot|Y,X)}[-\log P_c(Y|S(X,z))]}_{\text{Cross-Entropy loss between (S and Y)}} - \underbrace{\beta}_{\text{Weighting factor}} * \underbrace{D_{KL}(Q(z|Y,X)||P(z|X))}_{\text{Kullback-Leibler Divergence between posterior Q and prior P}}}^{-\text{ ELBO}}$$

Network Architecture: Training (Figure 1. b)



Performance Measures

Use **G**eneralized **E**nergy **D**istance to compare distributions of segmentations

$$D^2_{GED}\left(P_{gt}, P_{out}\right) = \underbrace{2\mathrm{E}[\mathrm{d}(\mathrm{S},\mathrm{Y})]}_{\text{Disagreement between gt and predicted sample}} - \underbrace{\mathrm{E}[\mathrm{d}(\mathrm{S},S')]}_{\text{Disagreement between a pair of predicted masks}} - \underbrace{\mathrm{E}[\mathrm{d}(\mathrm{Y},\mathrm{Y}')]}_{\text{Disagreement between a pair of gt masks}}$$

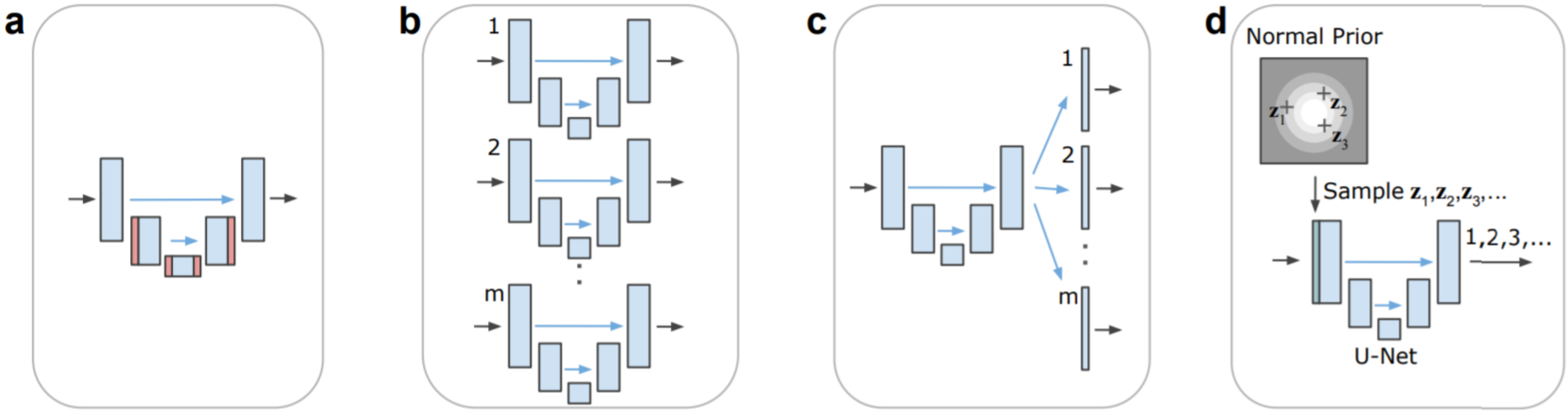
$\mathrm{S},\ S'$ – independent samples from predicted distribution

$\mathrm{Y},\ Y'$ – independent samples from ground truth masks

$d(x,y) = 1 - IOU(x,y)$ distance measure

When S and Y are empty $d(S,Y) = 0$

Baseline Methods



Dropout U-Net **U-Net Ensemble** **M-Heads** **Image2Image VAE**

Figure 2: Baseline architectures
Arrows: flow of operations
blue blocks: feature maps
red blocks: feature maps with dropout with probability p=0.5
green block broadcasted latents.
Note that the number of feature map blocks shown is reduced for clarity of presentation.

Results: Lung Abnormalities Segmentation

- Setup:
- 1018 lung CT scans
- from 1010 lung patients
- For each scan 4 radiologists (from a total of 12) provided annotation masks

- Resampled CT scans to 0.5 mm × 0.5 mm in-plane resolution
- Cropped 2D images (180 × 180 pixels) centered at the lesion positions

Training set	722 patients	8882 images
Validation set	144 patients	1996 images
Test set	144 patients	1992 images

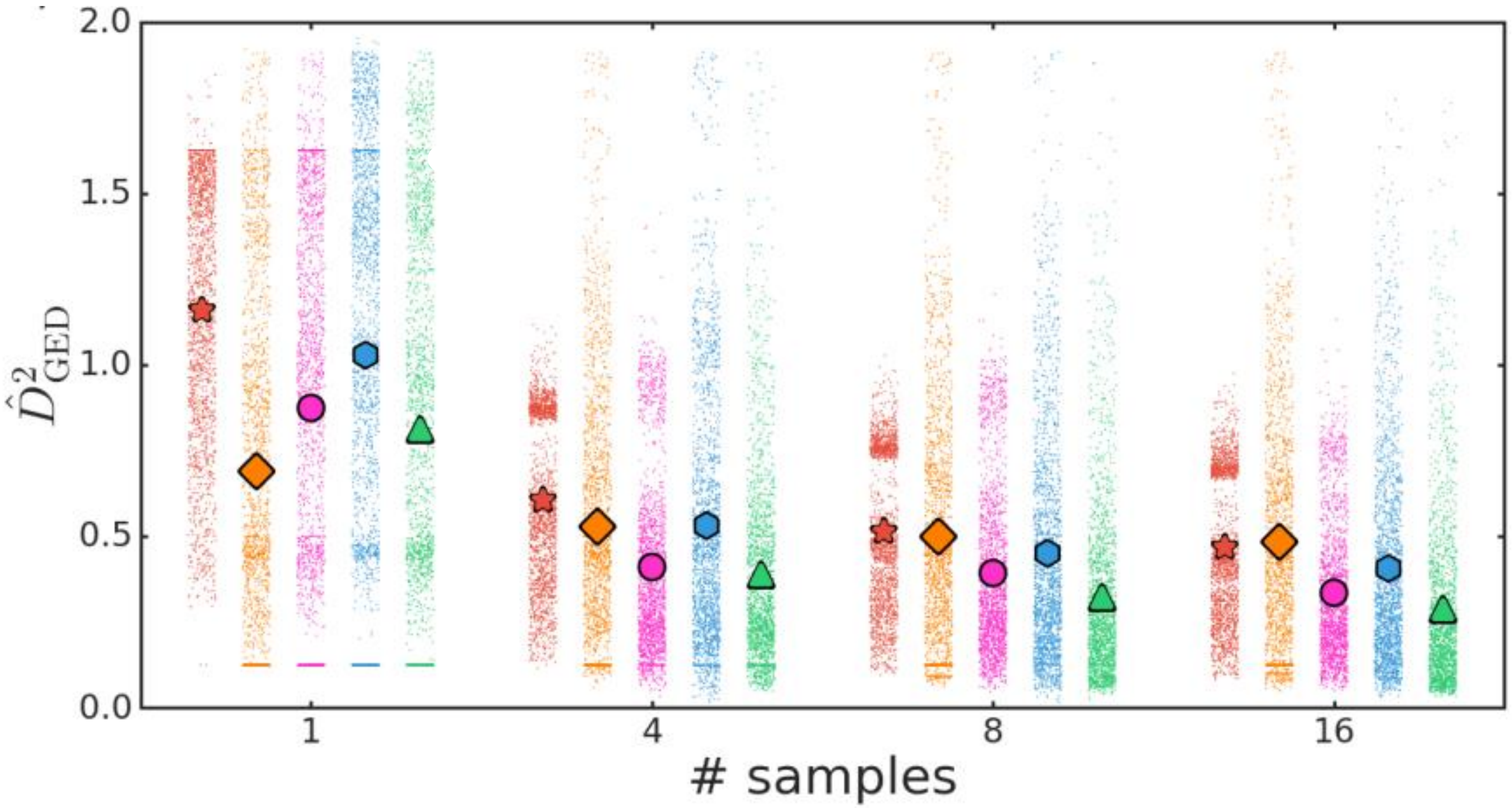
- 3 masks per image can be empty as experts can disagree

Results: Lung Abnormalities Segmentation

$$\widehat{D}^2_{GED}(P_{gt}, P_{out}) = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} d(S_i, Y_j) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d(S_i, S'_j) - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} d(Y_i, Y'_j)$$

m = 4

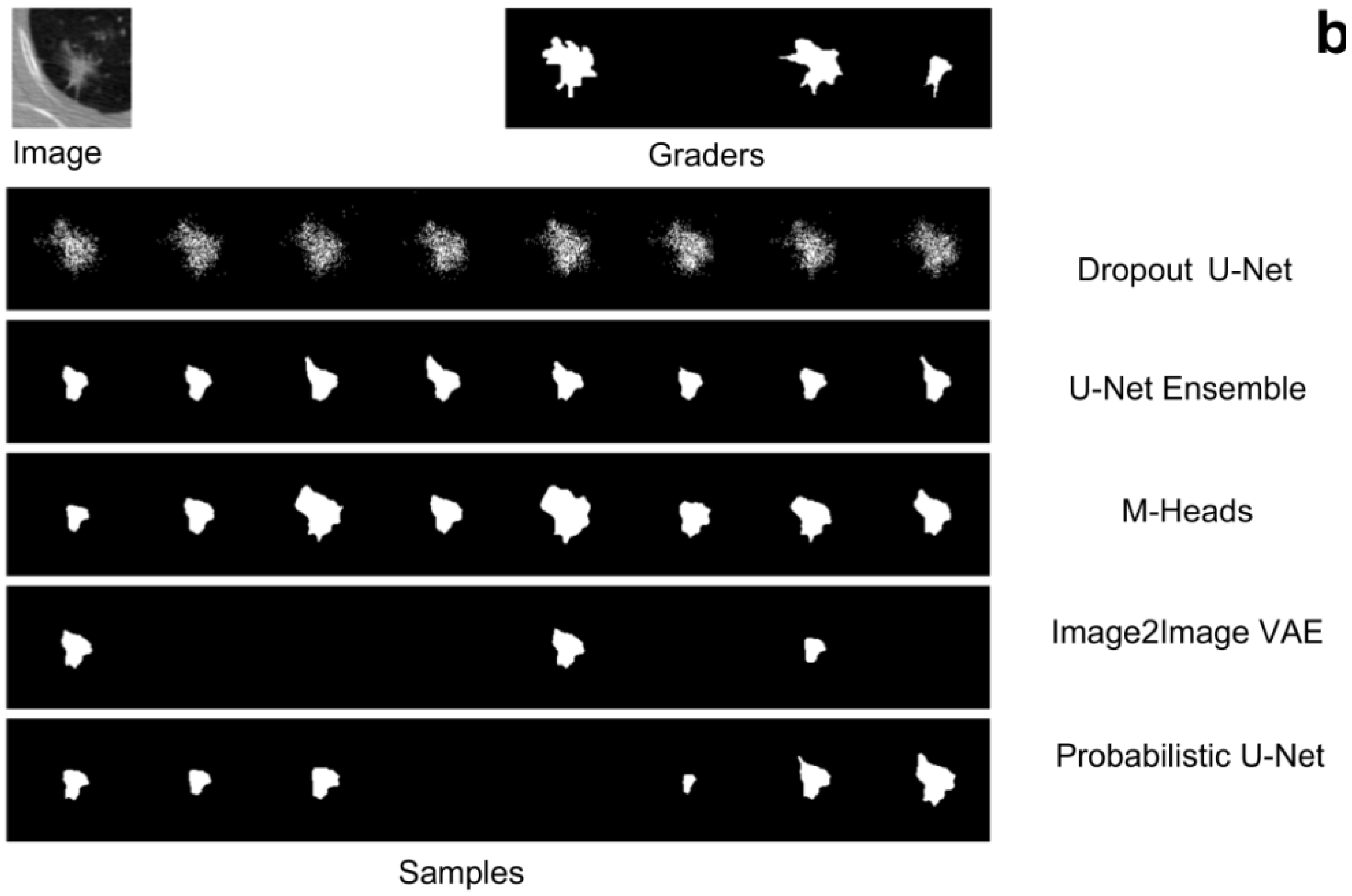
n = 1, 4, 8, 16



Dropout U-Net U-Net Ensemble M-Heads Image2Image VAE Probabilistic U-Net

Results: Lung Abnormalities Segmentation

Qualitative



Results: Cityscapes Semantic Segmentation

- ## Setup:
- images of street scenes taken from a car with corresponding semantic segmentation maps
- 19 classes
- 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)

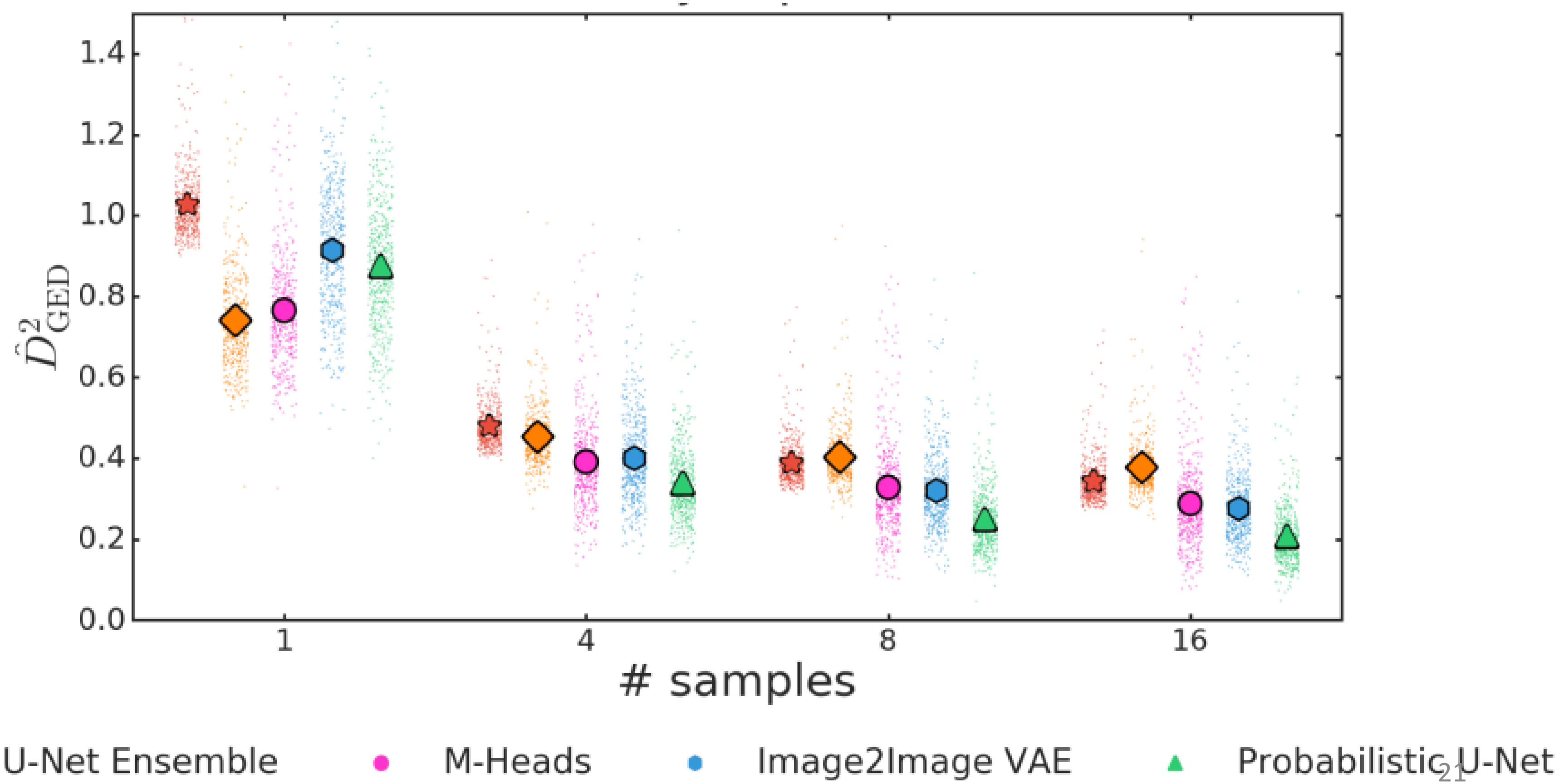
Training set	2975 images
Validation set	274 images (3 cities from official test set)
Test set	500 images (official test set)

Results: Cityscapes Semantic Segmentation

$$\widehat{D}^2_{GED}(P_{gt}, P_{out}) = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{M} d(S_i, Y_j)\,\omega_j - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d(S_i, S'_j) - \frac{1}{m^2}\sum_{i=1}^{M}\sum_{j=1}^{M} d(Y_i, Y'_j)\omega_i\omega_j$$

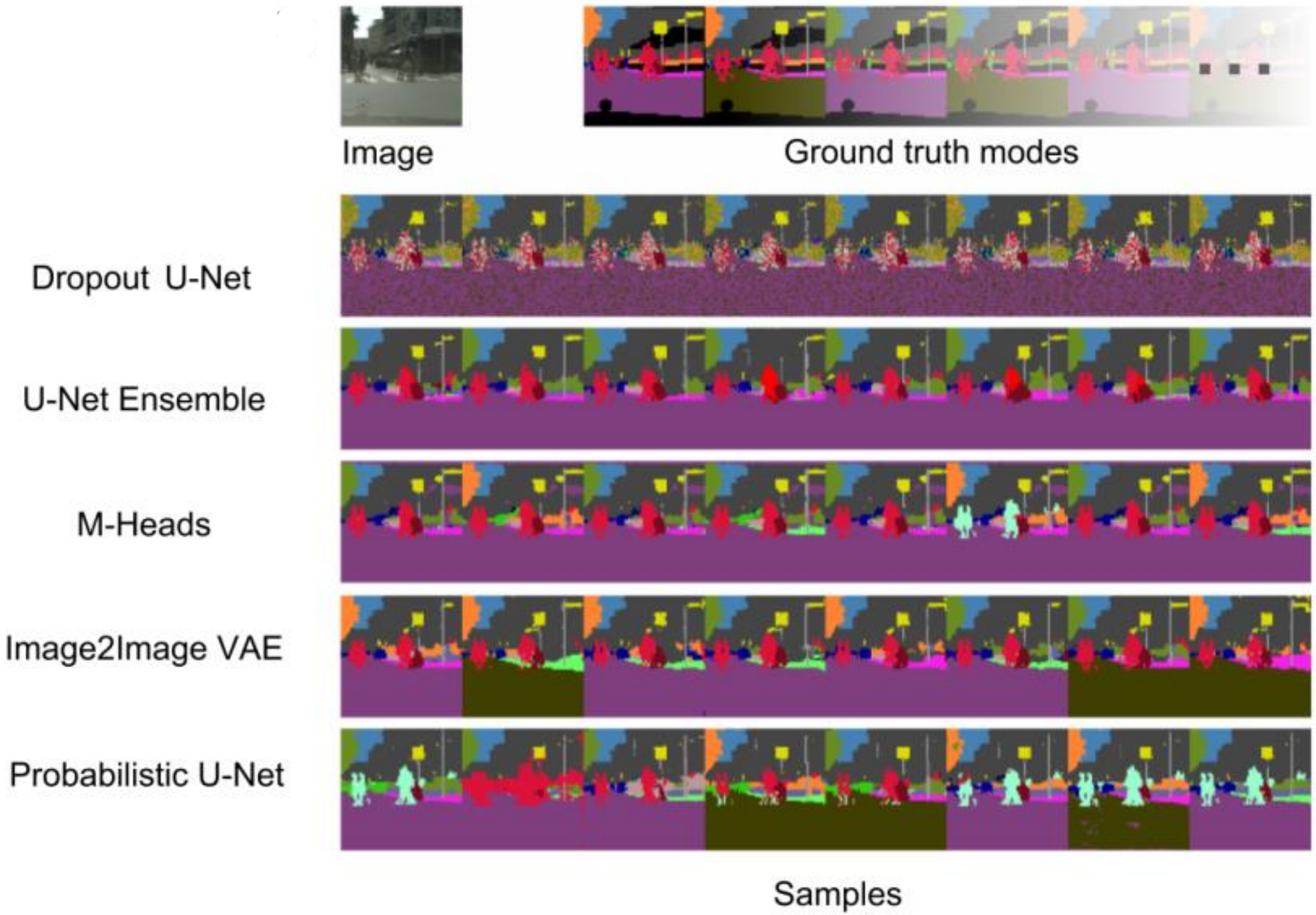
$M = 32$ Dirac delta distributions, used directly in the estimator; n = 1, 4, 8, 16;

ω_j - weight of the j-th mixture



Results: Cityscapes Semantic Segmentation

Qualitative



Reproducing segmentation probabilities (Cityscapes)

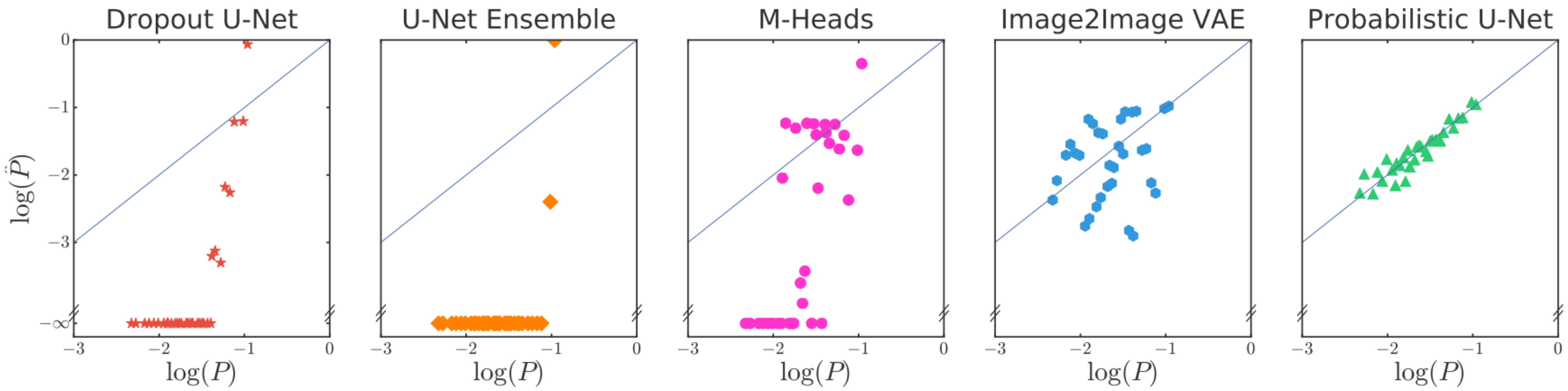
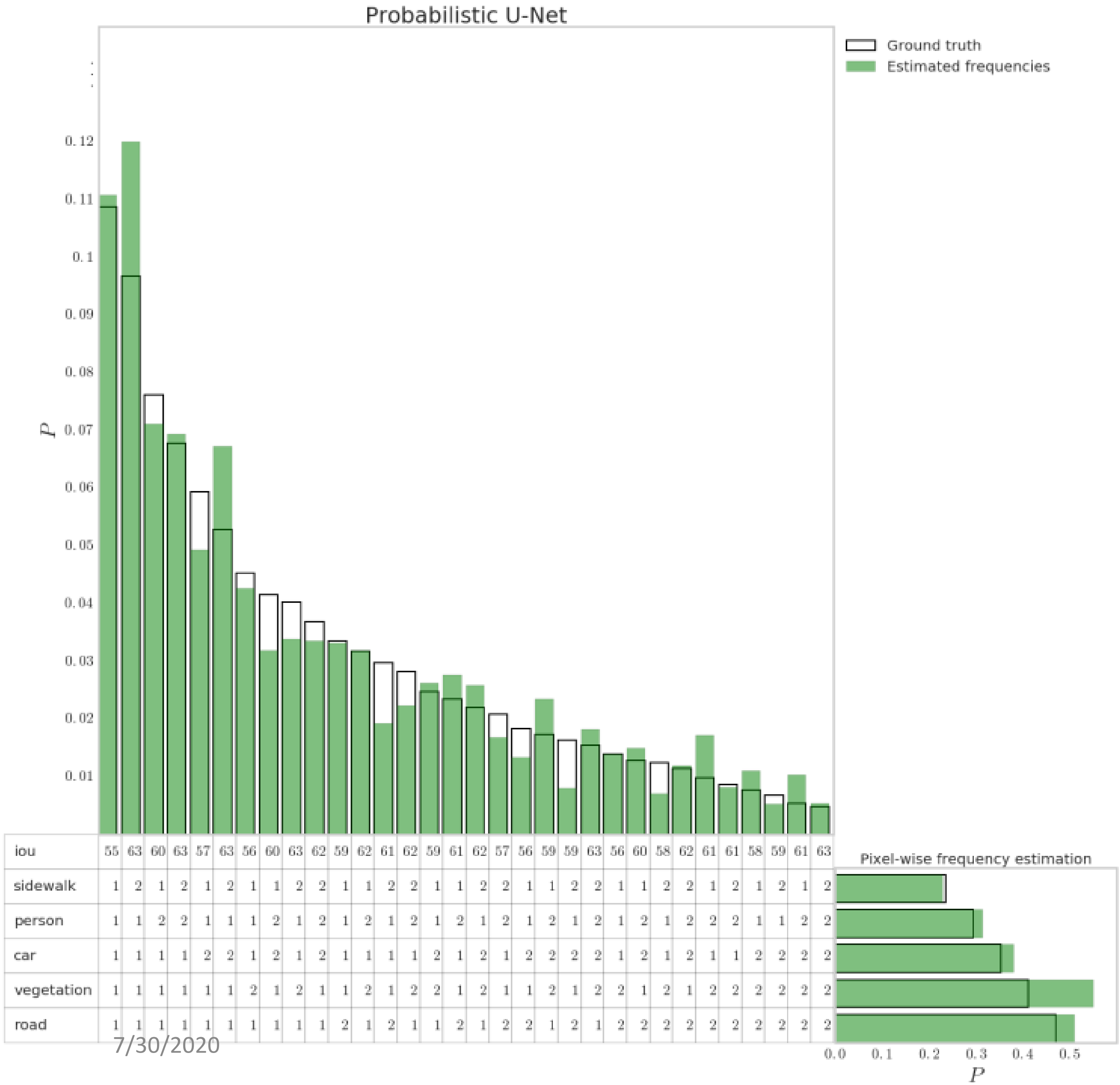


Figure 5.
The artificial flipping of 5 classes results in 32 modes with different ground truth probability (x-axis). The y-axis shows the frequency of how often the model predicted this variant in the whole test set. Agreement with the bisector line indicates calibration quality.

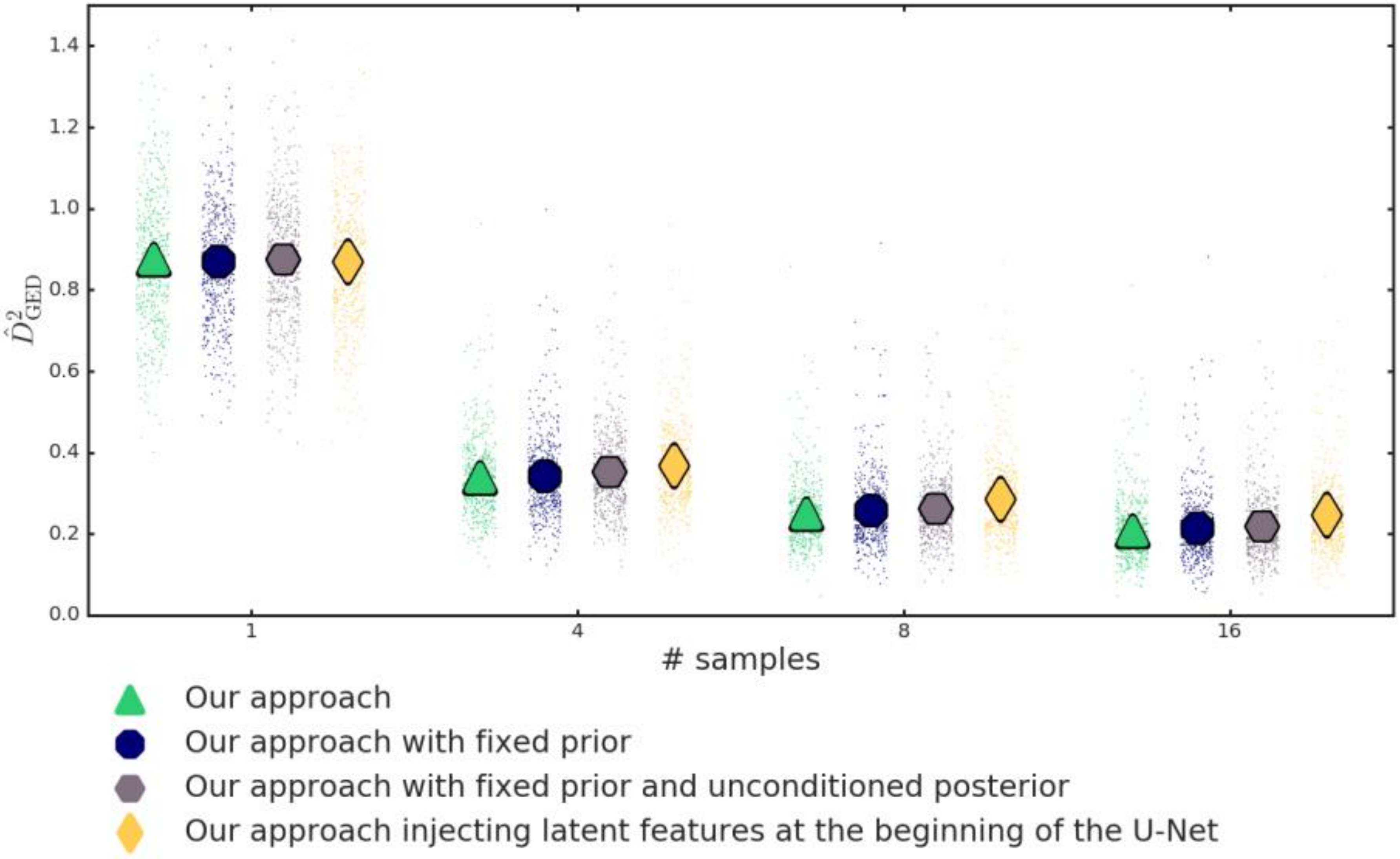


How the model fits ground truth distribution

Figure 10: 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)

Ablation analysis



Conclusions

- ✓ Probabilistic U-Net provides consistent segmentation masks that closely match multi-modal gt distributions.
- ✓ Captures complex output distributions including rare modes.
- ✓ Outperformed the baselines in 4,8 and 16 sample cases, by a significant (Wilcoxon signed-rank test yielding small p-value) but thin margin.
- ✓ Disentangles prior and segmentation net , conditioning on the entire image while allowing low computational cost.
- ✓ Architecture allows to inspect its latent space bc of VAE component.
- ✓ Experiment setup allows for in-depth performance evaluation.

– More support for the good model calibration claim would be welcome.

– Lack of description behind the choice of architecture components (e.g. why conditioning Prior and Posterior, why training Prior and Posterior networks, why Beta in the Loss).

Thank you!

Interesting posts

- Probabilistic U-Net's implementation on github (original)

Background

- Conditional Variational Autoencoders
- Conditional VAE
- Variational Inference

Training details

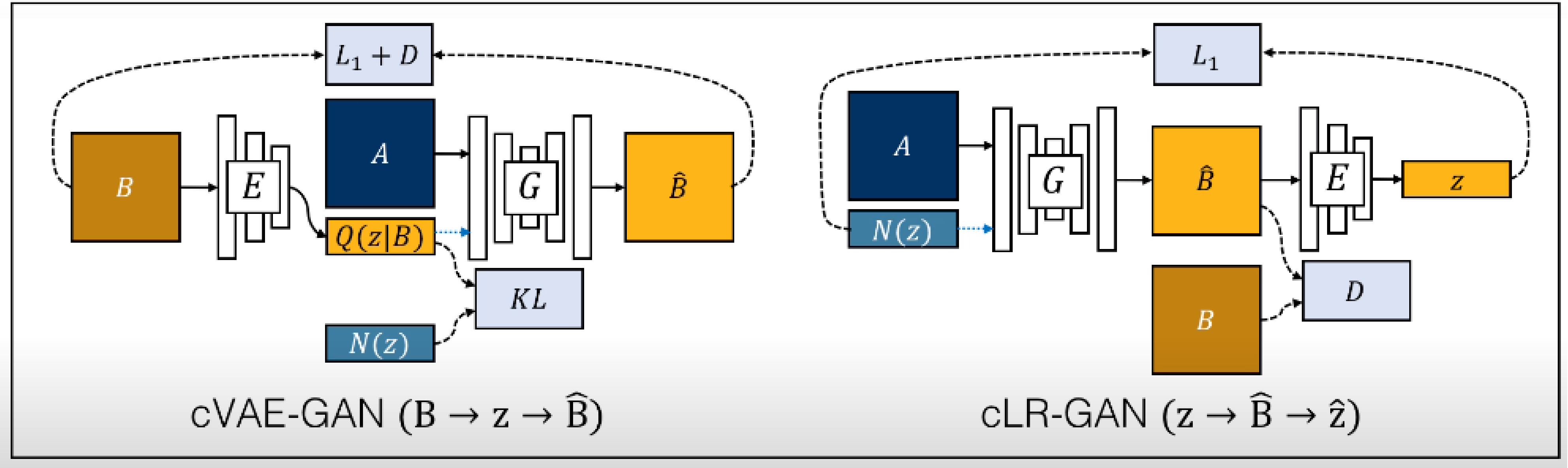
Lung Abnormalities

- Image-grader pairs drawn randomly
- Augmentations: random elastic deformation, rotation, shearing, scaling and a randomly
  translated crop
- U-Net:
  - 4 down- and up-sampling operations
  - Each block contains 3 Conv layers with 3X3 kernels → ReLU
  - Prior and posterior nets have same architecture as U-Net's encoder
- Training schedule:
  - 240 k iterations
  - Decaying learning rate: from $1e^{-4}$ to $1e^{-5}$ in 5 steps
  - Batch size: 32
  - Weight decay with weight $1e^{-5}$
  - Optimizer: Adam
- KL weight (β), latent space dimension (N):
  - For Image2Image β=10, N=3
  - For Probabilistic U-Net β=1, N=6

Cityscapes

- Augmentations: same + random color augmentations
- U-Net:
  - 5 down- and up-sampling operations
  - The rest is the same
- Training schedule:
  - 240 k iterations
  - Decaying learning rate: from $1e^{-4}$ to $1e^{-5}$ in 3 steps
  - Batch size: 16
  - Weight decay with weight $1e^{-5}$
  - Optimizer: Adam
- KL weight (β), latent space dimension (N):
  - For Image2Image β=1, N=3
  - For Probabilistic U-Net β=1, N=6

BicycleGAN



cVAE-GAN starts from a ground truth target image **B** and encode it into the latent space. The generator then attempts to map the input image **A** along with a sampled **z** back into the original image **B**.

cLR-GAN randomly samples a latent code from a known distribution, uses it to map **A** into the output $\widehat{\boldsymbol{B}}$, and then tries to reconstruct the latent code from the output.

BicycleGAN method combines constraints in both directions