Calibrated Adversarial Refinement for Stochastic Semantic Segmentation

Elias Kassapis[1,2,*]   Georgi Dikov[2]   Deepak K. Gupta[1]   Cedric Nugteren[2]

[1] Informatics Institute, University of Amsterdam, The Netherlands
[2] TomTom, Amsterdam, The Netherlands
[*] kassapiselias@hotmail.co.uk

Abstract

In semantic segmentation tasks, input images can often have more than one plausible interpretation, thus allowing for multiple valid labels. To capture such ambiguities, recent work has explored the use of probabilistic networks that can learn a distribution over predictions. However, these do not necessarily represent the empirical distribution accurately. In this work, we present a strategy for learning a calibrated predictive distribution over semantic maps, where the probability associated with each prediction reflects its ground truth correctness likelihood. To this end, we propose a novel two-stage, cascaded approach for calibrated adversarial refinement: (i) a standard segmentation network is trained with categorical cross entropy to predict a pixelwise probability distribution over semantic classes and (ii) an adversarially trained stochastic network is used to model the inter-pixel correlations to refine the output of the first network into coherent samples. Importantly, to calibrate the refinement network and prevent mode collapse, the expectation of the samples in the second stage is matched to the probabilities predicted in the first. We demonstrate the versatility and robustness of the approach by achieving state-of-the-art results on the multigrader LIDC dataset and on a modified Cityscapes dataset with injected ambiguities. In addition, we show that the core design can be adapted to other tasks requiring learning a calibrated predictive distribution by experimenting on a toy regression dataset. We provide an open source implementation of our method at https://github.com/EliasKassapis/CARSSS.

1. Introduction

Real-world datasets are often riddled with ambiguities, allowing for multiple valid solutions for a given input. These can emanate from an array of sources, such as sensor noise, occlusions, inconsistencies during manual data annotation, or an ambiguous label space [38]. Despite the fact that the empirical distribution can be multimodal, the majority of the
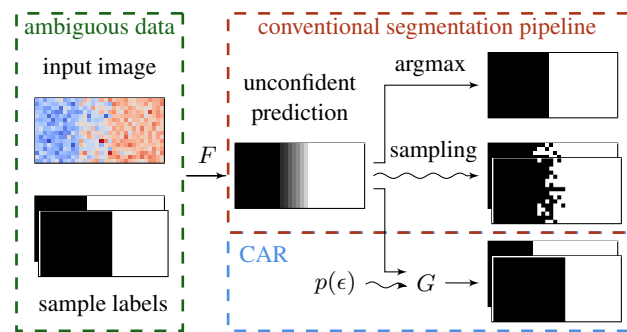

Figure 1: Conceptual diagram of stochastic semantic segmentation: blue and red pixels are separable by several different vertical boundaries, resulting in multiple valid labels. F is a network parametrising a factorised categorical likelihood, which captures the pixelwise data ambiguity. Extracting the mode via the argmax operation deterministically yields a single coherent prediction, while direct sampling gives multiple incoherent ones. Instead, calibrated adversarial refinement (CAR) uses a second stochastic adversarial network, G, which refines the output of F into diverse, coherent labels.

research encompassing semantic segmentation focuses on optimising models that assign only a single solution to each input image [47, 26, 53, 9, 10, 11, 7, 8], and are thus often incapable of capturing the entire empirical distribution.

These approaches typically model each pixel independently with a factorised categorical likelihood, and therefore do not consider inter-pixel correlations during sampling (see Fig. 13b in Appendix B.3). Further, since maximising the likelihood on noisy datasets leads to unconfident predictions in regions of label inconsistencies, direct sampling yields incoherent semantic maps. Alternatively, coherent predictions can be obtained by applying the argmax function, essentially extracting the mode of the likelihood. This, however, comes at the cost of limiting the model's representation capabilities to deterministic, one-to-one mappings between inputs and outputs (see Fig. 1).

Here we consider the problem of *stochastic semantic segmentation*: the task of semantically segmenting ambiguous images with an arbitrary number of valid labels, each with a distinct probability of occurrence. To this end, an ideal model should capture joint pixel dependencies and leverage uncertainty information to sample multiple coherent hypotheses. Further, it is important that the empirical occurrence frequency of each sampled segmentation variant reflects its ground truth correctness likelihood; that is, the predictive distribution should be calibrated [17, 34]. Such a system would be especially useful for semi-automatic safety-critical applications, e. g. medical imaging and map making, where it is crucial to identify ambiguous input and cross-examine all possible interpretations and their corresponding likelihoods before making an important decision [1, 42, 39].

In this work, we introduce calibrated adversarial refinement (CAR): a two-stage, cascaded framework for learning a calibrated, multimodal predictive distribution. In the first stage, we train a standard network with categorical cross entropy to estimate pixelwise class probabilities, as well as the associated aleatoric uncertainty estimates [28]. In the second, an adversarial network, capable of modelling the inter-pixel dependencies, is used to sample realistic, coherent predictions (see bottom of Fig. 1). The sample diversity is then calibrated relatively to the distribution predicted in the first stage, via an additional loss term. Our key contributions are:

- We propose a novel cascaded architecture for adversarial refinement that allows sampling of an arbitrary number of coherent segmentation maps.
- We introduce a novel loss term, called the *calibration loss*, that facilitates learning of calibrated stochastic mappings and mitigates mode collapse in conditional adversarial learning.
- Our model can be trained independently or used to augment any black-box semantic segmentation model.

2. Related work

Straightforward strategies towards learning multiple predictions include ensembling [35, 27] or using multiple prediction heads [49]. Even though these approaches can capture a diverse set of sampled predictions, they are limited to only a fixed number of samples. Alternatively, a probability distribution over the outputs can be induced by activating dropout during test time [13]. This method does offer useful uncertainty estimates over the pixel-space [45], however, it has been demonstrated [25, 31] that it introduces only minor stochasticity in the output and returns incoherent samples.

Bhattacharyya et al. (2018) [5] identify the maximum likelihood learning objective as the cause for this issue in dropout Bayesian neural networks [12]. They postulate that under cross entropy optimisation, all sampled models are forced to explain the entirety of the data, thereby converging to the mean solution. They propose to mitigate this

issue using variational inference and replacing cross entropy with an adversarial loss term parametrising a synthetic likelihood [48]. This renders the objective function conducive to multimodality but, unlike our method, requires the specification a weight prior and variational distribution family.

Kohl et al. (2018) [31] take an orthogonal approach in combining a U-Net [47] with a conditional variational autoencoder (cVAE) [30] to learn a distribution over semantic labels. Hu et al. (2019) [21] build on [31] by leveraging inter-grader variability as additional supervision. Even though this improves performance, a major limitation of this approach is the requirement of a priori knowledge of all the modes in the data distribution, often unavailable in real-world datasets. Alternatively, subsequent work in [32] and [4] improve the diversity of the samples of [31] by modelling the data on several scales of the image resolution. In more recent work, Monteiro et al. (2020) [44] take a different path, proposing a single network to parametrise a low-rank multivariate Gaussian distribution which models the inter-pixel and class dependencies in the logit space. This method does improve efficiency during inference, however, a low-rank parametrisation imposes a constraint on the sample complexity. In contrast to all these methods, ours uses an adversarial loss term which has been shown to elicit superior structural qualities than cross entropy [41, 14, 50].

In the more general domain of image-to-image translation, hybrid models have been proposed using adversarially trained cVAEs [58, 3] to learn a distribution over a latent code that encodes multimodality, allowing sampling of diverse yet coherent predictions. A common hurdle in such conditional generative adversarial networks (cGANs) is that simply incorporating a noise vector as an additional input often leads to mode collapse. This occurs due to the lack of regularisation between noise input and generator output, allowing the generator to learn to ignore the noise vector [25]. This is commonly resolved by using supplementary cycle-consistency losses [23, 37, 58, 3], as proposed by Zhu et al. (2017) [57], or with alternative regularisation losses on the generator [55]. Nonetheless, these do not address the challenge of calibrating the predictive distribution.

3. Method

3.1. Motivation

Semantic segmentation refers to the task of predicting a pixelwise class label $y \in \{1, \ldots, K\}^{H \times W}$ given an input image $x \in \mathbb{R}^{H \times W \times C}$. For a dataset of N image and label pairs, $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$, the empirical distribution $p_{\mathcal{D}}(y \mid x)$ can be explicitly modelled through a likelihood $q_\theta(y \mid x)$, parametrised by a softmax-activated convolutional neural network F with weights θ [47, 26]. One simple, yet effective way to learn the class probabilities is to express $y \in \{0, 1\}^{H \times W \times K}$ in a one-hot encoded representation

and set q_θ as a pixelwise factorised categorical distribution:

$$q_\theta(y \mid x) = \prod_i^H \prod_j^W \prod_k^K F_\theta(x)_{i,j,k}^{y_{i,j,k}}. \quad (1)$$

The parameters θ are then optimised by minimising the cross entropy between $p_\mathcal{D}$ and q_θ, defined as:

$$\mathcal{L}_{\text{ce}}(\mathcal{D}, \theta) = -\mathbb{E}_{p_\mathcal{D}(x,y)}[\log q_\theta(y \mid x)]. \quad (2)$$

When trained with Eq. (2), F_θ learns an approximation of $\mathbb{E}_{p_\mathcal{D}}[y \mid x]$ [6], thereby capturing the per-pixel class probabilities over the label that corresponds to a given input. At this point, the aleatoric uncertainty can be obtained by computing the entropy of the output of F_θ, $\mathbb{H}(F_\theta(x))$ [28].

As discussed in the introduction section and exemplified in Fig. 1, neither sampling from the likelihood in Eq. (1), nor extracting its mode are adequate solutions for stochastic semantic segmentation, where multiple valid predictions are sought. This issue can be partially addressed by adapting the framework of generative adversarial networks (GANs) [15] to the context of semantic segmentation, as proposed by [41]. Formally, this involves training a binary discriminator network D to optimally distinguish between ground truth and predictions, while concurrently training a conditional generative network G to maximise the probability that prediction samples $G(x)$ are perceived as real by D. Importantly, in contrast to explicit pixelwise likelihood maximisation, the adversarial setup learns an implicit sampler through G, capable of modelling the joint pixel configuration of the synthesised labels, and capturing both local and global consistencies present in the ground truth [41].

In practice, the generator loss is often complemented with the pixelwise loss from Eq. (2) to improve training stability and prediction quality [41, 14, 50]. However, we argue that the two objective functions are not well aligned in the presence of noisy data. While categorical cross entropy optimises for a single, mode averaging solution for each input x, thus encouraging high entropy in $q_\theta(y \mid x)$ within noisy regions of the data, the adversarial term optimises for low-entropy, label-like output, and allows multiple solutions. Therefore combining these losses in an additive manner, and enforcing them on the same set of parameters can be suboptimal—this prompts the generator to collapse to a deterministic output, as we show in Section 4 experimentally by using $\mathcal{L}_{\text{ce}}$-regularised baselines.

3.2. Calibrated adversarial refinement

In this work, we propose to avert potential conflict between the cross entropy and adversarial losses by decoupling them in a two-stage, cascaded architecture. This consists of a *calibration network* F_θ, optimised with $\mathcal{L}_{\text{ce}}$ from Eq. (2), the output of which is fed to a *refinement network* G_ϕ, optimised with an adversarial loss, which is in turn parametrised by

an auxiliary discriminator D_ψ trained with a binary cross entropy loss.

To account for the multimodality in the labels, we condition the refinement network on an additional extraneous noise variable $\epsilon \sim \mathcal{N}(0, 1)$, as done in the original GAN framework proposed by Goodfellow et al. [15]. In practice, we also condition the refinement network and the discriminator on the inputs x, however, we do not show this explicitly for notational convenience. More formally, using the non-saturated version of the adversarial loss [15], the objectives for the refinement and discriminator networks are given by:

$$\mathcal{L}_{\text{adv}}(\mathcal{D}, \theta, \phi) = -\mathbb{E}_{p_\mathcal{D}, p_\epsilon}[\log D_\psi(G_\phi(F_\theta(x), \epsilon))], \quad (3)$$

$$\mathcal{L}_{\text{D}}(\mathcal{D}, \theta, \phi, \psi) = -\mathbb{E}_{p_\mathcal{D}, p_\epsilon}[\log(1 - D_\psi(G_\phi(F_\theta(x), \epsilon)))] \\ - \mathbb{E}_{p_\mathcal{D}}[\log D_\psi(y)]. \quad (4)$$

To calibrate the predictive distribution, we impose diversity regularisation on G_ϕ by introducing a novel loss term, which we call the *calibration loss*, that encourages the sample average $\overline{G}_\phi(F_\theta(x)) := \mathbb{E}_{p(\epsilon)}[G_\phi(F_\theta(x), \epsilon)]$ to match the pixelwise class probabilities predicted by $F_\theta(x)$. Here, $\overline{G}_\phi(F_\theta(x))$ serves as a factorised approximation to the implicit predictive distribution of the refinement network. To this end, we define an auxiliary categorical likelihood q_ϕ as:

$$q_\phi(y \mid F_\theta(x)) = \prod_i^H \prod_j^W \prod_k^K \overline{G}_\phi(F_\theta(x))_{i,j,k}^{y_{i,j,k}}, \quad (5)$$

and optimise ϕ using the proposed calibration loss, formulated as:

$$\mathcal{L}_{\text{cal}}(\mathcal{D}, \theta, \phi) = \mathbb{E}_{p_\mathcal{D}, q_\phi}\left[\log \frac{q_\phi(y \mid F_\theta(x))}{q_\theta(y \mid x)}\right]. \quad (6)$$

This loss term expresses the the Kullback-Leibler divergence, $\text{KL}(q_\phi \parallel q_\theta)$[1]. Since both q_ϕ and q_θ are categorical distributions, the divergence can be computed exactly.

Notice that $\mathcal{L}_{\text{cal}}$ optimises through an approximation of the expectation $\overline{G}_\phi(F_\theta(x))$, rather than a single sampled prediction, therefore the model is not restricted to learning a mode-averaging solution for each input x. Consequently, $\mathcal{L}_{\text{cal}}$ is more compatible with $\mathcal{L}_{\text{adv}}$ than $\mathcal{L}_{\text{ce}}$ when a multimodal predictive distribution is desired. The total loss for the refinement network then becomes:

$$\mathcal{L}_{\text{G}}(\mathcal{D}, \theta, \phi) = \mathcal{L}_{\text{adv}}(\mathcal{D}, \theta, \phi) + \lambda \mathcal{L}_{\text{cal}}(\mathcal{D}, \theta, \phi), \quad (7)$$

where $\lambda \geq 0$ is a hyperparameter. Fig. 2 shows the interplay of F_θ, G_ϕ and D_ψ and the corresponding loss terms.

Intuitively, the calibration network F_θ serves three main purposes: (i) it sets a calibration target used by $\mathcal{L}_{\text{cal}}$ to regularise the predictive distribution of G_ϕ, (ii) it provides G_ϕ

---

[1]The choice of divergence is heuristically motivated and can be changed to fit different use-case requirements.
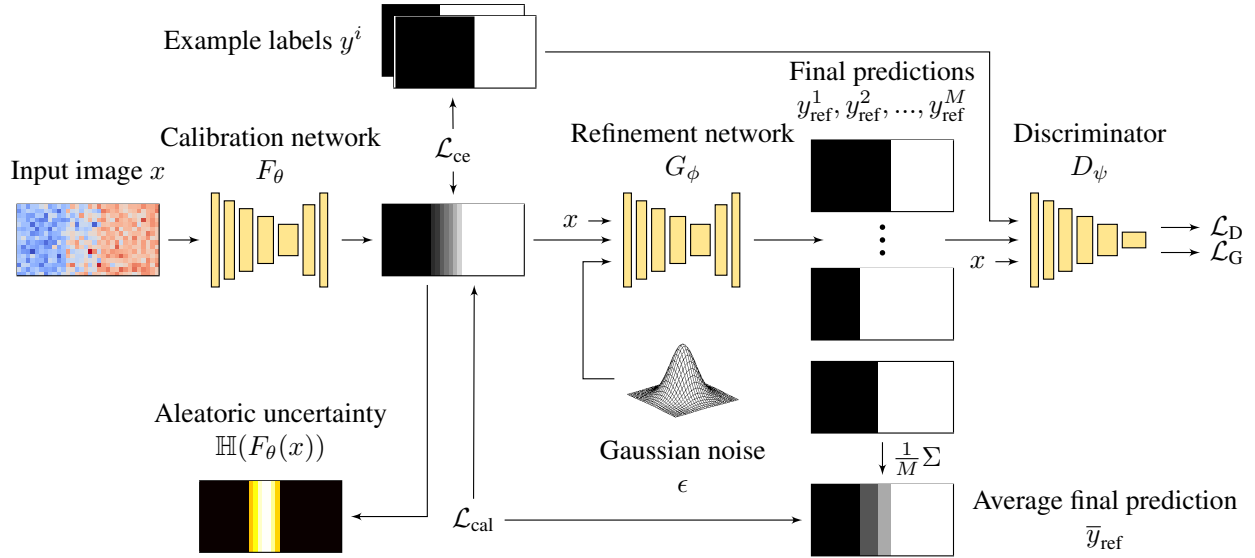
Figure 2: CAR model diagram based on the example in Fig. 1. First, the calibration network maps the input to a pixelwise distribution over the labels. This is then fed into the refinement network which samples an arbitrary number of diverse, crisp label proposals $y_{\text{ref}}^1, \ldots, y_{\text{ref}}^M$. To ensure calibration, the average of the final predictions is matched with the calibration target from the first stage through the $\mathcal{L}_{\text{cal}}$ loss. Additionally, the aleatoric uncertainty can be readily extracted from the calibration target, e. g. by computing the entropy $\mathbb{H}(F_\theta(x))$.

with an augmented representation of x enclosing probabilistic information about y, (iii) it accommodates the extraction of sample-free aleatoric uncertainty maps. The refinement network can therefore be interpreted as a stochastic sampler, modelling the inter-pixel dependencies to draw realistic samples from the explicit likelihood provided by the calibration network. Thus both the pixelwise class probability and object coherency are preserved. This approach leads to improved mode coverage, training stability and increased convergence speed, as demonstrated in Section 4.

3.3. Practical considerations

The gradients generated from the refinement network's loss function $\mathcal{L}_{\text{G}}$ are prevented from flowing into the calibration network, to ensure that F_θ learns an unbiased estimate of $\mathbb{E}_{p_\mathcal{D}}[y \mid x]$. As a consequence the weights of F_θ can be kept fixed, while the adversarial pair G_ϕ and D_ψ is being trained. This allows F_θ to be pretrained in isolation, thereby lowering the overall peak computational and memory requirements and improving training stability (see Algorithms 1 and 2 in Appendix A.1 for an outline of the training and inference procedures). Further, computing $\mathcal{L}_{\text{cal}}$ requires a Monte Carlo estimation of $\overline{G}_\phi(F_\theta(x))$, where the quality of the loss feedback increases with the sample count. This introduces a trade-off between training speed and prediction quality, however, modern deep learning frameworks allow for the samples to be subsumed in the batch dimension, and can therefore be efficiently computed on GPUs. Finally, our method

can augment any existing black-box model B for semantic segmentation, furnishing it with a calibrated multimodal predictive distribution. This can be done by conditioning F_θ on the output of B, as we demonstrate in Section 4.2.2.

4. Experiments

4.1. 1D bimodal regression

We give intuitive insight into the mechanics of the proposed calibration loss by designing and experimenting on a simple one-dimensional regression task. To create the dataset, an input $x \in [0, 1]$ is mapped to $y \in \mathbb{R}$ as follows:

$$y = \begin{cases} 0.5 - b + \epsilon, & x \in [0, 0.4) \\ (-1)^b(-1.25x + 1) + \epsilon, & x \in [0.4, 0.8) \\ \epsilon, & x \in [0.8, 1] \end{cases} \quad (8)$$

where $b \sim \text{Bernoulli}(\pi)$ and $\epsilon \sim \mathcal{N}(0, \sigma)$. We generate 9 different scenarios by varying the degree of mode selection probability $\pi \in \{0.5, 0.6, 0.9\}$ and the mode noise $\sigma \in \{0.01, 0.02, 0.03\}$.

For every data configuration, we use a 4-layer MLP for each of F_θ, G_ϕ and D_ψ, and train with and without calibration loss by setting the coefficient λ in Eq. (7) to 1 and 0, respectively. Note that unlike the categorical likelihood used in semantic segmentation tasks, we employ a Gaussian likelihood with fixed scale of 1. This changes the formulation of both Eqs. (2) and (6) to mean squared error losses
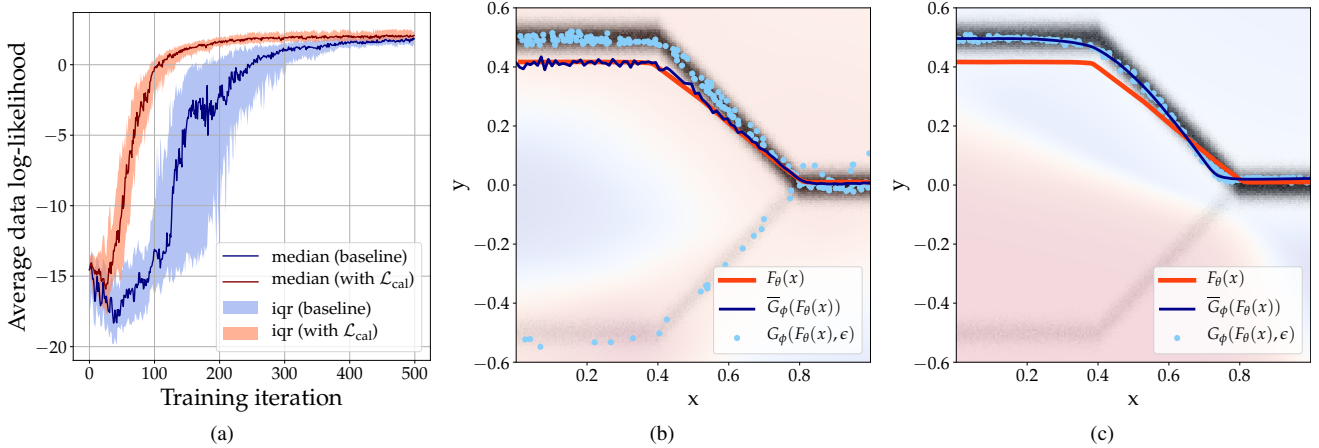
Figure 3: **(a)** Median and interquartile range (iqr) over the data log-likelihood, averaged over all 9×5×2 experiments. **(b)** High bias and noise configuration ($\pi = 0.9$, $\sigma = 0.03$) with calibration loss. The ground truth target is shown as black dots and the predicted samples as light blue dots. The predictions average in dark blue matches the calibration target in red. The discriminator output is shown in the background in shades of red (real) and blue (fake). **(c)** The same experiment configuration but without the proposed calibration loss, resulting in a mode collapse.

between ground truth labels y and predictions $\hat{y}$ for $\mathcal{L}_{\text{ce}}$, and between the output of the calibration network $F_\theta(x)$ and the average of multiple predictions $\overline{G}_\phi(F_\theta(x))$ for $\mathcal{L}_{\text{cal}}$ (see Appendix A.2). Finally, all runs are trained with a learning rate of $1\text{e}{-4}$ and each experiment is repeated five times.

The results, depicted in Fig. 3, illustrate that when using calibration loss, the optimisation process shows improved training stability, converges faster, and results in better calibrated predictions in comparison to the non-regularised baseline. Notice that $\mathcal{L}_{\text{cal}}$ also serves as an effective mechanism against mode collapse. This effect is more pronounced in data configurations with higher bias. More plots of the individual experiments are shown in Appendix B.1.

4.2. Stochastic semantic segmentation

In this section we examine the capacity of our model to learn shape and class multimodality in real-world segmentation datasets. We begin by sharing essential implementation details below.

Network architectures For the calibration network F_θ, we use the SegNet [2] encoder-decoder architecture. For G_ϕ, we designed a U-Net-style [47] architecture with 4 down- and upsampling blocks, each consisting of a convolutional layer, followed by a batch normalisation layer [24], a leaky ReLU activation, and a dropout layer [51] with 0.1 dropout probability. We use a base number of 32 channels, doubled or halved at every down- and upsampling transition. To propagate the sampled noise vector ϵ to the output, we inject it into every upsampling block of the network in a linear manner. To do so, we project ϵ using two fully connected

layers into scale and residual matrices with the same number of channels as the feature maps at the points of injection, and use these matrices to adjust the channel-wise mean and variance of the activations. This is similar to the mechanism used for adaptive instance normalisation [22]. We base the architecture for D_ψ on that used in DC-GAN [46] except that we remove batch normalisation. Any deviations from this setup are described in the corresponding sections.

Training details We utilise the Adam optimiser [29] with an initial learning rate of $2\text{e}{-4}$ for F_θ and G_ϕ, and $1\text{e}{-5}$ for D_ψ. The learning rates are linearly decayed over time and we perform scheduled updates to train the networks. Additionally, the discriminator loss is regularised by using the R_1 zero-centered gradient penalty term [43]. For a detailed list of hyperparameter values, see Appendix A.1.

Metrics Following [31, 32, 23, 4], we use the Generalised Energy Distance (GED) [52] metric:

$$D_{\text{GED}}^2(p_\mathcal{D}, q_\phi) = 2\mathbb{E}_{s \sim q_\phi, y \sim p_\mathcal{D}}[d(s, y)] - \mathbb{E}_{s, s' \sim q_\phi}[d(s, s')] - \mathbb{E}_{y, y' \sim p_\mathcal{D}}[d(y, y')],$$
(9)

where $d(s, y) = 1 - \text{IoU}(s, y)$. As an additional metric, we follow [32] in using the Hungarian-matched IoU (HM-IoU) [33]. In contrast to GED, which naively computes diversity as $1 - \text{IoU}$ between all possible pairs of ground truth or sampled predictions, HM-IoU finds the optimal 1:1 matching between all labels and predictions, and therefore is more representative of how well the learnt predictive distribution fits the ground truth.

All experiments are performed in triplicate and we report results as mean and standard deviation. Further details regarding the exact implementation of the GED and HM-IoU metrics for each experiment can be found in Appendix A.3.

4.2.1 Learning shape diversity on the LIDC dataset

To assess the accuracy and diversity of samples generated by our model, we use the Lung Image Database Consortium (LIDC) [1] dataset which consists of 1018 thoracic CT scans from 1010 lung cancer patients, graded independently by four expert annotators. We use the 180×180 crops from the preprocessed version of the LIDC dataset used and described in [31]. The dataset is split in 8882, 1996 and 1992 images in the training, validation and test sets respectively. All models are trained on lesion-centered 128×128 crops where at least one of the four annotations indicates a lesion. The final evaluation is performed on the provided test set.

In this task, we pretrain $F_\theta(x)$ with $\mathcal{L}_{ce}$ in isolation, fix its weights, and then train G_ϕ with $\mathcal{L}_G$, estimating $\mathcal{L}_{cal}$ with 20 samples from G_ϕ. Note that we disclose further experiments with varying sample size in Table 3 in Appendix B.2. As a control, we train using the same architecture but replace $\mathcal{L}_{cal}$ in the refinement network loss function with a cross entropy loss $\mathcal{L}_{ce}$, as done in [41, 14]. We denote this baseline throughout the manuscript as cGAN+$\mathcal{L}_{ce}$.

Our results show that the CAR model performs on par with other state-of-the-art methods w. r. t. the GED score, and outperforms them on the HM-IoU score (only available for [31, 32]). Numerical results are summarised in Table 1 and the diversity and fidelity of sampled predictions are illustrated in Fig. 4. In contrast, the $\mathcal{L}_{ce}$-regularised baseline collapses the predictive distribution, showing no perceptible diversity between samples (see Fig. 11b in Appendix B.2.3), which results in a stark increase in the mean GED score, and decrease in the HM-IoU score.

Method	GED ↓ (16)	GED ↓ (50)	GED ↓ (100)	HM-IoU ↑ (16)
Kohl et al. (2018) [31]	0.320 ± 0.030	—	$0.252 \pm$ N/A[1]	0.500 ± 0.030
Kohl et al. (2019) [32]	0.270 ± 0.010	—	—	0.530 ± 0.010
Hu et al. (2019) [21]	—	0.267 ± 0.012	—	—
Baumgartner et al. (2019) [4]	—	—	$\mathbf{0.224 \pm N/A}$[2]	—
Monteiro et al. (2020) [44]	—	—	0.225 ± 0.002[2]	—
cGAN+$\mathcal{L}_{ce}$	0.639 ± 0.002	—	—	0.477 ± 0.004
CAR (ours)	$\mathbf{0.264 \pm 0.002}$	$\mathbf{0.248 \pm 0.004}$	0.243 ± 0.004[2]	$\mathbf{0.592 \pm 0.005}$

[1] This score is taken from [4].
[2] Note that when following the data split methodology used in [4, 44] and computing the GED (100) metric, we achieve a score of 0.228 ± 0.009 instead of 0.243 ± 0.004 (see Appendix A.3).

Table 1: Mean GED and HM-IoU scores on LIDC. Top section: approaches using the original data splits defined by [31], which we also adhere to; middle: approaches using random data splits; bottom: the $\mathcal{L}_{ce}$-regularised baseline and our CAR model. The three central columns show the GED score computed with 16, 50 and 100 samples, respectively. The rightmost column shows the HM-IoU score, computed with 16 samples. The arrows ↑ and ↓ indicate if higher or lower score is better.
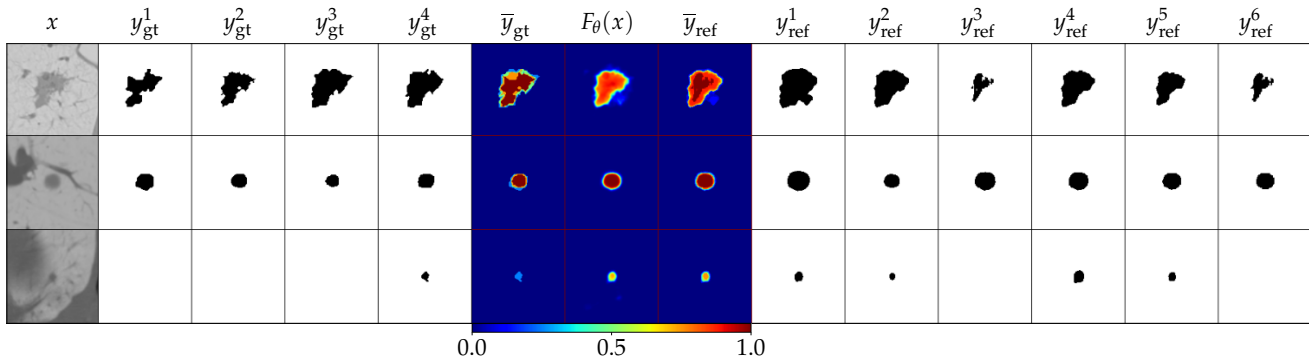

Figure 4: LIDC validation samples. From left to right: an input image x, followed by the four ground truth annotations $y_{gt}^1 \dots y_{gt}^4$, the mean of the labels $\bar{y}_{gt}$, the output of the calibration network $F_\theta(x)$, the mean of the six refinement network samples $\bar{y}_{ref}$, shown in columns $y_{ref}^1 \dots y_{ref}^6$.

4.2.2 Learning a calibrated distribution on a multimodal Cityscapes dataset

The Cityscapes dataset contains 1024×2048 RGB images of urban scenes, and corresponding segmentation maps. It consists of 2975 training, 500 validation and 1525 test images. Following [31], we use a stochastic version of the Cityscapes dataset with 19 semantic classes, downsample images and segmentation maps to a spatial resolution of 256×512, and report results on the validation set. Controlled multimodality is established by augmenting the dataset with 5 new classes: *sidewalk2*, *person2*, *car2*, *vegetation2* and *road2*, introduced by flipping their original counterparts with probabilities $^8/_{17}$, $^7/_{17}$, $^6/_{17}$, $^5/_{17}$ and $^4/_{17}$, respectively (see Fig. 5a), giving the dataset a total of 24 semantic classes.

To demonstrate that our approach can be easily integrated on top of any existing black-box segmentation model B, we employ the network from [54], trained on the official Cityscapes dataset, achieving a mIoU of 0.79 on the test set. We utilise its predictions as input to a smaller version of our calibration network F_θ comprising 5 convolutional blocks, each composed of a 3×3 convolutional layer followed by batch normalisation, a leaky ReLU activation, and a dropout layer with 0.1 dropout rate. We pretrain the calibration network in isolation, and subsequently apply it in inference mode while adversarially training the refinement network. We use a batch size of 16, and train with $\mathcal{L}_G$, estimating $\mathcal{L}_{\text{cal}}$ with 7 samples from G_ϕ. The same baseline from the LIDC experiment is employed, where $\mathcal{L}_{\text{cal}}$ is replaced with $\mathcal{L}_{\text{ce}}$. As a second control experiment, we completely omit the calibration network and instead condition the refinement network on the known ground truth pixelwise categorical distribution over the label, representing the ideal output of F_θ. This baseline allows us to directly evaluate the quality of sampling administered from the refinement network, and we denote it as cGAN+$\mathcal{L}_{\text{cal}}$ (ground truth).

When training the refinement network with a cross entropy loss instead of the calibration loss $\mathcal{L}_{\text{cal}}$, the predictive distribution collapses, making the output deterministic. Conversely, when we train our refinement network with $\mathcal{L}_{\text{cal}}$, the learnt predictive distribution is well adjusted, with high diversity and reconstruction quality, significantly outperforming the current state-of-the-art, as shown in Table 2. Fig. 5b displays representative sampled predictions from our model for three input images and Fig. 5c illustrates the corresponding aleatoric uncertainty maps extracted from $F_\theta(x)$. The learnt multimodality and noise in the dataset are reflected by regions of high entropy, where objects belonging to the different stochastic classes consistently display distinct shades of red, corresponding to their respective flip probabilities. Finally, we show that when using the ground truth distribution as the input and calibration target to the refinement network, we attain an almost perfect GED score (0.038 ± 0.00).

Since we manually set the flip probabilities for each stochastic class in this dataset, we can directly assess the calibration of our model by comparing the ground truth probabilities to the predicted probabilities from the calibration or refinement network. For F_θ we use the mean confidence values for each class, and for G_ϕ we obtain the empirical mean class probabilities via $\overline{G}_\phi(F_\theta(x))$, computed from 16 samples (see Appendix A.4 for more details). The ensuing results are shown graphically in Fig. 6, which illustrates the calibration of our models on the stochastic classes, evaluated over the entire dataset. This demonstrates that our models are well calibrated, with the calibration offset, computed as the absolute difference between the ground truth and predicted probabilities, being approximately 6% in the worst case (class "car2" for the calibration network). Note that the average calibration offset for $F_\theta(x)$ across the stochastic classes is 1.6%. Further, the ground truth conditioned baseline is almost perfectly calibrated, in accord to the near-optimal GED score reported in Table 2. Thus, we demonstrate that G_ϕ
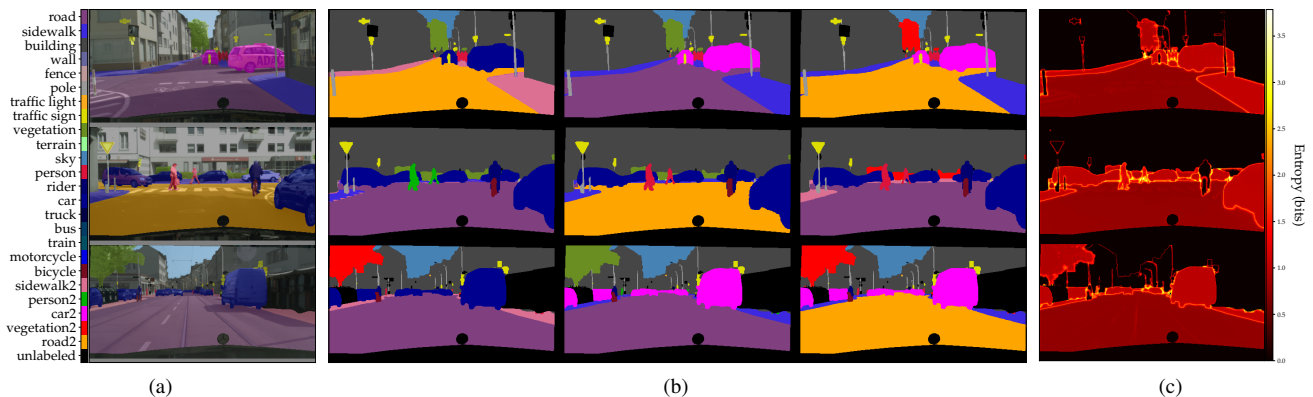

Figure 5: **(a)** Input images overlaid with the corresponding labels. **(b)** Samples obtained from the refinement network. **(c)** Aleatoric uncertainty computed as the entropy of the calibration output.
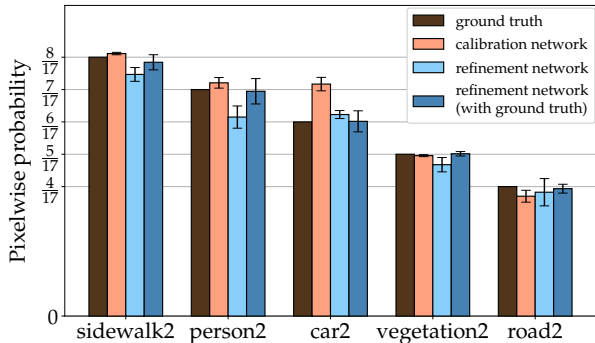
Figure 6: Calibration of the pixelwise probabilities of the five stochastic classes. Note that the calibration network (in orange) is conditioned on black-box predictions.

Method	GED
Kohl et al. (2018) [31]	$0.206 \pm$ N/A
cGAN+$\mathcal{L}_{ce}$	0.632 ± 0.07
CAR (ours)	$\mathbf{0.164 \pm 0.01}$
cGAN+$\mathcal{L}_{cal}$ (ground truth)	0.038 ± 0.00

Table 2: Mean GED scores on the modified Cityscapes. Top section: competing model; middle: $\mathcal{L}_{ce}$-regularised baseline and CAR model; bottom: ground truth calibrated refinement network (cGAN). GED scores are computed using 16 samples.

successfully learns calibrated refinement of the predictions from F_θ, where the quality of the final predictive distribution depends on the quality of $F_\theta(x)$.

In order to further scrutinise the calibration quality of $F_\theta(x)$, we construct a reliability diagram and compute the corresponding expected calibration error (ECE), following [17]. To create the diagram, each pixel is considered independently, and the class confidences are binned into 10 equal intervals of size 0.1. We then compute the accuracy for all predictions in each bin. Fig. 7 shows the reliability diagram for the calibration network, where the orange bars depict the calibration gap, defined as the difference between the mean confidence for each interval and the corresponding accuracy. The corresponding ECE score amounts to 2.15%. Note that this also considers the average calibration error computed for the stochastic classes, where we randomly sample the labels according to the predefined probabilities. Hence, we confirm that $F_\theta(x)$ is well calibrated.

An important outstanding issue in our approach is that the calibration network may not perfectly capture the class-probabilities, e. g. as seen from the *car2* category in Fig. 6. Since we calibrate the refinement network relative to the calibration target provided by F_θ, such errors can be propagated
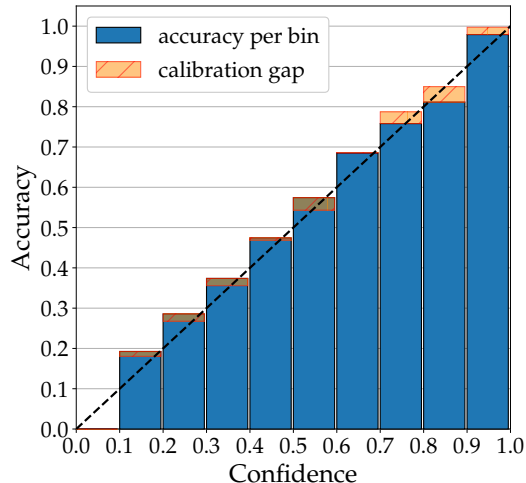

Figure 7: Reliability diagram for the calibration network. The corresponding ECE score is 2.15%.

into G_ϕ. Calibration issues in modern neural networks are well documented [17, 34, 56]—these have been attributed to several factors, such as long-tailed data distributions, out-of-distribution inputs, specific network architectural elements or optimising procedures. Even though we did not find it necessary, ad hoc solutions to miscalibration, such as temperature scaling [20], Dirichlet calibration [34] etc., can be readily applied to the calibration network to improve the predicted probabilities, and thus the overall calibration of our model. Therefore, an important direction for future work is towards improving the calibration of deep neural networks.

5. Conclusion

In this work, we developed a novel framework for semantic segmentation capable of learning a calibrated multimodal predictive distribution, closely matching the ground truth distribution of labels. We attained improved results on a modified Cityscapes dataset and competitive scores on the LIDC dataset, indicating the utility of our approach on real-world datasets. We also showed that our approach can be easily integrated into any off-the-shelf, deterministic, black-box semantic segmentation model, enabling sampling an arbitrary number of plausible segmentation maps. By highlighting regions of high data uncertainty and providing multiple valid label proposals, our approach can be used to identify and resolve ambiguities, diminishing risk in safety-critical systems. Therefore, we expect our approach to be particularly beneficial for applications such as map making for autonomous driving or computer-assisted medical diagnostics. Finally, even though the primary focus of this work is semantic segmentation, we demonstrated the versatility of our method with an illustrative toy regression problem, alluding to a broader applicability beyond semantic segmentation.

References

[1] Samuel G Armato III, Geoffrey FMcLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. 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(2):915–931, 2011. 2, 6

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 5, 2

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2745–2754, 2017. 2

[4] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer, 2019. 2, 5, 6, 7

[5] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Bayesian prediction of future street scenes using synthetic likelihoods. In *International Conference on Learning Representations*, 2018. 2

[6] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006. 3

[7] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *CVPR*, 2016. 1

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 1

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[11] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 1

[12] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *ICLR workshop track*, 2016. 2

[13] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 2

[14] Mohsen Ghafoorian, Cedric Nugteren, Nóra Baka, Olaf Booij, and Michael Hofmann. El-gan: Embedding loss driven generative adversarial networks for lane detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2, 3, 6

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[16] Alex Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356, 2011. 7

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017. 2, 8, 6

[18] Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. Noise contrastive priors for functional uncertainty. *arXiv preprint arXiv:1807.09289*, 2018. 7

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Iclr*, 2(5):6, 2017. 7

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 8

[21] 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*, pages 137–145. Springer, 2019. 2, 6, 7

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 5

[23] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 2, 5

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 5

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[26] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 11–19, 2017. 1, 2

[27] Konstantinos Kamnitsas, Wenjia Bai, Enzo Ferrante, Steven McDonagh, Matthew Sinclair, Nick Pawlowski, Martin Rajchl, Matthew Lee, Bernhard Kainz, Daniel Rueckert, et al.

Ensembles of multiple models and architectures for robust brain tumour segmentation. In *International MICCAI Brainlesion Workshop*, pages 450–462. Springer, 2017. 2

[28] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 2, 3, 7, 8

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5, 1

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 7

[31] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018. 2, 5, 6, 7, 8, 3

[32] Simon AA Kohl, Bernardino Romera-Paredes, Klaus H Maier-Hein, Danilo Jimenez Rezende, SM Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*, 2019. 2, 5, 6, 7

[33] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52(1):7–21, 2005. 5

[34] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, pages 12295–12305, 2019. 2, 8, 6

[35] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017. 2

[36] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. 7

[37] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2

[38] Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. In *Advances in Neural Information Processing Systems*, pages 2119–2127, 2016. 1

[39] Geert Litjens, Oscar Debats, Wendy van de Ven, Nico Karssemeijer, and Henkjan Huisman. A pattern recognition approach to zonal segmentation of the prostate on mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 413–420. Springer, 2012. 2

[40] Christos Louizos, Xiahan Shi, Klamer Schutte, and Max Welling. The functional neural process. *arXiv preprint arXiv:1906.08324*, 2019. 7

[41] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016. 2, 3, 6

[42] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 2

[43] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018. 5, 1

[44] Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *arXiv preprint arXiv:2006.06015*, 2020. 2, 6

[45] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 2

[46] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 5

[47] 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*, pages 234–241. Springer, 2015. 1, 2, 5

[48] Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017. 2

[49] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3591–3600, 2017. 2

[50] Laurens Samson, Nanne van Noord, Olaf Booij, Michael Hofmann, Efstratios Gavves, and Mohsen Ghafoorian. I bet you are wrong: Gambling adversarial networks for structured semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 3

[51] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 5

[52] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013. 5, 2

[53] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5229–5238, 2019. 1

[54] TensorFlow DeepLab Model Zoo. Pretrained model: xception71_dpc_cityscapes_trainfine, 2020. 7

[55] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In *International Conference on Learning Representations*, 2018. 2

[56] Jize Zhang, Bhavya Kailkhura, and T Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. *arXiv preprint arXiv:2003.07329*, 2020. 8

[57] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[58] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017. 2