Retained Singular Values in Probabilistic Image Segmentation with Normalizing Flows and Optimal Transport

Anonymous authors
Paper under double-blind review

Abstract

Latent probabilistic models are a popular choice for quantifying aleatoric uncertainty in image segmentation tasks. Nevertheless, we find that the singular values of the model distributions can vanish and result in a poor latent space. The retainment of latent singular values has been successful in state-of-the-art deterministic self-supervised models by optimizing embeddings on a projected space. In this work, we extend this approach to the probabilistic setting by introducing the Conditional Sinkhorn Auto-encoder (cSAE). It is shown that with Normalizing Flows and Optimal Transport theory, we can project the latent space and improve the learned embeddings of supervised conditional probabilistic segmentation models. This is because the singular values of the learned Normal densities are better retained, thereby improving the ability to accurately quantify the data uncertainty.

1 Introduction

A good understanding of the input data is important for developing successful deep learning models. Data is almost always obtained with inherent ambiguity, which is commonly referred to as the *aleatoric uncertainty*. Ignorance of this uncertainty can result in an incorrect interpretation of model predictions. Probabilistic segmentation enabls learning of the aleatoric uncertainty in the data. This work introduces the problem of vanishing singular values in probabilistic image segmentation and proposes an improved model, i.e. the Conditional Sinkhorn Autoencoder (cSAE), with an implementation that enhances segmentation performance. We show that by using Normalizing Flows (NFs) and Optimal Transport (OT), it is possible to significantly improve the encapsulation of the aleatoric uncertainty in the data.

In the context of vision, ambiguity in images can be due to occlusions, shadows, sensor noise, insufficient resolution or other equivocal information. Kohl et al. (2018) proposed to capture the ambiguity in image segmentation problems through the use of a conditional latent variable density model. Their work, the Probabilistic U-Net (PU-Net), learns latent embeddings of the ambiguity as amortized axis-aligned Normal densities. At test time, the learned ambiguity can be expressed by sampling from them. Although this model has been successful in modeling the aleatoric uncertainty, such as in the work of Valiuddin et al. (2021); Hu et al. (2019), we have found that it does not capture the true ambiguity that exists in the data with sufficient accuracy. We consider that the limiting factor of the PU-Net is in the latent space. In particular, the Normal densities have singular values that vanish to an extent during training. Unfortunately, not much work has been presented on the understanding of the learned densities. Recent work by Selvan et al. (2020); Valiuddin et al. (2021) successfully proposed augmenting the posterior of the PU-Net with an NF. Nevertheless, interpretation and explanation of its performance gains remain unaddressed. In this work, we suggest that the NFs aid in a better latent space representation of the ambiguity in the image. We highlight the resemblance of the PU-Net with joint embedding methods (JEMs) in self-supervised learning (SSL). Recently, in the context of SSL, it has been reported by Hua et al. (2021); Tian et al. (2021) that the models can suffer from partial dimensional collapse caused by vanishing singular values. State-of-the-art deterministic SSL methods such as that of Chen et al. (2020), make use of a so-called *projector* that improves the behaviour of the embeddings by optimizing on an auxiliary space. We state that the NFs augmented to the model posterior serve a similar purpose as the projector head in SSL. Nevertheless, the works of Selvan et al. (2020); Valiuddin et al. (2021) do not augment the prior density with an NF, which is in contrast

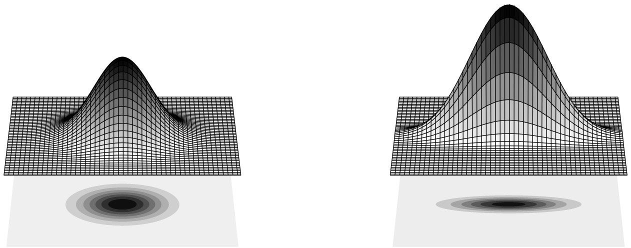

Figure 1: Normal densities with retained singular values (left) and a vanished singular value (right).

to JEMs in SSL, where both networks use projector heads. Given this, we extend upon the PU-Net by projecting both the prior and posterior distributions with NFs, offering a performance increase compared to the earlier methods. In particular, the singular values of the axis-aligned Normal densities are better retained, resulting in more accurate and expressive samples.

We observe that vanishing singular values, a form of dimensional collapse, can occur in the PU-Net. Recent work has shown that projecting the latent space can improve the embeddings and reduce partial dimensional collapse. To achieve this in the probabilistic setting, we combine the theory of OT with NFs to optimize on the projected latent space. This results in better latent Normal densities with retained singular values. Our contributions are as follows:

- The theory of NFs and OT is provided and used to propose the cSAE framework in Section 3.
- The development of the cSAE and its implementation is discussed in Section 4.
- Strong performance gains on competitive baselines on multi-annotated medical image segmentation datasets are presented, in terms of encapsulation of the aleatoric uncertainty and retainment of the singular values, in Section 5.

2 Related work

2.1 Dimensional collapse in self-supervised learning

Mode collapse is a problem that exists in many branches of machine learning. Hua et al. (2021) show that this can also happen partially, i.e. along particular dimensions, in SSL due to vanishing singular values. Jing et al. (2021) have investigated the effect of the commonly used multi-layered perceptron head (also referred to as a *projector*) in JEMs (Chen et al. (2020); Bachman et al. (2019); Wu et al. (2018)). Empirically, it is found that a simple projector improves the representation of the preceding layer. Furthermore, they argue that the projector reduces the degree of dimensional collapse caused by implicit regularization (Ji & Telgarsky (2018); Gunasekar et al. (2017); Arora et al. (2019)) due to over-parameterization (Saxe et al. (2019); Neyshabur et al. (2018); Soudry et al. (2018); Barrett & Dherin (2020)) and strong data augmentation. Chen et al. (2020) have motivated the importance of projection heads by pointing to the induced information loss caused by the training objective. Our work shows that vanishing singular values, a form of dimensional collapse, also occurs in conditional probabilistic latent variable models. Similar to SSL, we show that projecting the latent distributions to an auxiliary embedding retains the singular values and therefore improves the embeddings.

2.2 Probabilistic segmentation with Normalizing Flows

Kohl et al. (2018) introduced the PU-Net for image segmentation by combining the conditional Variational Autoencoder (cVAE) of Sohn et al. (2015) and the U-Net of Ronneberger et al. (2015). This enabled the model to provide multiple plausible segmentation hypotheses and, as a result, capture the ambiguity in the data.

With the increasing interest in methods for quantifying uncertainty, the PU-Net gained received significant attention and various improvements (Hu et al. (2019); Selvan et al. (2020); Valiuddin et al. (2021)). Hu et al. (2019) use the inter-observer variability as a target in the training objective. Selvan et al. (2020) and Valiuddin et al. (2021) use NFs to augment the posterior network, originally introduced by Rezende Rezende & Mohamed (2015) in the non-conditional setting. Valiuddin et al. (2021) hypothesize that the improvement in performance is due to the non-Gaussian nature of the posterior density enabled by the NF augmentation. While this can certainly be true, we argue in this work that a major contribution to the increased performance can be accredited to the regularizing effect of the augmented projection space.

2.3 Generative modeling with Optimal Transport

Bousquet et al. (2017) proposed building a generative model from the perspective of minimizing the Wasserstein OT plan between the model and target distribution. Patrini et al. (2020) have extended the analysis of Bousquet et al. (2017) and have introduced a training objective that uses Sinkhorn iterations (Sinkhorn & Knopp (1967); Cuturi (2013b); Feydy et al. (2019)) to approximate the Wasserstein distance. We also exploit the Sinkhorn iterations for similar purposes. A detailed explanation can be found in Section 3.2.

3 Theory

We use calligraphic letters ($\mathcal{X}$) to denote sets, capital letters (X) for random variables and lower case letters x for specific values. We denote marginals as P_{X}, probability distributions as $P(\mathrm{X})$ and densities as $p(x)$. We specifically distinguish vectors and matrices with boldface characters. The fundamentals of Normalizing Flows are explained in Section 3.1. The Optimal Transport problem is formulated in Section 3.2. Furthermore, it shown that conditional latent variable architectures are suitable to be used for OT problems and the choice of the Wasserstein distance as a divergence measure is justified in Section 3.2.1. Finally, an approximation to the Wasserstein distance, the Sinkhorn Algorithm, is discussed in Section 3.2.2.

3.1 Normalizing Flows

Normalizing Flows are a sequence of bijective transformations, typically starting from a complex distribution, transforming into a Normal distribution. The log-likelihood $\log p(\mathbf{x})$ of a sample $\mathbf{x} \in \mathbb{R}^{D \times D}$ from a Normal distribution subject to the NF with transformation $f_i : \mathbb{R} \mapsto \mathbb{R}$ is

$$\log p(\mathbf{x}) = \log p_0\left(\mathbf{z}_0\right) - \sum_{i=1}^{K} \log\left(\left|\det \frac{df_i}{d\mathbf{z}_{i-1}}\right|\right), \tag{1}$$

where the latent sample $\mathbf{z}_i$ is from the i-th transformation in the K-step NF and p_0 the base Normal probability distribution. A planar flow, introduced by Rezende & Mohamed (2015), is a specific type of bijection that possesses the property to expand and contract distributions along a specific direction with the following transformation

$$f(\mathbf{x}) = \mathbf{x} + \mathbf{u}h(\mathbf{w}^T\mathbf{x} + b), \tag{2}$$

with h being any smooth element-wise non-linear function with parameters $\mathbf{w} \in \mathbb{R}^D$, $\mathbf{u} \in \mathbb{R}^D$ and $b \in \mathbb{R}$.

3.2 Optimal Transport

Let $\mathcal{Y}$ and $\hat{\mathcal{Y}}$ be the two separable metric spaces. We adopt the Monge-Kantorovich formulation of Villani (2008) for the OT problem, by specifying

$$W_c^*(\mu, \nu) := \inf\left\{\int_{\mathcal{Y}\times\hat{\mathcal{Y}}} c(\mathbf{y}, \hat{\mathbf{y}})\, d\gamma(\mathbf{y}, \hat{\mathbf{y}}) \middle| \gamma \in \Gamma(\mu, \nu)\right\}, \tag{3}$$

where $\Gamma(\mu, \nu)$ denotes the tight collection of all probability measures on $\mathcal{Y} \times \hat{\mathcal{Y}}$ with marginals μ and ν, respectively, coupling γ, and $c(\mathbf{y}, \hat{\mathbf{y}})$: $\mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}_+$ being any lower semi-continuous measurable cost function. The usual context of this formulation is in finding the lowest cost of moving samples from the probability measures in $\mathcal{Y}$ to the measures in $\hat{\mathcal{Y}}$. In the case of probabilistic segmentation, the aim is to learn the ground-truth distribution P_{Y} by matching it with the prediction distribution $P_{\hat{\mathrm{Y}}}$.

3.2.1 Latent variable optimization

Let $(\mathrm{X}, \mathrm{Y}) \sim P_{\mathrm{X},\mathrm{Y}}$ be an input-target pair taking values in $\mathbb{R}^{D\times D} \times \mathbb{R}^{D\times D}$, and let P_{X} and P_{Y} be the marginal of X and Y, respectively. We define P_Z on $\mathcal{P}$ as a latent distribution of a lower-dimensional representation Z of X, taking values in $\mathbb{R}^d$ with a well-defined density $p_{\mathrm{Z}}(\mathbf{z}) = \mathbb{E}_{\mathrm{X}\sim P_{\mathrm{X}}}[p_{\mathrm{Z}|\mathrm{X}}(\mathbf{z}|\mathrm{X})]$. The conditional latent variable model with the transformation $\mathrm{G} : \mathcal{Z} \times \mathcal{X} \to \hat{\mathcal{Y}}$ is defined as

$$p_{\hat{\mathrm{Y}}|\mathrm{X}}(\hat{\mathbf{y}}|\mathbf{x}) := \int_{\mathcal{Z}} p_{\mathrm{G}}(\hat{\mathbf{y}}|\mathbf{z}, \mathbf{x}) p_{\mathrm{Z}|\mathrm{X}}(\mathbf{z}|\mathbf{x}) dz, \qquad \forall\, \mathbf{x} \in \mathcal{X}, \hat{\mathbf{y}} \in \hat{\mathcal{Y}}, \tag{4}$$

assuming all involved densities are properly defined. Furthermore, $\hat{\mathrm{Y}} \sim P_{\hat{\mathrm{Y}}}$ is the prediction with a density given by $p_{\hat{\mathrm{Y}}}(\hat{\mathbf{y}}) = \mathbb{E}_{\mathrm{X}\sim P_{\mathrm{X}}}[p_{\hat{\mathrm{Y}}|\mathrm{X}}(\hat{\mathbf{y}}|\mathrm{X})]$. Given this, Bousquet et al. (2017) state the upper bound

$$W_c^*(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}}) \leq W_c(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}}) := \inf_{P \in \mathcal{P}_{\mathrm{Y},\hat{\mathrm{Y}}}} \mathbb{E}_{(\mathrm{Y},\hat{\mathrm{Y}})\sim P}[c(\mathrm{Y}, \hat{\mathrm{Y}}))]. \tag{5}$$

The expressions in Equation (5) are equal when limiting P_{G} to Dirac measures (i.e. G is a deterministic map) and in that case, $W_c(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}})$ is a valid surrogate objective. We consider the family of encoding distributions Q_{Z} defined on $\mathcal{Q}$ with well-defined density $q_{\mathrm{Z}}(\mathbf{z}) = \mathbb{E}_{(\mathrm{X},\mathrm{Y})\sim P_{\mathrm{X},\mathrm{Y}}}[q_{\mathrm{Z}|\mathrm{X},\mathrm{Y}}(\mathbf{z}|\mathrm{X}, \mathrm{Y})]$. We can rephrase the first theorem in the work of Bousquet et al. (2017) as follows.

Theorem 1 *Let* $\mathrm{G} : \mathcal{Z} \times \mathcal{X} \to \hat{\mathcal{Y}}$ *be purely deterministic, then*

$$W_c^*(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}}) = W_c(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}}) = \inf_{Q : Q_{\mathrm{Z}} = P_{\mathrm{Z}}} \mathbb{E}_{\mathrm{X},\mathrm{Y}\sim P_{\mathrm{X},\mathrm{Y}}} \mathbb{E}_{\mathrm{Z}\sim Q}[c(\mathrm{Y}, \mathrm{G}(\mathrm{Z}, \mathrm{X}))]. \tag{6}$$

Notice that the generator is dependent on both the latent variable Z as well as the input image X. As proposed by Bousquet et al. (2017), the equality condition on Q_{Z} and P_{Z} can be relaxed by constraining the latent densities as stated in the second theorem below.

Theorem 2 *Given a convex penalty* $F : \mathcal{Q} \times \mathcal{P} \to \mathbb{R}_+$ *such that* $F(Q_{\mathrm{Z}}, P_{\mathrm{Z}}) = 0$ *if and only if* $Q_{\mathrm{Z}} = P_{\mathrm{Z}}$ *and* $\lambda > 0$*, then*

$$W_c^{\lambda}(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}}) = \inf_{Q \in \mathcal{Q}} \mathbb{E}_{\mathrm{X},\mathrm{Y}\sim P_{\mathrm{X},\mathrm{Y}}} \mathbb{E}_{\mathrm{Z}\sim Q}[c(\mathrm{Y}, \mathrm{G}(\mathrm{Z}, \mathrm{X}))] + \lambda \cdot F(Q_{\mathrm{Z}}, P_{\mathrm{Z}}) \leq W_c(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}}), \tag{7}$$

where the left-hand side approaches the upper bound with increasing λ*.*

Hence, $W_c^{\lambda}(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}})$ is a lower bound to $W_c(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}})$ and minimization of the former does not guarantee minimization of the latter objective. Nevertheless, Tolstikhin et al. (2017) show that the objective improves upon the VAE with specific choices for F. Patrini et al. (2020) use in the penalty F the $\tilde{p}$-Wasserstein distance, which is defined by

$$W_p(Q_{\mathrm{Z}}, P_{\mathrm{Z}}) := \inf \left\{ \left(\int_{\mathcal{Q}\times\mathcal{P}} d(\mathbf{q}, \mathbf{p})^p \, d\gamma(\mathbf{q}, \mathbf{p}) \right)^{\frac{1}{p}} \middle| \gamma \in \Gamma(Q_{\mathrm{Z}}, P_{\mathrm{Z}}) \right\}. \tag{8}$$

Here, the cost function is a distance metric d. According to Patrini et al. (2020), this choice of F in Equation (2) makes $W_c^{\lambda}(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}})$ an upper bound on $W_p(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}})$. Thus, we rephrase Theorem 2.1 by Patrini et al. (2020) as follows.

Theorem 3 *Let* $p \geq 1$ *and* $\mathrm{G} : \mathcal{Z} \times \mathcal{X} \to \hat{\mathcal{Y}}$ *be any deterministic function that is* γ*-Lipschitz, then we obtain the equality*

$$W_p(P_{\mathrm{Y}}, P_{\hat{\mathrm{Y}}}) = \inf_{Q \in \mathcal{Q}} \sqrt[p]{\mathbb{E}_{\mathrm{X},\mathrm{Y}\sim P_{\mathrm{X},\mathrm{Y}}} \mathbb{E}_{\mathrm{Z}\sim Q}[c(\mathrm{Y}, \mathrm{G}(\mathrm{Z}, \mathrm{X}))^p]} + \gamma \cdot W_p(Q_{\mathrm{Z}}, P_{\mathrm{Z}}), \tag{9}$$

where $\mathcal{Q}$ *is any class of probabilistic encoders that at least contains a class of universal approximators.*

This objective is resembles that of a VAE, which is similarly composed of a reconstruction and divergence term. This theorem justifies the use of the Wasserstein distance as a divergence measure in our conditional probabilistic segmentation model. Note that in contrast to the VAE and previous work in generative modeling with OT, the prior density P_{Z} is learned and conditioned on X.

3.2.2 Sinkhorn Algorithm as a Wasserstein approximation

In practice, the intricacies of the Wasserstein distance complicates its calculation. It has been proposed by Wilson (1968) to introduce entropic regularization to the OT problem. This is achieved by using the entropy of the coupling matrix as a regularizing function, which is specified by

$$\tilde{S}_\epsilon(\mu, \nu) := \inf \left\{ \int_{\mathcal{Y}\times\hat{\mathcal{Y}}} \left(d(\mathbf{y}, \hat{\mathbf{y}}) + \epsilon \log \frac{\gamma(\mathbf{y}, \hat{\mathbf{y}})}{d\mu d\nu} \right) d\gamma(\mathbf{y}, \hat{\mathbf{y}}) \middle| \gamma \in \Gamma(\mu, \nu) \right\}, \tag{10}$$

where $\epsilon \in \mathbb{R}_+$. As mentioned by Cuturi (2013a), the entropy term can be expanded to $\log(\gamma(\mathbf{y}, \hat{\mathbf{y}})) - \log(d\mu) - \log(d\nu)$. In this way, this formulation can be understood as constraining the joint probability to have *sufficient entropy* or contain small enough *mutual information* with respect to $d\mu$ and $d\nu$. This entropic regularization allows optimization over a Lagrangian dual for faster computation with the iterative Sinkhorn matrix scaling algorithm. The entropic bias is removed from the OT problem to obtain the Sinkhorn Divergence, specified as

$$S_\epsilon(\mu, \nu) = \tilde{S}_\epsilon(\mu, \nu) - \frac{1}{2} \left(\tilde{S}_\epsilon(\mu, \mu) + \tilde{S}_\epsilon(\nu, \nu) \right). \tag{11}$$

The Sinkhorn divergence interpolates between W_p ($\epsilon \to 0$) with $\mathcal{O}(\epsilon \log(\frac{1}{\epsilon}))$ deviation and Maximum Mean Discrepancy ($\epsilon \to \infty$), which favours dimension-independent sample complexity (Genevay et al. (2019)). A viable option is to approximate the Sinkhorn Divergence via sampling with weights $\alpha, \beta \in \mathbb{R}_+$. The regularized Sinkhorn algorithm performance lacks in lower temperature settings. To alleviate this limitation, as well as to increase efficiency, Kosowsky & Yuille (1994) introduce *ϵ-scaling* or *simulated annealing* to the Sinkhorn algorithm.

4 Methods

In the context of density modeling, several criteria need to be satisfied in order to employ latent space projection and to use the preceding embedding layer after training. Firstly, we need to constrain the projected latent spaces during training. For (un-)normalized Gaussians, the KL-divergence can be used. However, when the two underlying distributions are unknown (i.e. only a set of samples are observed), we can use the Sinkhorn Divergence for an approximate Wasserstein distance and employ this as a latent space constraint (Theorem 3). To achieve this, the sample likelihood needs to remain known after the projecting function. NFs are a suitable transformation due to their bijective property that allows keeping track of the sample probabilities with Equation (1). Finally, the preceding latent space needs to be sufficiently close to the projected latent space, such that samples result in accurate segmentation reconstructions.

In the cSAE, the projections originate from the base Normal distributions which can theoretically be unconstrained. Nevertheless, we empirically found that this leads to overfitting. In this case, the preceding embedding layers are far away from the projected space and serve as poor latent representations. Also, the successes with projection heads (see Section 2) are with JEMs, where both networks have identical weights. In the cSAE, sharing weights is not desired because the two networks have different tasks. Namely, the posterior network aims to learn the mutual information between the input and ground-truth images. This is achieved by constraining its density with that of the prior network, which is only conditioned on the input image and extracts from it relevant semantic information. As a solution to overfitting, we additionally penalize the KL-divergence between the base Normal distributions. This has several merits. Firstly, this constraint forces the posterior network to learn the mutual information between the input and ground-truth images in the Normal densities. Without the constraint, a significant part of the learning is attempted in the augmented NFs, which is not possible due to their simple nature. Furthermore, this design allows for the NFs to eventually converge the densities to be identical. In other words, initially, during training the projected space is far from the base Normal densities. As training continues, the projected latent spaces coincide with the base Normal densities. This enables us to remove the augmented NFs and sample accurate predictions from the preceding base Normal distributions. We hypothesize that this method results in better latent densities with retained singular values, due to the regularizing effect of the projected space.

Combining the aforementioned approaches, we can implement the conditional Sinkhorn Autoencoder, as presented in Figure 2. The posterior network Q, which attempts to learn the ambiguity in the image, is

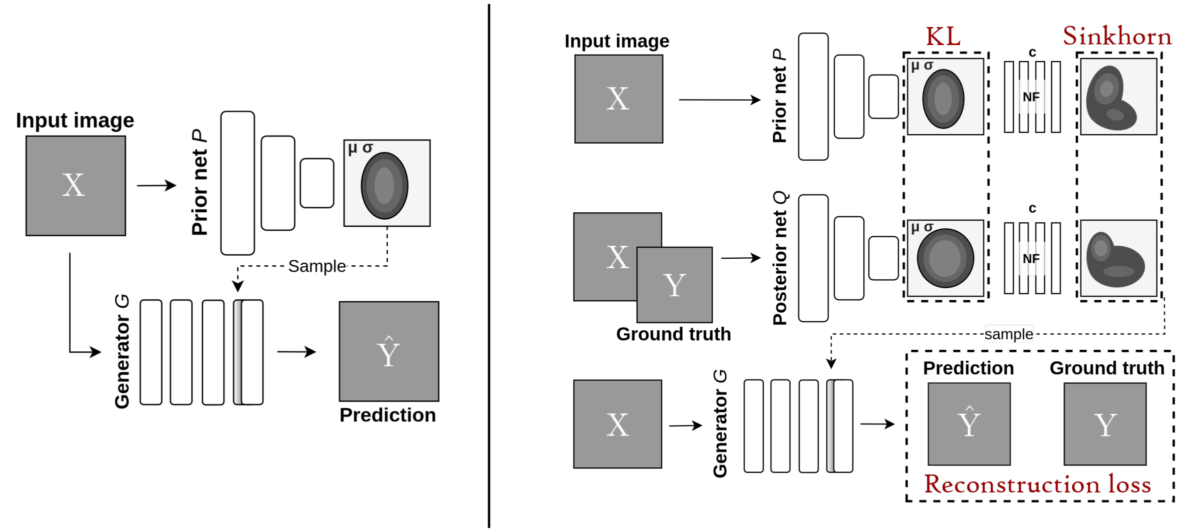

Figure 2: The Conditional Sinkhorn Autoencoder during testing (left) and training (right). Note how during training samples are taken from the projected, NF transformed latent space, and at test time this projection head is discarded.

conditioned on the input image X and ground-truth labels Y. The posterior network Q also provides the amortized parameters $\mathbf{c}$ for the NFs. During training, samples are drawn from the posterior and combined with the features from the generator (in our case the U-Net of Ronneberger et al. (2015)), to construct a segmentation. The optimization can be viewed from the perspective of a OT problem and we can use this to formulate a suitable constraint on the NF-subjected latent distributions, as discussed in Section 3.2.1. We state the training objective of the cSAE as

$$\mathcal{L} = -\mathbb{E}_{\mathrm{Z}\sim Q}\mathbb{E}_{\mathrm{X}\sim P_{\mathrm{X}}}[\log p_{\mathrm{G}}(Y|Z,X)] + \beta \cdot \mathrm{KL}\left(\mathcal{N}_Q^{\mu,\sigma}||\mathcal{N}_P^{\mu,\sigma}\right) + \gamma \cdot S_{c,\epsilon}(\hat{Q}_{\mathrm{Z}}, \hat{P}_{\mathrm{Z}}). \tag{12}$$

The Normal densities are denoted as $\mathcal{N}_Q^{\mu,\sigma}$ and $\mathcal{N}_P^{\mu,\sigma}$ with amortized parameters mean μ and standard deviation σ. The samples after the NF, conditioned on $\mathbf{c}$, are denoted as $\hat{Q}_{\mathrm{Z}}$ and $\hat{P}_{\mathrm{Z}}$. Furthermore, β and γ are tuneable hyperparameters introduced to regulate influence of the KL and Sinkhorn Divergence.

4.1 Dataset details

Several post-processed multi-annotated public datasets are used for training. Firstly, we use the LIDC-IDRI (Armato III et al. (2011)) dataset containing CT scans of lung nodules with up to four annotations. Secondly, we use the QUBIQ 2021 (Menze et al. (2021)) dataset consists of 7 multi-annotated subsets containing CT and MRI imaging of the prostate, brain, kidney and pancreas. We have found that the PU-Net already generalizes well on the prostate and kidney datasets, leaving little room for improvements. Therefore, we have experimented with the remaining datasets. More details on the datasets and training procedure can be found in Appendices A and B, respectively.

4.2 Performance evaluation

To demonstrate the effectiveness of our approach, we compare the cSAE with the PU-Net and its NF-augmented variant of Valiuddin et al. (2021). We keep the authors naming convention of and refer to the posterior augmented model as $2p$-planar, indicating a two-step planar NF on the posterior only. A two-layer non-linear projector head (i.e. a two-step NF) is used, since it has yielded substantial success (Jing et al. (2021); Valiuddin et al. (2021)).

We consider the inter-observer variability (i.e. disagreement between the annotators) of the multi-annotated datasets as the target for the learned aleatoric uncertainty. We use several evaluation metrics on our test sets to measure the similarity of the prediction and ground-truth images. Similarly to Kohl et al. (2019), Monte Carlo sampling (i.e. sampling from the latent prior for reconstruction) can be used to uniquely match all elements in the prediction with an element in the ground-truth set and is commonly referred to as Hungarian Matching. We refer to this method as the Empirical Wasserstein metric, because $\widehat{W}_k$, can be regarded as a Monte Carlo approximation of the original Wasserstein objective defined in Equation (3). We calculate the Empirical Wasserstein metric as follows

$$\widehat{W}_k(\mathcal{Y}, \hat{\mathcal{Y}}) = \min_{\gamma \in \Gamma(\mathcal{Y}, \hat{\mathcal{Y}})} \sum_i k(\gamma_i). \tag{13}$$

Here, a unique coupling γ is found between the ground-truth and prediction sets $\mathcal{Y}$ and $\hat{\mathcal{Y}}$. The optimal coupling minimizes the cost defined by the kernel k over all individual couples γ_i. To match the number of elements in both sets, we duplicate the number of ground-truth images to match the sample size of the predictions. When both prediction and ground-truth image are empty, we assign maximum score (i.e. minimal distance). For the Empirical Wasserstein metric, four different kernels are used. We apply the Intersection over Union (IoU), also known as the Jaccard index, defined by

$$\text{IoU} := \frac{TP}{TP + FP + FN}, \tag{14}$$

and the Dice score

$$\text{Dice} := \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \tag{15}$$

where TP, FP, TN, FN stand for the pixel-wise true positive, false positive, true negative and false negative, respectively, between the ground-truth $\mathbf{y}$ and prediction images $\hat{\mathbf{y}}$. Furthermore, we make use of the Hausdorff distance. Let $\mathcal{C}$ (resp. $\hat{\mathcal{C}}$) be the set containing all non-zero pixel-coordinate vectors of ground-truth $\mathbf{y}$ (resp. prediction $\hat{\mathbf{y}}$), then the Hausdorff distance can be defined as

$$d_{HD}(\mathbf{y}, \hat{\mathbf{y}}) := \max\{\max_{\mathbf{c} \in \mathcal{C}} d(\mathbf{c}, \hat{\mathcal{C}}), \max_{\hat{\mathbf{c}} \in \hat{\mathcal{C}}} d(\mathcal{C}, \hat{\mathbf{c}})\}, \tag{16}$$

where

$$d(\mathbf{c}, \hat{\mathcal{C}}) = \min_{\hat{\mathbf{c}} \in \hat{\mathcal{C}}} d(\mathbf{c}, \hat{\mathbf{c}}) \tag{17}$$

and

$$d(\mathcal{C}, \hat{\mathbf{c}}) = \min_{\mathbf{c} \in \mathcal{C}} d(\mathbf{c}, \hat{\mathbf{c}}), \tag{18}$$

and d denotes the Euclidean distance. Finally, we also evaluate the negative log-likelihood (cross-entropy) defined as

$$H_{ce}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i,j} y_{ij} \log(\hat{y}_{ij}), \tag{19}$$

Furthermore, to quantify the vanishing of the singular values we make use of the Gini index (Gini (1921); Lorenz (1905)). This metric sufficiently adheres to all requirements for an accurate sparsity measure, as is shown in Hurley & Rickard (2009). To calculate the Gini index, we obtain the singular values of the prior Normal densities. Because the Normal densities are axis-aligned, Singular Value Decomposition is not required and we can directly obtain the singular value vector $\boldsymbol{\sigma}$ from the trace of the covariance matrix. Then, singular values are normalized and ranked in an ascending order. We measure the sparsity S, of the L-dimensional vector as

$$S(\boldsymbol{\sigma}) = 1 - 2 \sum_{k=1}^{L} \frac{\sigma_k}{\|\boldsymbol{\sigma}\|_1} \left(\frac{L - k + \frac{1}{2}}{L} \right). \tag{20}$$

All evaluations are done on a test set, for each of the cross-validation folds, to rule out convenient splitting of the dataset. To qualitatively depict the model performance, we present the mean and standard deviation of the sample reconstructions. Even though it can benefit the quality of the predictions, we refrain from using temperature scaling during inference as this provides a more accurate representation of the learning capability of the models.

Dataset	Subset	Model	$\widehat{W}_k(\mathcal{Y}, \hat{\mathcal{Y}}) \downarrow$ 1 - IoU	 1 - Dice	 d_{HD}	 H_{ce}	**Gini**↓
LIDC-IDRI		PU-Net	0.308 ± 0.053	0.272 ± 0.053	21.426 ± 4.616	0.330 ± 0.068	0.457 ± 0.063
		$2p$-planar	0.305 ± 0.033	0.269 ± 0.033	21.044 ± 3.139	0.329 ± 0.054	0.289 ± 0.082
		cSAE	**0.291 ± 0.036**	**0.255 ± 0.036**	**19.896 ± 3.161**	**0.303 ± 0.052**	**0.192 ± 0.018**
QUBIQ 2021	pancreas	PU-Net	0.130 ± 0.077	0.092 ± 0.062	11.987 ± 5.989	0.753 ± 0.310	0.470 ± 0.026
		$2p$-planar	0.145 ± 0.051	0.100 ± 0.040	12.103 ± 3.792	0.692 ± 0.222	0.248 ± 0.058
		cSAE	**0.126 ± 0.070**	**0.089 ± 0.051**	**9.804 ± 4.854**	**0.550 ± 0.194**	**0.198 ± 0.062**
	pancreatic lesion	PU-Net	0.131 ± 0.054	0.081 ± 0.037	5.873 ± 2.034	0.566 ± 0.172	0.539 ± 0.057
		$2p$-planar	0.145 ± 0.044	0.097 ± 0.037	5.942 ± 2.884	0.544 ± 0.149	0.344 ± 0.063
		cSAE	**0.101 ± 0.057**	**0.062 ± 0.042**	**5.003 ± 2.216**	**0.513 ± 0.188**	**0.205 ± 0.042**
	brain growth	PU-Net	0.154 ± 0.005	0.084 ± 0.003	6.767 ± 0.164	3.061 ± 0.108	0.143 ± 0.066
		$2p$-planar	0.153 ± 0.009	0.084 ± 0.006	**6.528 ± 0.274**	3.177 ± 0.254	0.190 ± 0.071
		cSAE	**0.149 ± 0.003**	**0.081 ± 0.002**	6.753 ± 0.327	**2.952 ± 0.094**	**0.104 ± 0.046**

Table 1: Comparison of the proposed cSAE with the PU-Net and its posterior NF-augmented variant ($2p$-planar) on various datasets. We present the mean and standard deviation (due to threefold cross-validation) of the Empirical Wasserstein metric for multiple kernels and the Gini index of the singular values. All results are test set evaluations. The cSAE has better performance on almost all metrics.

5 Results and discussion

5.1 Quantitative evaluation

We present the test set evaluation of the three-fold cross-validation training in Table 1. Our main findings are twofold. Firstly, it is confirmed that by augmenting the PU-Net with NFs, the singular values are better retained. This can be understood from the Gini indices of the evaluations on the different datasets. In particular, it is observed that the cSAE has the lowest Gini index, indicating the least sparse singular values. The Gini indices of the $2p$-planar models generally lie in between that of the PU-Net and the cSAE. This shows that only projecting the posterior can already retain the singular values but the best performance is obtained when projecting both densities. Secondly, a correlation is found between the Gini index and the Empirical Wasserstein score on all datasets. This clearly indicates a relationship with the singular values and the model performance. Namely, that the retained singular values enable better sample accuracy according to the distribution of the ground-truth images.

The latent space behaviour during training time is also investigated. We depict the Gini indices of the prior density singular values as the model converges in Figure 4. Interestingly enough, it can be observed that in most instances, the PU-Net prior latent space has singular values vanishing (increased Gini index) early in training. On the other hand, it can also be observed that the singular values of the $2p$-planar and cSAE models always have a lower Gini index and are more volatile. We see a trend where more volatile Gini indices during training are also generally lower. We hypothesize that this volatility is due to the mechanism that retains the singular values during training time. We leave the details of this exact mechanism for future work. From the results it can be confirmed that the cSAE retains the singular values the most.

5.2 Qualitative evaluation

Qualitative samples of the various dataset are shown in Appendix C. It is clearly visible that the cSAE better encapsulates the aleatoric uncertainty. For the LIDC dataset, it is observed that the PU-Net considers ambiguity present at the center of the lesions, while the cSAE indicates most ambiguity almost exclusively around the edges. As a result of a better latent space, we observe stronger mean values in the areas where the annotators agree and stronger standard deviation values where the annotators disagree. For the QUBIQ Brain growth and Brain tumor subsets, the PU-Net underestimates the aleatoric uncertainty. In contrast, the Pancreas and Pancreatic lesion subsets result in overestimation by the PU-Net. The cSAE aligns the closest with the inter-observer variability and continuously the aleatoric uncertainty in the data.

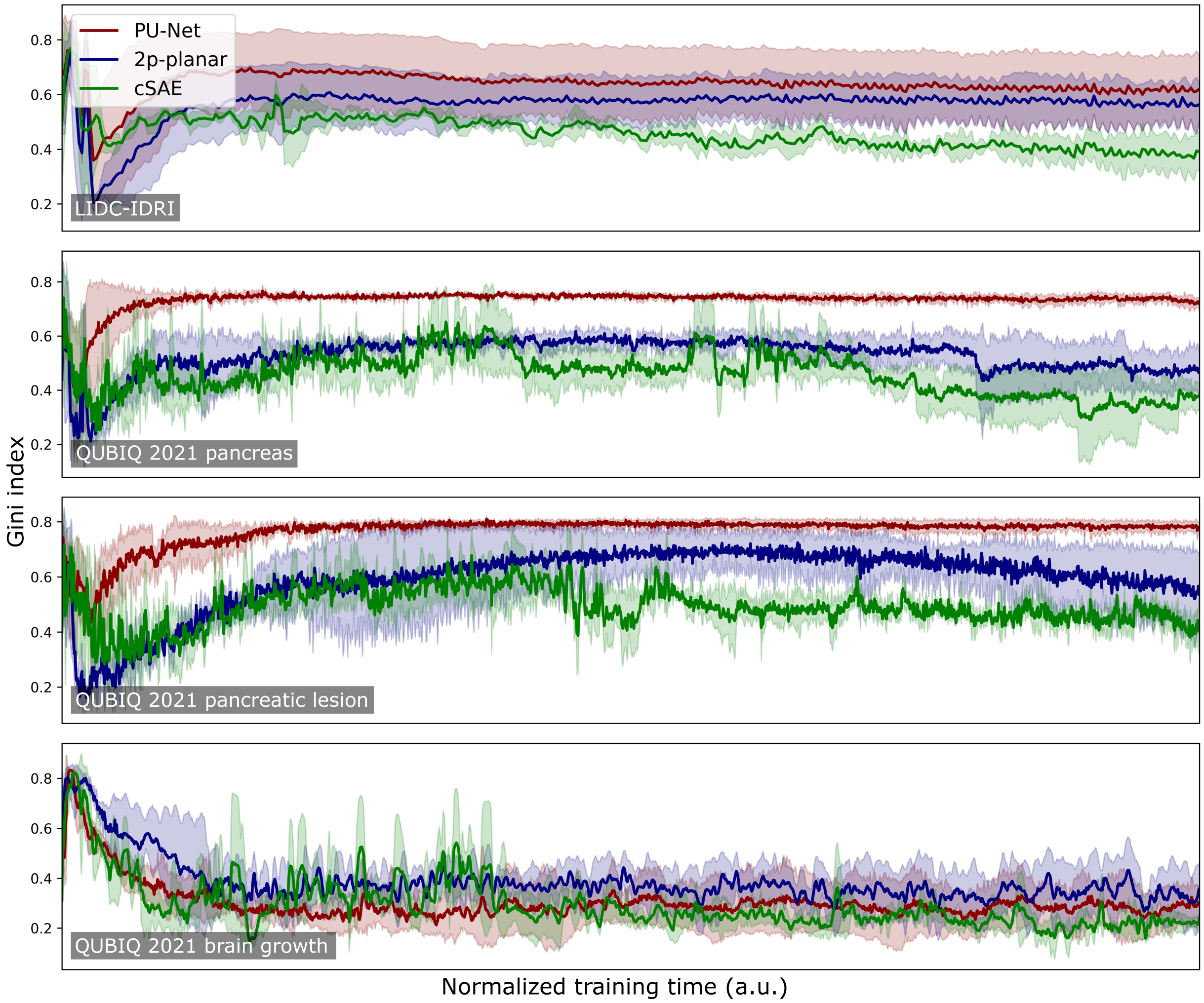

Figure 3: Gini indices of the prior singular values over normalized training time for the LIDC-IDRI and QUBIQ 2021 datasets. Lower is better. The cSAE has the lowest Gini index for all experiments, followed by the $2p$-planar and PU-Net. This shows that latent space projections retain the density singular values.

Dataset	Task	Model	$\widehat{W}_k(\mathcal{Y}, \hat{\mathcal{Y}}) \downarrow$				**Gini**↓
			1 - IoU	1 - Dice	Hausdorff	NLL	
QUBIQ 2021 brain tumor	task 1	PU-Net	0.098 ± 0.036	0.054 ± 0.021	9.049 ± 4.100	0.843 ± 0.259	0.224 ± 0.103
		2*p*-planar	0.086 ± 0.026	0.047 ± 0.015	**6.270 ± 3.787**	**0.737 ± 0.126**	**0.103 ± 0.013**
		cSAE	**0.073 ± 0.015**	**0.039 ± 0.008**	7.755 ± 3.470	0.892 ± 0.255	0.174 ± 0.091
	task 2	PU-Net	0.365 ± 0.098	0.291 ± 0.088	18.505 ± 5.574	1.048 ± 0.284	0.533 ± 0.158
		2*p*-planar	0.308 ± 0.092	0.236 ± 0.088	13.016 ± 7.982	0.835 ± 0.286	**0.399 ± 0.108**
		cSAE	**0.279 ± 0.043**	**0.206 ± 0.035**	**9.219 ± 2.286**	**0.778 ± 0.174**	0.402 ± 0.139
	task 3	PU-Net	0.126 ± 0.025	0.071 ± 0.015	2.449 ± 0.163	0.167 ± 0.010	0.229 ± 0.046
		2*p*-planar	0.111 ± 0.015	0.062 ± 0.008	2.228 ± 0.104	0.181 ± 0.019	**0.112 ± 0.047**
		cSAE	**0.105 ± 0.010**	**0.058 ± 0.006**	**2.131 ± 0.258**	**0.153 ± 0.005**	0.217 ± 0.129

Table 2: Comparison of the proposed cSAE with the PU-Net and its posterior NF-augmented variant (2*p*-planar) on the Brain tumor subset of the QUBIQ 2021 dataset. We present the mean and standard deviations (due to threefold cross-validation) of the Empirical Wasserstein metric for multiple kernels and the Gini indices of the singular values. All results are test set evaluations.

5.3 Outlier cases

The QUBIQ 2021 Brain tumor dataset deviates in terms of Gini index from the described trend (see Table 2). For that subset, the cSAE is significantly better by the Empirical Wasserstein metric while the Gini index is not the lowest of all three models. This can suggest that the retained singular values do not exclusively contribute to the increased performance, but play an important role in it. Also, this can indicate that for this dataset and with the particular chosen latent space dimensionality, the sparsity of the singular values is not the bottleneck of the performance. The difference in Gini indices across models over training time is very low. This, combined with the fact that the values vary substantially per fold challenges the statistical significance of the results and reliability of any conclusions therefrom. We highlight two possible reasons for the deviating behaviour. Firstly, it is important to note that the Brain tumor subset is significantly smaller (only around 20 images) which is likely too low to sufficiently generalize on the problem. Secondly, it is the only multi-modal MRI dataset used in the experiments, where the input data consisted of different acquisition modalities, concatenated channel-wise.

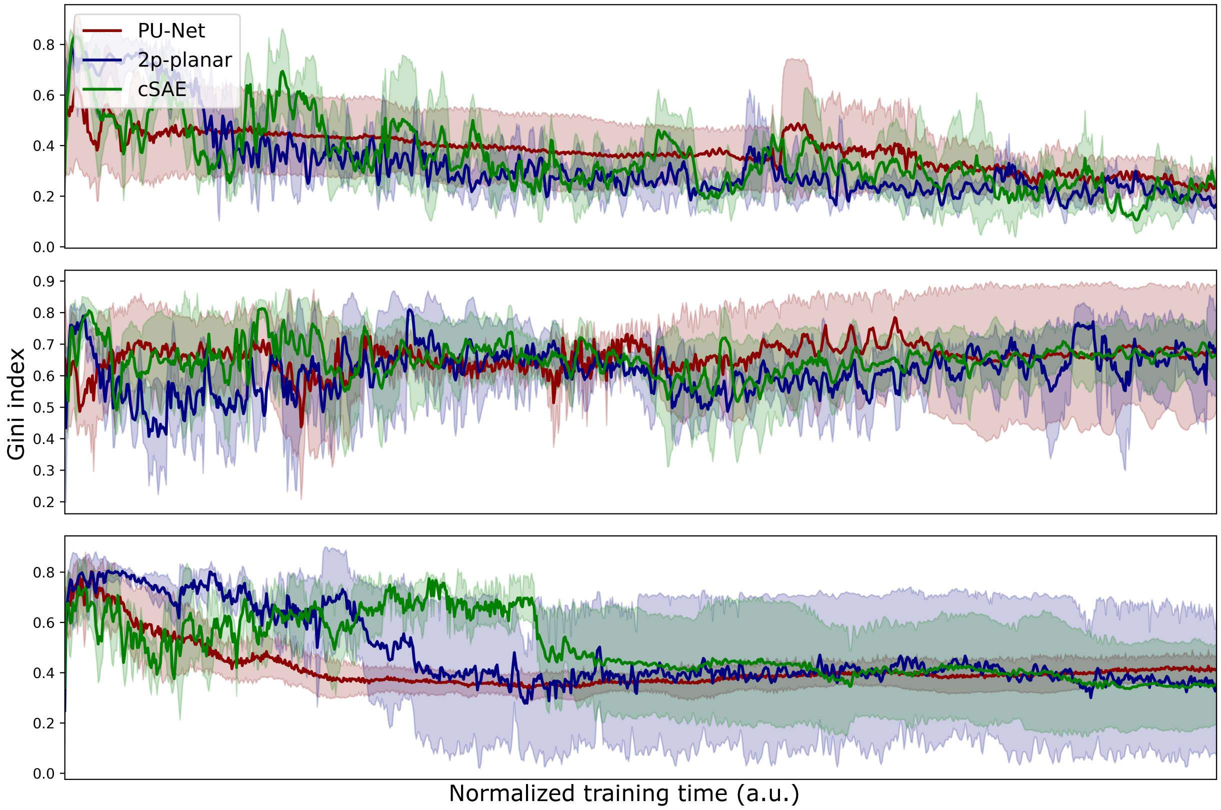

Figure 4: Gini indices of the prior singular values over normalized training time for the three tasks of the QUBIQ 2021 Brain tumor dataset. Lower is better. The various models do not portray a relationship with the Gini index, since their respective values are close and have high variance.

6 Conclusion

Quantifying uncertainty in image segmentation is quintessential for well-informed decision making. In this paper, we propose to use Normalizing Flows and the theory of Optimal Transport to improve the latent modeling of the ambiguity in the data. Our approach, the cSAE, prevents vanishing of the singular values of modeled axis-aligned Normal densities, which is observed in the former variants of the Probabilistic U-Net. In our proposed framework, the absence of vanishing singular values results in significant improvements of the aleatoric uncertainty quantification. Certainly, any model that aims to capture distributions in an axis-aligned Normal can suffer from this. Therefore, we propose that the projection with NFs and OT should be experimented with throughout other ambiguous modalities. Our future work will focus on gaining a deeper understanding on the exact mechanism that retains the singular values. Furthermore, we did not consider architectures that model densities at multiple resolutions. We are convinced that extending this work in the multi-resolution setting will similarly result in performance gains.

References

Samuel Armato III, Geoffrey Mclennan, Luc Bidaut, Michael McNitt-Gray, Charles Meyer, Anthony Reeves, Binsheng Zhao, Deni Aberle, Claudia Henschke, Eric Hoffman, Ella Kazerooni, Heber Macmahon, Edwin Beek, David Yankelevitz, Alberto Biancardi, Peyton Bland, Matthew Brown, Roger Engelmann, Gary Laderach, and Laurence Clarke. The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics*, 38:915–931, 01 2011. doi: 10.1118/1.3528204.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

David GT Barrett and Benoit Dherin. Implicit gradient regularization. *arXiv preprint arXiv:2009.11162*, 2020.

Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From optimal transport to generative modeling: the vegan cookbook, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013a. URL `https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf`.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013b.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1574–1583. PMLR, 2019.

Corrado Gini. Measurement of inequality of incomes. *The economic journal*, 31(121):124–126, 1921.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 137–145. Springer, 2019.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9598–9608, 2021.

Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *NeurIPS*, 2018.

Simon A. A. Kohl, Bernardino Romera-Paredes, Klaus H. Maier-Hein, Danilo Jimenez Rezende, S. M. Ali Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities, 2019.

Jeffrey J Kosowsky and Alan L Yuille. The invisible hand algorithm: Solving the assignment problem with statistical physics. *Neural networks*, 7(3):477–490, 1994.

M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905. doi: 10.1080/15225437.1905.10503443.

Bjoern Menze, Leo Joskowicz, Spyridon Bakas, Andras Jakab, Ender Konukoglu, Anton Becker, Amber Simpson, and Richard Do. Quantification of Uncertainties in Biomedical Image Quantification 2021, March 2021. URL `https://doi.org/10.5281/zenodo.4575204`.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Giorgio Patrini, Rianne van den Berg, Patrick Forre, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. In *Uncertainty in Artificial Intelligence*, pp. 733–743. PMLR, 2020.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

Raghavendra Selvan, Frederik Faye, Jon Middleton, and Akshay Pai. Uncertainty quantification in medical image segmentation with normalizing flows. In *MLMI@MICCAI*, 2020.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pp. 10268–10278. PMLR, 2021.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders, 2017. URL `https://arxiv.org/abs/1711.01558`.

M. M. Amaan Valiuddin, Christiaan G. A. Viviers, Ruud J. G. van Sloun, Peter H. N. de With, and Fons van der Sommen. Improving aleatoric uncertainty quantification in multi-annotated medical image segmentation with normalizing flows. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pp. 75–88, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87735-4.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. URL `https://books.google.nl/books?id=hV8o5R7_5tkC`.

A.G. Wilson. *The Use of Entropy Maximising Models in the Theory of Trip Distribution, Mode Split and Route Split*. Working papers // Centre for Environmental Studies. Centre for Environmental Studies, 1968. URL `https://books.google.nl/books?id=bn3wAAAAMAAJ`.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

A Dataset details

The LIDC-IDRI dataset is preprocessed from 1,018 thoracic CT scans into 15,096 128×128-pixel patches. Each image has up to four annotators. The pancreas subset of the QUBIQ 2021 dataset contains 76 cases with 2 annotations per slice. We process the data such that each slice has at least one non-zero value in one of the associated annotations. This results in a total of 869 128×128-pixel patches. The brain-tumor subset has four input channels where each channel corresponds to a different acquisition method (multimodal MRI) for the same subject. For the other subsets, i.e. pancreatic lesion and brain growth, we only normalize and resize to 128×128-dimensional patches. The information on the datasets are summarized in Table 3

Dataset	Datapoints	Dimensionality	Annotators	Input channels	Tasks
LIDC-IDRI	15,096	128×128	4	1	1
Pancreas	869	128×128	2	1	1
Pancreatic lesion	156	128×128	2	1	1
Brain growth	39	128×128	7	1	1
Brain tumor (all)	32	128×128	7	4	4

Table 3: Details of each applied dataset.

B Training details

Training details across datasets are kept as similar as possible to rule out biases towards favourable test-set evaluations. We have summarized the hyperparameter settings per dataset in Table 4. The models are trained with the Adam optimizer, weight decay and early stopping applied to the validation loss. The hyperparameters β and γ weigh the contribution of the KL and Sinkhorn divergence towards the loss function. The KL divergence of the L-dimensional densities are calculated analytically and the Sinkhorn iterations are based upon N_{OT} samples. All Monte-Carlo evaluations are done with N_{eval} samples. Tuning of the number of iterations for early stopping is based upon the validation loss during training. For the aforementioned reason, we have also kept the data augmentation consistent throughout the datasets and models (see Table 5 for details). Splitting of the different datasets are also performed in the same manner. For testing, the first 20% of the dataset is used. The remaining 80% is used for training with threefold cross-validation. We have chosen cross-entropy for the reconstruction loss. For the Sinkhorn Divergence, the `GeomLoss` Feydy et al. (2019) library has been used with $p = 2$ and $Diameter = 100$. The other parameters for the Sinkhorn Divergence are left to default. The gradients are clipped to have unitary norm. Training has been done on an 11GB NVIDIA RTX 2080TI. Our implementation of the cSAE with KL regularization will be made public.

Dataset	β	γ	Batch size	Learning rate	Weight decay	Patience	L	N_{OT}	N_{eval}
LIDC-IDRI	10	10	32	10^{-4}	$5e^{-5}$	40	6	16	16
Pancreas	10	1	32	10^{-4}	$5e^{-5}$	200	6	16	16
Pancreatic lesion	10	10	32	10^{-4}	$5e^{-5}$	500	6	16	15
Brain growth	10	10	32	10^{-4}	$5e^{-5}$	300	6	16	28
Brain tumor (all)	10	1	32	10^{-4}	$5e^{-5}$	300	6	16	15

Table 4: Training hyperparameters settings for the training of the different datasets.

Random augmentation	Min	Max
Rotation	-180	180
Translation	-0.1	0.1
Scaling	0.8	1.2
Shear (x and y)	-30	30
Brightness	0.8	1.2
Gamma	0.5	1.5

Table 5: Data augmentation settings for all datasets.

C Qualitative samples

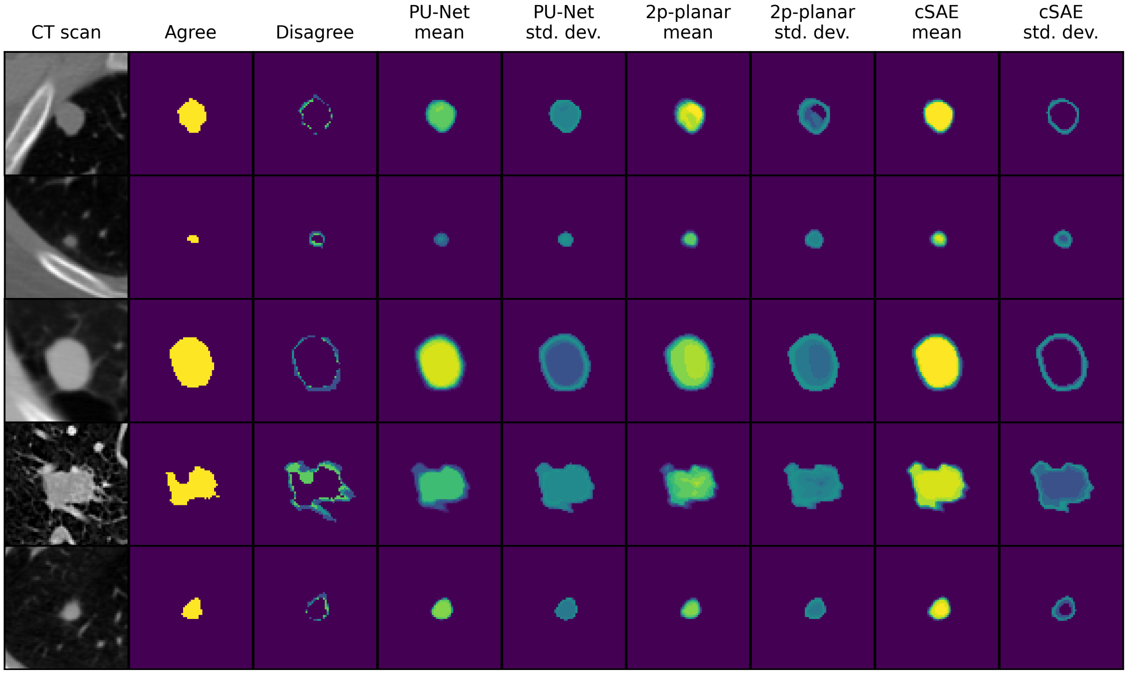

Figure 5: Test set samples of the LIDC IDRI dataset

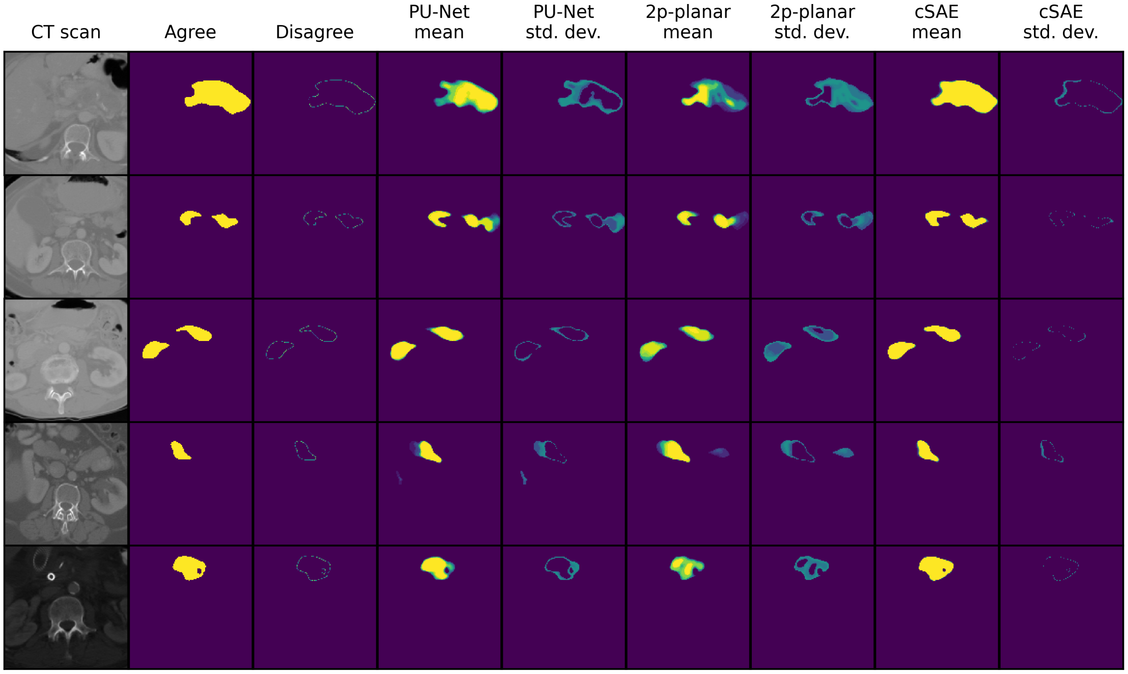

Figure 6: Test set samples of the QUBIQ 2021 pancreas subset

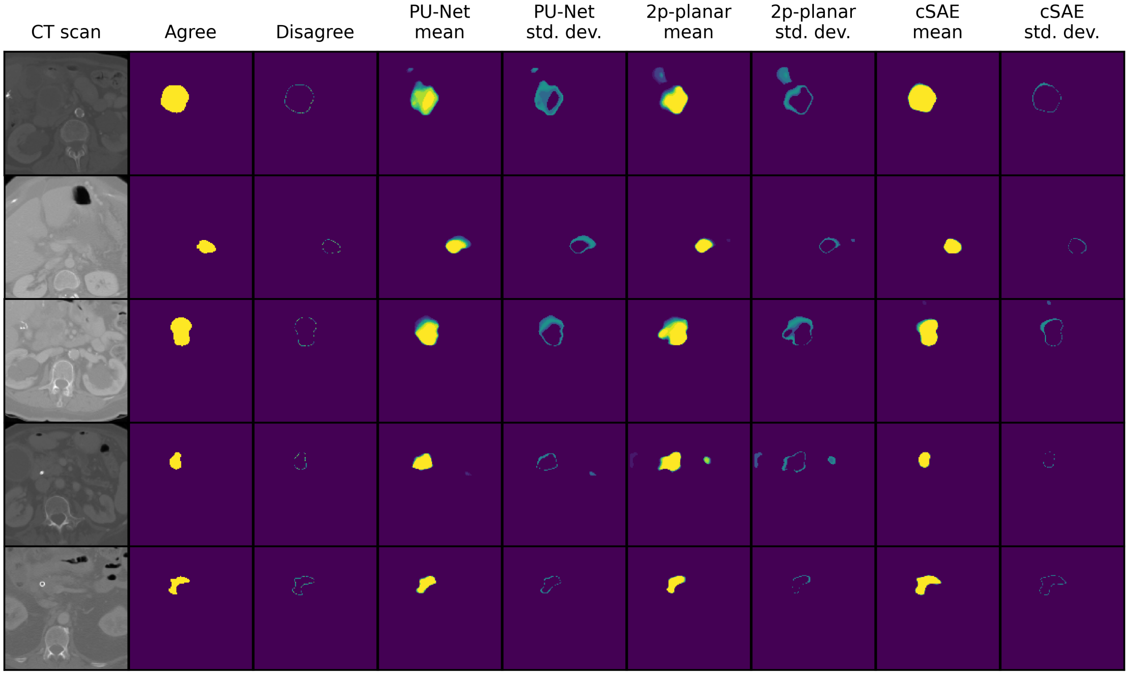

Figure 7: Test set samples of the QUBIQ 2021 pancreatic-lesion subset

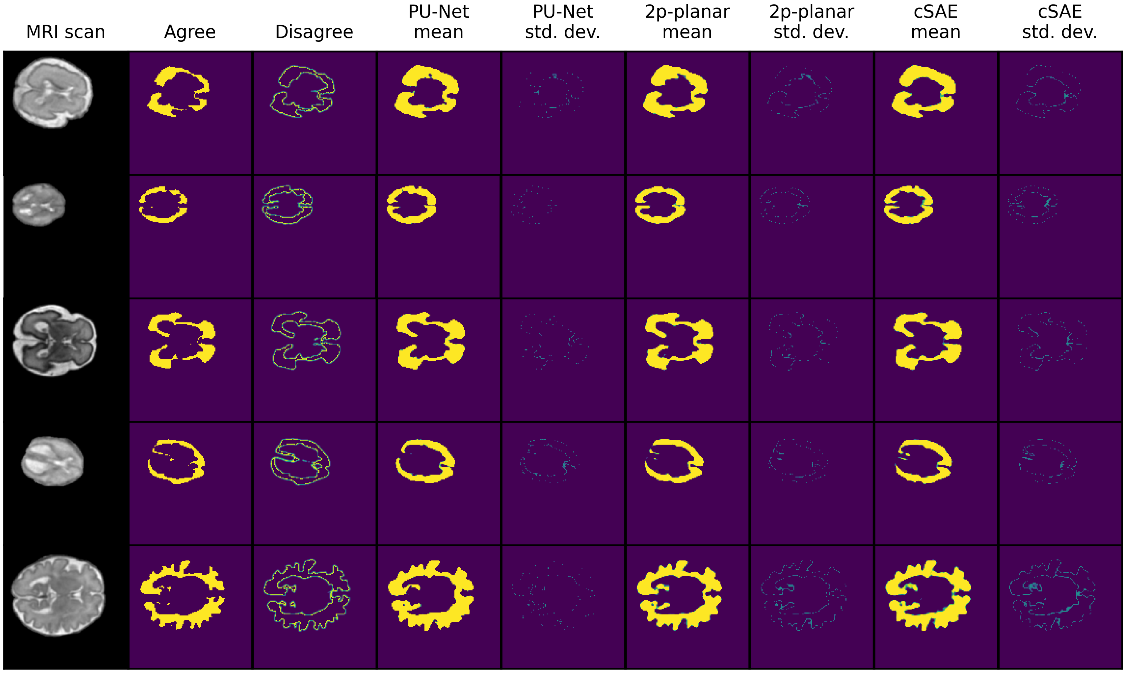

Figure 8: Test set samples of the QUBIQ 2021 brain-growth subset