Temperature-Scaled Uncertainty-Aware Selective Segmentation with MC Dropout

Vignesh Mulbagal Venkataravanappa
Final Term Project

Abstract

Deep segmentation networks can achieve strong mean intersection-over-union (mIoU) while still producing silent, overconfident pixel-level failures. This report studies *selective segmentation*: predicting only where the model is confident and abstaining elsewhere. I implement a U-Net with test-time Monte Carlo (MC) Dropout to approximate Bayesian inference and to compute pixel-wise uncertainty maps. To improve probability calibration without retraining, I fit a single *temperature scaling* parameter on a held-out validation set using pixel-wise negative log-likelihood. I evaluate reliability using (i) pixel-wise expected calibration error (ECE), (ii) fixed-coverage *selective risk* curves (error rate on accepted pixels), and (iii) selective mIoU as a function of coverage. On Oxford-IIIT Pet trimap segmentation, temperature scaling improves ECE from 0.0098 to 0.0084 and selective segmentation reduces risk from 0.132 at full coverage to 0.024 at ≈ 0.60 coverage, while increasing selective mIoU from 0.677 to a peak of 0.789 at ≈ 0.70 coverage. I provide qualitative uncertainty maps that localize ambiguity around object boundaries and challenging background regions.

1 Introduction

Semantic segmentation assigns a label to every pixel and is a core component of modern perception pipelines in robotics, medical imaging, autonomous navigation, and content understanding. While the community commonly reports average accuracy metrics such as mIoU, many real deployments require a stronger guarantee: the system should *know when it does not know*. In practice, segmentation models can be wrong yet confident, particularly near object boundaries, under low-quality imaging, or in visually ambiguous regions. Because pixel-level outputs are dense and high-dimensional, even a small fraction of incorrect pixels can cause substantial downstream harm.

This project focuses on **reliability** rather than only average accuracy. I develop a segmentation system that can:

1. **Quantify uncertainty** via MC Dropout sampling at test time.
2. **Calibrate confidence** via post-hoc temperature scaling fitted on validation data.
3. **Abstain** on high-uncertainty pixels and evaluate the resulting **risk–coverage trade-off**.

Contributions.

- I implement MC Dropout uncertainty for a U-Net segmentation model and compute both predictive entropy and mutual information style uncertainty scores.
- I adapt **temperature scaling** to dense segmentation by fitting a single temperature T using pixel-wise negative log-likelihood on the validation set.
- I introduce a **fixed-coverage** evaluation protocol for selective segmentation, reporting **selective risk** (pixel error rate on accepted pixels) and selective mIoU at target coverage levels.

Preprint.

- I report quantitative and qualitative results on Oxford-IIIT Pet trimap segmentation, demonstrating improved calibration and actionable abstention behavior.

Problem statement. Given an image x and ground-truth mask y, I want a predictor that outputs both a mask $\hat{y}$ and an uncertainty map u. A downstream decision rule uses u to accept or reject pixels, trading coverage for reduced error.

2 Related Work

2.1 Segmentation architectures

Encoder–decoder architectures with skip connections, such as U-Net, remain strong baselines for segmentation because they preserve spatial information while learning high-level semantic features [Ronneberger et al., 2015]. In this project I use U-Net to isolate the effects of uncertainty and calibration from architectural complexity.

2.2 Uncertainty estimation with MC Dropout

Bayesian deep learning formalizes predictive uncertainty as a distribution over model parameters. MC Dropout approximates Bayesian inference by retaining dropout at test time and averaging predictions over multiple stochastic passes [Gal and Ghahramani, 2016]. This approach is computationally attractive because it requires no architectural changes beyond standard dropout layers and avoids training multiple separate models.

2.3 Calibration and temperature scaling

Modern deep networks are often miscalibrated, meaning predicted probabilities do not match empirical correctness. Temperature scaling is a widely used post-hoc method that fits a single scalar temperature T on validation data to improve calibration without changing the predicted class labels [Guo et al., 2017]. For segmentation, confidence calibration is particularly important because thresholds are often used to filter predictions.

2.4 Selective prediction and abstention

Selective prediction (classification with a reject option) studies systems that trade coverage for lower risk by abstaining on uncertain inputs [Geifman and El-Yaniv, 2017]. This project adapts the same evaluation philosophy to segmentation by abstaining at the pixel level. The central reliability object becomes a **risk–coverage** curve that answers: *how accurate are the predictions I choose to keep?*

3 Methods

3.1 Segmentation model

Let $f_\theta(x) \in \mathbb{R}^{C\times H\times W}$ be the logit output of the model. I compute per-pixel probabilities:

$$p_\theta(c \mid x, i) = \operatorname{softmax}\left(f_\theta(x)_i\right)_c . \tag{1}$$

I use a U-Net with Dropout2D inserted inside convolution blocks.

3.2 MC Dropout predictive distribution

At test time, I keep dropout active and draw T stochastic forward passes:

$$p^{(t)}(c \mid x, i) = \operatorname{softmax}\left(\frac{f_\theta^{(t)}(x)_i}{T_{\text{temp}}}\right)_c , \quad t = 1, \ldots, T, \tag{2}$$

where T_{temp} is a temperature for calibration (Section 3.4). The predictive mean is:

$$\bar{p}(c \mid x, i) = \frac{1}{T}\sum_{t=1}^{T} p^{(t)}(c \mid x, i), \tag{3}$$

and the predicted label is $\hat{y}_i = \arg\max_c \bar{p}(c \mid x, i)$.

3.3 Uncertainty measures

I compute two pixel-wise uncertainty measures.

Predictive entropy (total uncertainty).

$$u_{\text{ent}}(x, i) = \mathrm{H}(\bar{p}(\cdot \mid x, i)) = -\sum_{c=1}^{C} \bar{p}(c \mid x, i) \log \bar{p}(c \mid x, i). \tag{4}$$

Mutual information (epistemic proxy).

$$u_{\text{MI}}(x, i) = \mathrm{H}(\bar{p}) - \frac{1}{T} \sum_{t=1}^{T} \mathrm{H}\left(p^{(t)}\right). \tag{5}$$

In our main experiments I report predictive entropy; MI can be added as an ablation.

3.4 Temperature scaling for segmentation

Temperature scaling fits a single scalar $T_{\text{temp}} > 0$ on validation data by minimizing pixel-wise negative log-likelihood (NLL):

$$\min_{T_{\text{temp}}>0} \sum_{(x,y)\in\mathcal{D}_{\text{val}}} \sum_{i} -\log \operatorname{softmax}\left(\frac{f_\theta(x)_i}{T_{\text{temp}}}\right)_{y_i}. \tag{6}$$

Unlike retraining, temperature scaling changes only the *softmax sharpness*, preserving the argmax predictions. I optimize $\log T_{\text{temp}}$ using Adam to avoid large-memory LBFGS graphs.

3.5 Selective segmentation and fixed-coverage risk

Given uncertainty $u(x, i)$ and threshold τ, define accepted pixels:

$$m_i(\tau) = \mathbb{I}[u(x, i) \le \tau]. \tag{7}$$

Coverage is the fraction of accepted pixels:

$$\mathrm{cov}(\tau) = \frac{1}{HW} \sum_i m_i(\tau). \tag{8}$$

Fixed-coverage thresholds. For a target coverage $\alpha \in (0, 1]$, I set τ_α as the α-quantile of uncertainty values (estimated on validation images). This allows evaluating methods at comparable coverage points.

Selective risk. Selective risk is the pixel error rate conditioned on acceptance:

$$\mathrm{risk}(\alpha) = \frac{\sum_i \mathbb{I}[\hat{y}_i \neq y_i] \cdot m_i(\tau_\alpha)}{\sum_i m_i(\tau_\alpha)}. \tag{9}$$

Lower is better at the same coverage.

Selective mIoU. I also compute selective mIoU on accepted pixels by treating rejected pixels as "ignore" in the IoU computation.

3.6 Calibration metric: pixel-wise ECE

I compute pixel-wise expected calibration error (ECE). Let confidence $s_i = \max_c \bar{p}(c \mid x, i)$. Partition confidences into B bins; then:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \left|\mathrm{acc}(S_b) - \mathrm{conf}(S_b)\right|, \tag{10}$$

where S_b contains pixels in bin b.

3.7 Algorithm

Algorithm 1 Temperature-scaled MC Dropout selective segmentation

Require: Image x, model f_θ, MC samples T, temperature T_{temp}, target coverage α

Enable dropout layers at inference
for $t = 1$ to T **do**
    $p^{(t)} \leftarrow \mathrm{softmax}(f_\theta^{(t)}(x)/T_{\text{temp}})$
end for
$\bar{p} \leftarrow \frac{1}{T}\sum_t p^{(t)}$, $\hat{y} \leftarrow \arg\max_c \bar{p}_c$
$u \leftarrow \mathrm{H}(\bar{p})$ ▷ or u_{MI}
$\tau_\alpha \leftarrow \text{Quantile}(u, \alpha)$
$m \leftarrow \mathbb{I}[u \le \tau_\alpha]$
Report $\mathrm{risk}(\alpha)$, selective mIoU using mask m

4 Experiments

4.1 Dataset and setup

I use the Oxford-IIIT Pet dataset with trimap segmentation annotations (3 classes: pet, border, background) [Parkhi et al., 2012]. I resize images to 256×256 and normalize with ImageNet mean/std. The dataset sizes from our run are:

- Train: 3128 images
- Validation: 552 images
- Test: 3669 images

4.2 Training details

I train for 10 epochs using AdamW with learning rate 2×10^{-4}, weight decay 10^{-4}, batch size 16, and AMP. The loss is a weighted combination: $0.7\cdot$ cross-entropy + $0.3\cdot$ Dice loss. The U-Net base width is 32 with dropout probability $p = 0.20$. The best validation mIoU achieved was 0.6769.

4.3 Temperature scaling results

I fit temperature on the validation set. The learned temperature was:

$$T_{\text{temp}} = 1.0104.$$

Validation NLL improved slightly from 0.3445 to 0.3444, consistent with the model already being well calibrated. On the test set, pixel ECE improved from 0.0098 to 0.0084, indicating a measurable calibration gain despite the small temperature change.

Table 1: Calibration results (test set).

Method	Pixel ECE ↓	Notes
No temperature scaling	0.0098	baseline confidence
Temperature scaling ($T = 1.0104$)	0.0084	post-hoc calibration

4.4 Selective segmentation at fixed coverage

I evaluate fixed target coverages $\alpha \in \{1.00, 0.95, \ldots, 0.60\}$. Thresholds τ_α are estimated from validation uncertainties and applied on the test set. Table 2 summarizes results. As coverage decreases, selective risk drops sharply (better reliability), while selective mIoU increases up to ≈ 0.70 coverage and then slightly decreases due to discarding too many informative pixels.

Table 2: Fixed-coverage selective segmentation results (test set).

Target coverage α	Realized coverage	Selective risk ↓	Selective mIoU ↑
1.00	1.0000	0.1322	0.6773
0.95	0.9489	0.1113	0.7046
0.90	0.8986	0.0939	0.7264
0.85	0.8474	0.0769	0.7483
0.80	0.7971	0.0611	0.7689
0.75	0.7476	0.0485	0.7830
0.70	0.6992	0.0386	0.7893
0.65	0.6516	0.0306	0.7870
0.60	0.6027	0.0242	0.7734

4.5 Risk–coverage and mIoU–coverage curves

Figure 1 shows the selective risk curve, and Figure 2 shows the selective mIoU curve (generated from the tabulated results). These curves are the primary reliability outputs of this project.

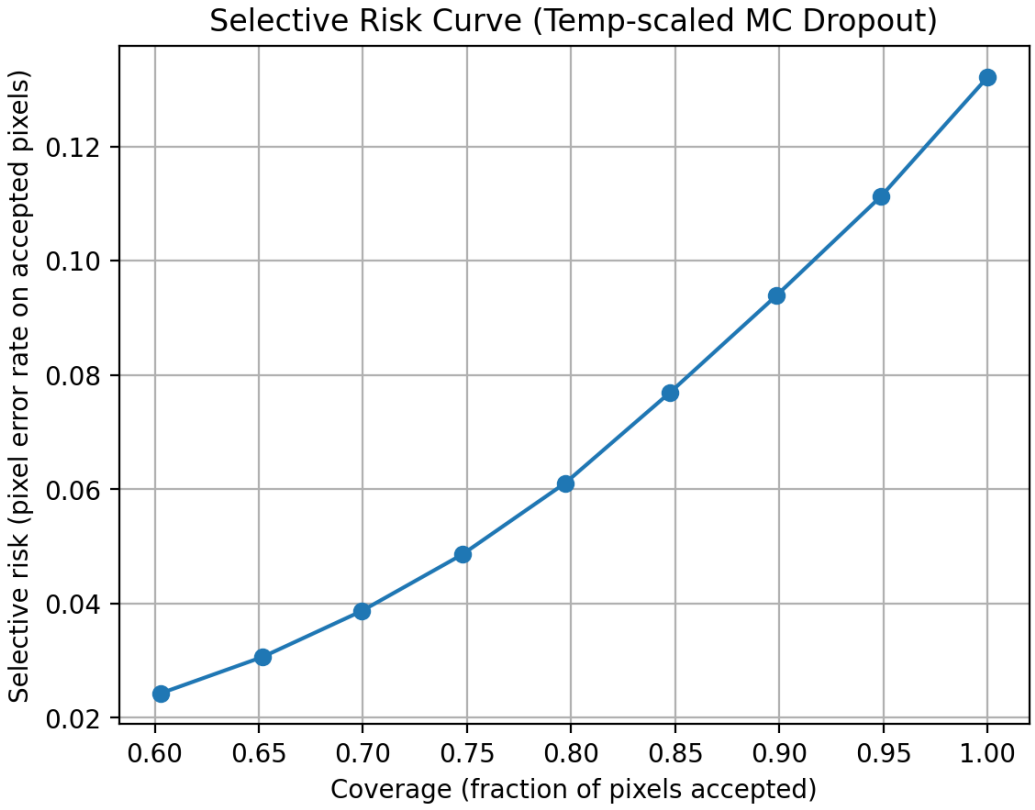

Figure 1: Selective risk vs coverage (temp-scaled MC Dropout). Risk is pixel error rate on accepted pixels.

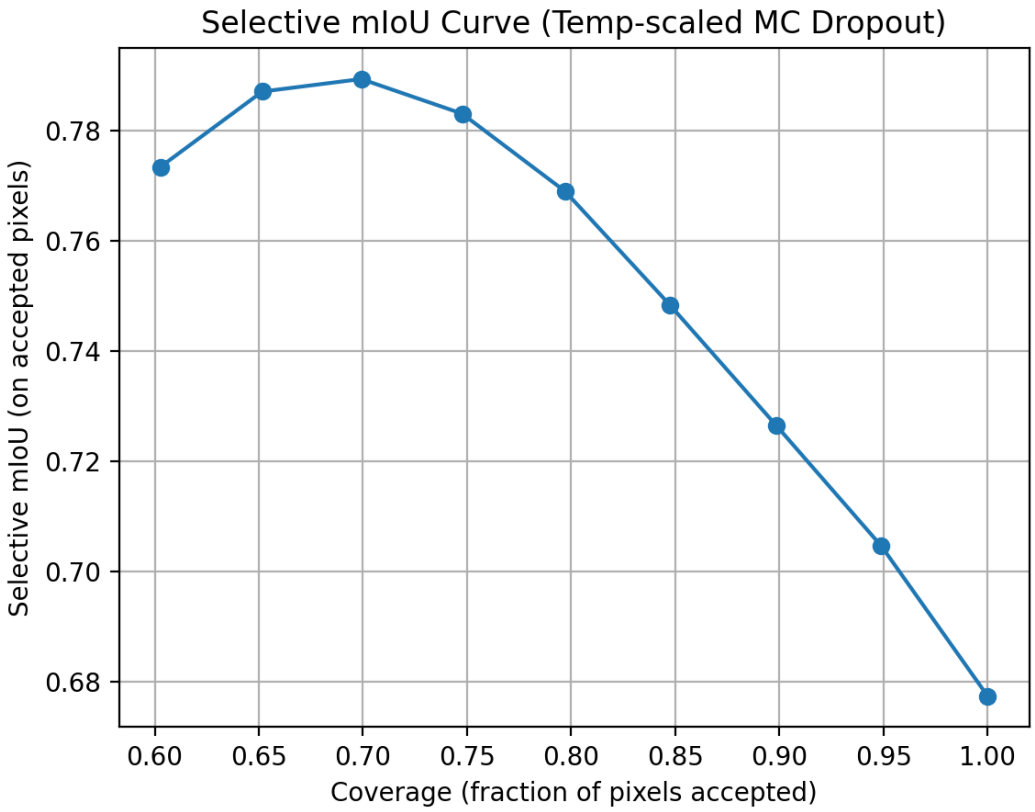

Figure 2: Selective mIoU vs coverage (temp-scaled MC Dropout). Selective mIoU peaks around 0.70 coverage.

4.6 Qualitative uncertainty analysis

Figure 3 presents a representative example with the input image, ground truth (GT), prediction, and predictive entropy. Uncertainty concentrates near boundaries and in ambiguous regions, which aligns with the intuition that the model is less certain where pixel evidence is mixed. In the shown example, the model misses some structure in the upper left and produces stray false positives; predictive entropy highlights these regions as uncertain, making abstention a reasonable decision.

4.7 Discussion

Reliability gains. At full coverage (≈ 1.0), risk is 0.132. By abstaining on the most uncertain 40% of pixels (coverage ≈ 0.60), risk reduces to 0.024, a relative reduction of $\approx 82\%$. This demonstrates that uncertainty estimates are *actionable*: they identify error-prone pixels.

Accuracy–coverage trade-off. Selective mIoU improves from 0.677 at full coverage to a peak of 0.789 near 0.70 coverage. This suggests that a moderate abstention rate can both reduce error and improve the quality of the remaining segmentation mask. Past a point, rejecting too many pixels reduces IoU because the evaluation ignores rejected pixels and the remaining set may over-represent easy regions.

Calibration. The small temperature ($T = 1.0104$) indicates that the base model is already relatively well calibrated; nevertheless, ECE improves. This matters because many decision rules rely on calibrated confidence. In larger models or under distribution shift, temperature scaling often yields larger gains.

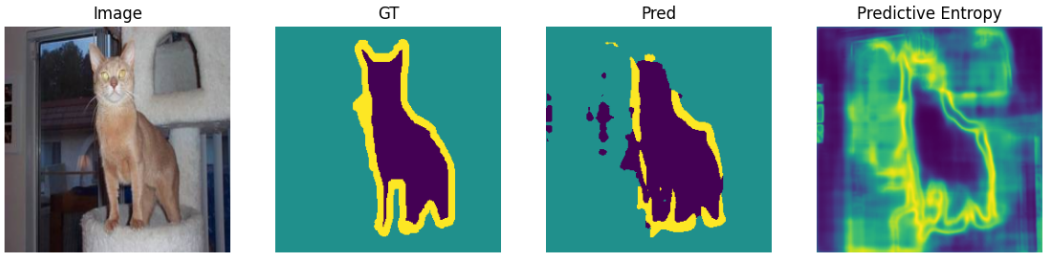

Figure 3: Qualitative example: input image, ground truth, prediction, and predictive entropy (uncertainty).

Limitations and future work. I focused on in-distribution evaluation plus selective reliability. Future work can extend this pipeline to explicit distribution shift by applying controlled corruptions (noise/blur/compression) and measuring how the risk–coverage curve changes. Another extension is spatially coherent abstention (reject regions/connected components) to avoid scattered "holes" in the mask.

5 Conclusion

I presented a reliability-focused segmentation system that combines MC Dropout uncertainty with temperature scaling and evaluates pixel-wise abstention using fixed-coverage selective risk. On Oxford-IIIT Pet trimap segmentation, temperature scaling improves test ECE from 0.0098 to 0.0084. More importantly, selective segmentation offers a strong risk–coverage trade-off: risk decreases from 0.132 at full coverage to 0.024 at ≈ 0.60 coverage, and selective mIoU improves up to 0.789 around ≈ 0.70 coverage. These results support a key conclusion: uncertainty estimation becomes most valuable when paired with explicit decision rules and reliability metrics, rather than being reported only as qualitative heatmaps.

A Additional Implementation Details

A.1 Compute and reproducibility

All experiments were run on a CUDA-enabled GPU in Google Colab. Random seeds were fixed for Python, NumPy, and PyTorch. Images were resized to 256×256. Training used AMP for efficiency.

A.2 Memory-stable MC inference

A naive implementation stacks a $T \times B \times C \times H \times W$ tensor of probabilities, which can be memory expensive. I instead use a streaming formulation that accumulates $\sum_t p^{(t)}$ and $\sum_t \mathrm{H}(p^{(t)})$ without storing all samples.

A.3 Why fixed-coverage evaluation

Thresholding uncertainty at arbitrary values can be hard to compare across methods and datasets. Fixed-coverage evaluation chooses thresholds by uncertainty quantiles, ensuring each point on the curve corresponds to a clear operating regime (e.g., "accept 80% of pixels").

B Extra analysis: selecting an operating point

In practice, an application chooses an operating point based on acceptable risk and available human review capacity. For example, if the goal is risk ≤ 0.05, Table 2 suggests operating at coverage around 0.75 (risk 0.0485), retaining roughly three quarters of pixels while maintaining low error on those pixels.

C Ethical considerations

Abstention can reduce silent failures but may shift burden to human reviewers or downstream fallback systems. In a real deployment, one should define clear procedures for rejected pixels (e.g., conservative default labels, routing to manual inspection, or additional sensing) and verify that abstention does not disproportionately affect particular classes or subgroups due to dataset bias.

References

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *NeurIPS*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.