QUTCC 🤗: Quantile Uncertainty Training and Conformal Calibration for Imaging Inverse Problems

Cassandra Tong Ye
cassye@cs.cornell.edu

Shamus Li
shamus@cs.cornell.edu

Tyler King
ttk22@cornell.edu

Kristina Monakhova
monakhova@cornell.edu

Abstract

While deep learning offers tremendous promise for scientific and medical imaging, any failures and hallucinations (predictions that do not coincide with reality) are hard to pinpoint and can have serious downstream consequences. Uncertainty estimation techniques, such as conformal prediction, can help by predicting statistically valid error bars for a model's prediction. However, popular conformal prediction methods were not designed for high-dimensional image-valued problems and do not take into account spatial correlations within an image during conformal calibration, resulting in larger-than-necessary uncertainty intervals. We propose a practical simultaneous quantile regression method that enables non-linear, spatially-adaptive scaling during conformal calibration. Our method, QUTCC uses a U-Net architecture with a quantile embedding to learn a full conditional quantile distribution during training, and then leverages this non-linear, learned function for *spatially-adaptive conformal calibration*. At test time, our method can efficiently estimate uncertainty intervals with pixel-marginal coverage guarantees. In addition, QUTCC can also predict pixel-wise conditional probability density estimates without any built-in distributional assumptions. We evaluate our method on several denoising problems, accelerated magnetic resonance imaging, and quantitative phase microscopy. Our method consistently produces tighter uncertainty intervals than prior conformal methods at the same coverage level, can predict plausible conditional distributions for different tasks, and in some cases, high-uncertainty regions can help us locate hallucinations in a model's prediction.

1 Introduction

Deep learning offers tremendous promise in scientific and medical imaging by enabling the algorithmic extraction of meaningful signals from sparser, noisier, and more underdetermined measurements [1–3]. For example, deep learning has enabled image restoration in fluorescence microscopy from 60-fold fewer photons [4], can accelerate magnetic resonance imaging by requiring fewer k-space measurements for compelling reconstruction [5], and can aid in black hole recovery from sparse interferometric measurements [6]. However, this promise is counteracted by significant risk: *can we trust the model's prediction? How can we determine if the model's prediction is incorrect?*

Deep learning models will often produce compelling, high-resolution reconstructions, but these predictions may not coincide with reality. Failures and hallucinations (i.e., predictions that do not coincide with reality) are incredibly hard to pinpoint. However, when using deep learning for scientific and medical imaging, it is critical for scientists and doctors to understand how and when a model's prediction may be wrong. In this paper, we introduce a method to predict both statistically valid error bars and pixel-wise conditional probability density functions for imaging inverse problems.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

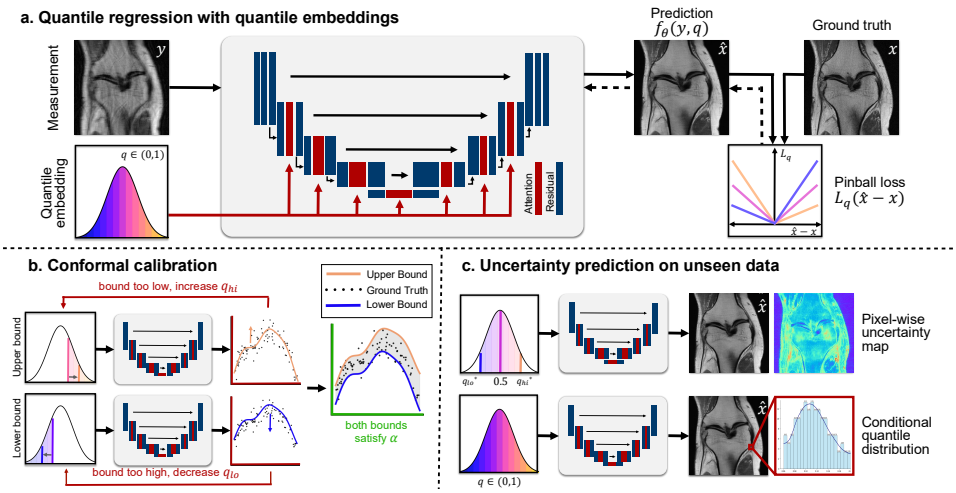


Figure 1: **QUTCC Overview. a) Quantile Regression with Quantile Embedding:** During training, a neural network with a quantile embedding predicts an image as a function of the measurement and quantile, q. The quantile embedding is randomly sampled ($q \in (0, 1)$) and the value of q determines the asymmetry of the pinball loss, enabling the model to learn a range of conditional quantiles. **b) Conformal Calibration:** During calibration, the predictive bounds ($q_{\text{lo}}, q_{\text{hi}}$) are iteratively adjusted on a held-out dataset to satisfy the desired miscoverage level α. **c) Uncertainty Prediction on Unseen Data:** At test time, the model can be queried to predict the mean image, a pixel-wise uncertainty map, or a conditional probability density function at each pixel.

These can alert users to regions of high predicted uncertainty where failures or hallucinations may be more likely, and allow them to probe the estimated distribution of possibilities.

Our method, Quantile Uncertainty Training and Conformal Calibration (QUTCC)[1], builds on previous work in conformal prediction for image regression [7, 8], but is specifically designed for high-dimensional imaging data. While prior methods use a linear, constant scaling to obtain calibrated uncertainty intervals and ignore spatial correlations in high-dimensional imaging data, our approach leverages a learned, nonlinear scaling that better captures the underlying distribution of such data. Our contributions are the following:

1. We introduce a practical simultaneous quantile regression method for high-dimensional imaging inverse problems based on a U-Net with quantile conditioning. We show how this can be paired with conformal calibration, such as risk-controlling prediction sets, to produce tighter uncertainty intervals than previous conformal methods [7, 8]. This method is fast and practical, producing tight, calibrated uncertainty-interval predictions with only two forward passes through the network.

2. We introduce a method to estimate and calibrate a pixel-wise conditional probability density function ($p(\mathbf{x}|\mathbf{y})$). By querying the same, quantile-conditioned network several times, we can obtain an estimate of the conditional probability density function. This conditional distribution does not include any Gaussian distributional assumptions, allowing us to visualize unknown pixel-wise conditional distributions.

3. We apply and evaluate our method on five imaging inverse problems. Two of these problems, synthetic Gaussian and Poisson denoising, allow us to probe and evaluate our method on a known noise distribution. The other three problems: microscopy denoising, compressive magnetic resonance imaging, and quantitative phase imaging, test our performance on challenging, real imaging data. Our method achieves state-of-the-art results on these tasks for uncertainty interval prediction, predicting smaller intervals than previous conformal methods, while being calibrated to the same coverage. Furthermore, we show examples in which our prediction of a high-uncertainty region helps us identify model hallucinations.

[1] Pronounced like cutesy, /ˈkjuːt.si/

2 Background and Notation

First, we briefly introduce the relevant background and notation. See A.1 for a comprehensive related works section. Our method is based on conformal prediction, which constructs predictive intervals with finite-sample marginal coverage guarantees at a user-specified level $(1-\alpha)$ by transforming uncertainty estimates using a held-out calibration dataset [9–11]. In general, conformal methods take heuristic or model-derived uncertainty estimates and transform them into statistically valid predictive intervals using a held-out calibration dataset. The simplicity, speed, and distribution-free nature of conformal methods have made them popular across classification [12, 13], language modeling [14, 15], robotics [16, 17], and protein design [18]. Recently, conformal prediction has been applied to image-to-image regression for pixel-wise uncertainty prediction [7, 19, 8]. However, existing methods, like Im2Im [7] and K-RCPS [8], use a fixed calibration procedure that is not spatially-adaptive. We describe existing methods below, and then outline how our method enables spatially-adaptive conformal calibration, resulting in smaller interval sizes for imaging problems.

2.1 Conformal prediction for image-valued problems

In imaging inverse problems, the goal is to predict an image, $\mathbf{x} \in \mathbb{R}^{d_x}$, from a measurement, $\mathbf{y} \in \mathbb{R}^{d_y}$, given some imaging operator/degradation and noise n, $\mathbf{y} = A(\mathbf{x}) + n$. This general framework applies to a wide range of imaging problems, including those that are ill-conditioned or underdetermined. We assume that we have access to matched input-output pairs, $\mathcal{D}_t = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_t}$, that are randomly sampled from the unknown joint distribution $p(\mathbf{x}, \mathbf{y})$. In addition, we assume access to an underlying predictor, $f(\mathbf{y})$[2], which maps a measurement, $\mathbf{y}$, to an image estimate, $\hat{\mathbf{x}}$.

On a high level, the goal of conformal prediction for image regression problems [7, 8] is to predict an uncertainty interval for each pixel in the image $\hat{\mathbf{x}}$ that contains the true pixel values with a user-specified probability. This uncertainty interval $[\hat{\mathbf{x}} - \hat{l}(\mathbf{y}), \hat{\mathbf{x}} + \hat{u}(\mathbf{y})]$ will have a width of $\hat{l}(\mathbf{y})$ in the lower direction, and a width of $\hat{u}(\mathbf{y})$ in the upper direction. But how do we practically obtain these intervals? Prior work has shown that pixel-wise quantile regression offers state-of-the-art performance and is practical to implement [7]. In quantile regression, a neural network is trained to estimate a conditional quantile instead of a mean estimate. This can be accomplished using an asymmetric pinball loss:

$$L_q(x, \hat{x}) = \begin{cases} q \cdot |x - \hat{x}| & \text{if } x - \hat{x} \geq 0 \\ (1-q) \cdot |x - \hat{x}| & \text{otherwise.} \end{cases} \tag{1}$$

where $\hat{x}$ denotes the predicted value, x represents the ground truth, and $q \in (0, 1)$ is the quantile of interest. When $q > 0.5$, the loss assigns a greater penalty to underestimations (i.e., $\hat{x} < x$), encouraging the model to predict higher values. Conversely, when $q < 0.5$, overestimations incur a larger penalty, biasing predictions downward. This asymmetry enables the model to learn conditional quantiles of the target distribution, in contrast to losses like mean squared error (MSE), which are symmetric and designed to estimate the conditional mean. To train a predictor for the 90% uncertainty interval, the lower and upper bounds can be set to estimate the 0.95 and 0.05 quantiles, respectively. Training a predictor with quantile loss results in a heuristic upper and lower interval prediction, $\mathcal{C}(\mathbf{y}) = [\hat{\mathbf{x}} - \tilde{l}(\mathbf{y}), \hat{\mathbf{x}} + \tilde{u}(\mathbf{y})]$. However, this interval does not necessarily contain the ground-truth pixel values with the desired probability. This can be remedied through conformal calibration, in which these heuristic bounds are scaled using a held-out calibration dataset, $\mathcal{D}_c = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_c}$, until they contain the desired fraction of ground truth pixels.

Im2Im uses the Risk Controlling Prediction Sets (RCPS) procedure [20], which determines a scaling factor $\hat{\lambda}$ such that the calibrated prediction intervals, $C_\lambda(\mathbf{y}) = [\hat{\mathbf{x}} - \hat{\lambda}\tilde{l}(\mathbf{y}), \hat{\mathbf{x}} + \hat{\lambda}\tilde{u}(\mathbf{y})]$, achieve the desired coverage $(1-\alpha)$ with high probability. RCPS frames calibration as a risk-control problem. The goal is to choose λ so that the population miscoverage risk $R(\lambda) = \mathbb{P}(\mathbf{X} \notin C_\lambda(\mathbf{Y}))$ stays below a user-chosen miscoverage level α. Since $R(\lambda)$ is unknown, we estimate it from the calibration set via the empirical risk:

$$\hat{R}(\lambda) = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbb{1}\{\mathbf{x}_i \notin C_\lambda(\mathbf{y}_i)\}.$$

[2]Note that we adopt the notation $\hat{\mathbf{x}} = f(\mathbf{y})$, which is commonly used in the field of inverse problems instead of $\hat{\mathbf{y}} = f(\mathbf{x})$, which is more common in the machine learning literature.

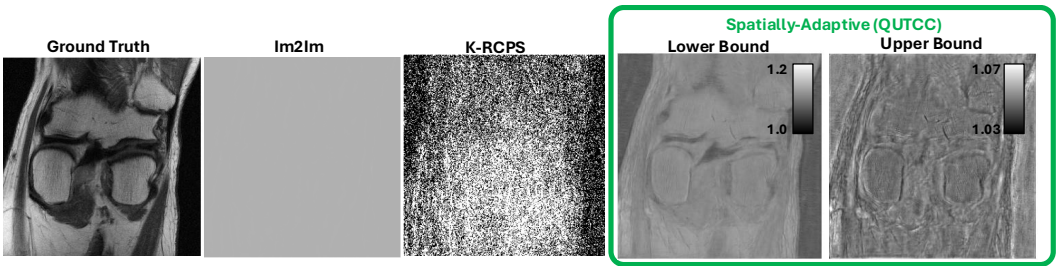


Figure 2: **Scaling parameter visualization**: Im2Im uses a single global scaling parameter, λ, K-RCPS calibrates K group-wise parameters λ_k over a fixed pixel partition; QUTCC produces an implicit pixel-wise scaling via direct quantile calibration for tighter, more adaptive bounds.

Because $\hat{R}(\lambda)$ is a finite-sample estimate of the population risk $R(\lambda)$, RCPS controls $R(\lambda)$ via a pointwise upper confidence bound $\hat{R}^+(\lambda)$ at confidence level $1-\delta$ (e.g., via the Hoeffding–Bentkus inequality). The RCPS scaling parameter is then chosen as

$$\hat{\lambda} = \inf\left\{\lambda : \hat{R}^+(\lambda') \leq \alpha \;\text{ for all } \lambda' \geq \lambda\right\},$$

where $\alpha \in (0,1)$. Since the prediction intervals are nested (larger λ yields wider, more conservative intervals), this selection guarantees $\mathbb{P}\big(R(\hat{\lambda}) \leq \alpha\big) \geq 1-\delta$. At inference time, the single calibrated constant $\hat{\lambda}$ is applied uniformly to both the predicted upper and lower bounds for every new measurement $\mathbf{y}$, producing the final prediction interval $C_{\hat{\lambda}}(\mathbf{y})$.

2.2 Non-spatially adaptive conformal calibration

While RCPS provides rigorous coverage guarantees, the calibrated parameter, $\hat{\lambda}$, acts as a uniform scalar applied identically to every pixel, regardless of local image content. This limitation was partially addressed by K-RCPS [8], which partitions the image into K groups via a fixed partition matrix (e.g. based on Otsu's method [21], or another heuristic) and assigns a separate scaling parameter $\hat{\lambda}_k$ to each group. Therefore, certain image regions can have different scaling parameters, yielding smaller uncertainty intervals compared to Im2Im at the same coverage level. However, the partition is fixed across all images and chosen independently of any individual measurement $\mathbf{y}$. An example of this global and K-partition scaling is shown in Fig. 2 for an MRI measurement. Neither method can effectively adapt its scaling to features within an image: high and low uncertainty regions within an image are calibrated the same way.

3 Methodology

What if we could predict how uncertainty varies across the image to determine continuous, spatially-adaptive scaling parameters? Rather than calibrating a global scalar parameter [12], or K-parameters [8] based on a pre-computed heuristic, we propose to calibrate the prediction interval based on a conditioning input to a neural network in which a neural network is conditioned on the upper and lower quantiles, $\hat{q}_{\text{hi}}, \hat{q}_{\text{lo}}$. This interval,

$$\mathcal{C}_{q_{\text{lo}},q_{\text{hi}}} \in [f_\theta(\mathbf{y}, \hat{q}_{\text{lo}}), f_\theta(\mathbf{y}, \hat{q}_{\text{hi}})], \tag{2}$$

is a function of a neural network, and enables continuous, *spatially-adaptive calibration* that leverages information across an image to determine how to best scale different image regions.

We obtain a heuristic uncertainty interval by setting the network's conditioning parameters to high and low quantile values, e.g., $\tilde{q}_{\text{hi}} = 0.95$, $\tilde{q}_{\text{lo}} = 0.05$. During conformal calibration with RCPS, these parameters $\hat{q}_{hi}$ and $\hat{q}_{lo}$ are updated until the desired coverage is reached on the calibration dataset (e.g. $\hat{q}_{\text{hi}} = 0.97$, $\hat{q}_{\text{lo}} = 0.06$). While we only calibrate these two input parameters, this results in a spatially-varying scaling across the entire image. This scaling is non-linear, spatially-varying, and learned from the training dataset. An example of our learned pixel-wise scaling is shown in Fig. 2(right), with full details on how we calculate this effective image-space scaling in A.4.

After calibration, the network can provide a pixel-wise uncertainty map for new, unseen data *or* can be queried to predict a conditional distribution for any pixel in the image. Our method, QUTCC, is summarized in Fig. 1. We elaborate on how we train our network through simultaneous quantile regression, how we calibrate the heuristic intervals through conformal calibration, and how we use the network to predict a conditional distribution in the sections below.

3.1 Simultaneous quantile regression

We leverage ideas from simultaneous quantile regression to train a single neural network, $f_\theta(\mathbf{y}, q)$, to predict any conditional quantile, q, of the joint distribution. The network $f_\theta(\mathbf{y}, q)$ – an attention U-Net [22, 23] with a quantile-embedding – is conditioned on the parameter q, to predict the conditional quantile. The image estimate is obtained by querying the network at $q = 0.5$, which gives us the median image:

$$\hat{\mathbf{x}} = f_\theta(\mathbf{y}, q = 0.5). \tag{3}$$

To train the neural network to predict an arbitrary conditional quantile image, we use pinball loss, (Eq. 1). At each training step, the quantile parameter, q is randomly sampled and used to both condition the network *and* as an input to the loss function. This allows the model to learn the full conditional quantile function, rather than a discrete, fixed quantile value as in prior image-to-image regression methods [7, 19]. The total loss is given by:

$$\mathcal{L}_{\text{total}}(\theta) = \mathbb{E}_{[(\mathbf{x},\mathbf{y})\sim D_{\text{t}}, q\sim\mathcal{U}(0,1)]}\Big[\mathcal{L}_q\big(\mathbf{x}, f_\theta(\mathbf{y}, q)\big)\Big], \tag{4}$$

where f_θ is a neural network with parameters θ, and $f_\theta(\mathbf{y}_i, q)$ is the output of the neural network given an input measurement, $\mathbf{y}_i$ and quantile value q. The loss is an expectation over both the data distribution $(x, y) \sim D_{\text{t}}$ and quantile levels $q \sim \mathcal{U}(0, 1)$, explicitly reflecting the uniform sampling used during training. This loss is minimized with the Adam optimizer [24] using backpropagation. The neural network weights are shared across different quantile predictions, limiting quantile crossing. The proposed network is shown in Fig. 1, and the full architecture and training details are described in Appendix A.6, A.2.

3.2 Conformal calibration

After the neural network is trained to obtain heuristic uncertainty interval predictions, we use an RCPS-style calibration to ensure statistical coverage. Rather than calibrating a scaling parameter that directly scales the predicted bounds, we find the neural network conditioning parameters (q_{lo}, q_{hi}) that achieve the target coverage, to obtain the final calibrated parameters ($\hat{q}_{lo}$, $\hat{q}_{hi}$). We calibrate the lower and upper quantile bounds independently to provide marginal pixelwise coverage across the calibration dataset. To satisfy a target total miscoverage rate of α, the calibration process allocates half of this error budget to each bound, so that the lower and upper bounds each capture violations at a marginal rate no greater than $\alpha/2$. Thus, if $\mathbf{x}_i[k]$ denotes the ground truth at pixel k of image i and $[f_\theta(\mathbf{y}_i, \hat{q}_{\text{lo}})[k], f_\theta(\mathbf{y}_i, \hat{q}_{\text{hi}})[k]]$ is the calibrated predicted interval, we enforce:

$$\frac{1}{N_c K}\sum_{i=1}^{N_c}\sum_{k=1}^{K}\mathbb{1}\{x_i[k] < f_\theta(\mathbf{y}_i, \hat{q}_{\text{lo}})[k]\} \;\le\; \frac{\alpha}{2}, \quad \frac{1}{N_c K}\sum_{i=1}^{N_c}\sum_{k=1}^{K}\mathbb{1}\{x_i[k] > f_\theta(\mathbf{y}_i, \hat{q}_{\text{hi}})[k]\} \;\le\; \frac{\alpha}{2}. \tag{5}$$

By a union bound, these two conditions together guarantee a total marginal miscoverage rate of at most α. Pseudocode for this calibration process is provided in **Algorithm 1**. At each step, we compute the marginal miscoverage from the quantile upper and lower bounds over the entire calibration dataset. If the violation rate for a bound exceeds $\alpha/2$, we relax the corresponding bound to widen the interval; otherwise, we tighten it. This process proceeds via a binary search over the quantile space until the desired coverage is reached. Conducting a binary search over the quantile space assumes that the learned quantile function is monotonic. In practice, this function is mostly monotonic with a few rare violations, which occur in background regions (See Appx Tbl. 5a). Note that we adjust the per-bound error rate, $\alpha' \leftarrow \frac{\alpha}{2} - \frac{1-\alpha/2}{N_c}$, to account for the finite calibration dataset size [9, 25]. At the end of this procedure, we can obtain a constructed interval $\mathcal{C} = [f_\theta(\mathbf{y}, \hat{q}_{\text{lo}}), f_\theta(\mathbf{y}, \hat{q}_{\text{hi}})]$ that controls the risk.

Algorithm 1 Calibrating Quantile Bounds

Require: Calibration risks $R_{\text{lower}}(\cdot)$, $R_{\text{upper}}(\cdot)$; target level α; calibration size N_c; tolerance ϵ

Compute adjusted error: $\alpha' \leftarrow \alpha - \frac{1-\alpha}{N_c}$

Set per-bound budget: $\alpha'' \leftarrow \alpha'/2$

$\hat{q}_{\text{lo}} \leftarrow$ largest $q \in [0,1]$ with $R_{\text{lower}}(q) \leq \alpha''$ (via binary search to tolerance ϵ)

$\hat{q}_{\text{hi}} \leftarrow$ smallest $q \in [0,1]$ with $R_{\text{upper}}(q) \leq \alpha''$ (via binary search to tolerance ϵ)

return $(\hat{q}_{\text{lo}},\ \hat{q}_{\text{hi}})$

3.3 Estimating the conditional distribution

Our approach offers an additional benefit: since it is trained to predict the full quantile function rather than a single fixed quantile, we can recover an estimate of the entire quantile function at each pixel. This is accomplished by querying the network $\hat{f}(\mathbf{y}, q)$ over a range of quantile levels $q \in (0,1)$, from which we construct a pixel-wise conditional PDF that serves as a standalone tool for predicting the underlying conditional distribution, $p(x|\mathbf{y})$ for each pixel in $\hat{\mathbf{x}}$ [3]. Under standard regularity conditions—namely, that the pixel-wise distribution admits a probability density function p_k, the quantile function is differentiable, and the density p_k is positive on its support—the following standard identity from probability theory holds: the PDF is inversely proportional to the rate of change of the quantile function. Applying this identity, we can estimate the conditional PDF $\hat{p}_k(x|\mathbf{y})$ for pixel intensity $x_j := \hat{f}(\mathbf{y}, q_j)$ as:

$$\hat{p}_k(x_j|\mathbf{y}) = \left(\frac{\partial \hat{f}(\mathbf{y}, q_j)}{\partial q} \right)^{-1}, \quad q_j \in \{q_1, \ldots, q_n\} \subset (0,1). \tag{6}$$

We obtain the derivative $\frac{\partial \hat{f}(\mathbf{y}, q_j)}{\partial q}$ through numerical approximation using finite differences. In practice, we query our neural network at different quantile levels $\{q_j\}_{j\in[n]}$, where j represents the quantile level being queried, to obtain predictions $x_j := \hat{f}(\mathbf{y}, q_j)$, which in turn can be used to compute the derivative via finite differences:

$$\frac{\partial \hat{f}(y, q_j)}{\partial q} \approx \frac{\hat{f}(\mathbf{y}, q_{j+1}) - \hat{f}(\mathbf{y}, q_{j-1})}{q_{j+1} - q_{j-1}}. \tag{7}$$

We finally obtain estimates of the PDF at points $x_1, \cdots, x_n$ by taking the inverse of $\frac{\partial \hat{f}(\mathbf{y}, q_j)}{\partial q}$ for all $j \in [n]$. While this provides a useful estimate of the underlying distribution, we can further enhance these estimates with statistical guarantees through conformal calibration. To construct a conformally calibrated PDF, we first specify the desired coverage levels for multiple quantile pairs. For example, for 90% coverage we initialize with quantiles 0.05 and 0.95, while for 60% coverage we use 0.20 and 0.80. These initial bounds are then calibrated using Algorithm 1 to produce bounds that guarantee the target coverage. By performing this calibration for multiple quantile levels $\{q_j\}_{j=1}^n$ and their associated coverage rates $\{1-\alpha_i\}_{j=1}^n$, we obtain a set of conformally calibrated predictions,

$$\{\hat{f}^{\text{conf}}(\mathbf{y}, \hat{q}_j)\}_{j=1}^n, \tag{8}$$

from which the pixel-wise conditional PDF is reconstructed via finite differences:

$$\hat{p}_k^{\text{conf}}\big(\hat{f}^{\text{conf}}(\mathbf{y}, \hat{q}_j)|\mathbf{y}\big) = \left(\frac{\hat{f}^{\text{conf}}(\mathbf{y}, \hat{q}_{j+1}) - \hat{f}^{\text{conf}}(\mathbf{y}, \hat{q}_{j-1})}{\hat{q}_{j+1} - \hat{q}_{j-1}} \right)^{-1}. \tag{9}$$

The resulting PDF then has the statistical guarantees of the conformal calibration procedure, providing pairwise coverage at each calibrated quantile level. Producing the pixel-wise PDF requires multiple forward passes through the network, increasing computational time. In addition, this procedure relies on numerical differentiation, which could be sensitive to the number and placement of the queried quantiles, see Fig. 13.

[3] In this section, since we predict pixel-wise PDFs, we refer to $\mathbf{x}[k]$ as x, the pixel value at index k

4 Results

For evaluation, we compare MC-Dropout [26], Deep Ensembles, [27], Im2Im[4] [7], Im2Im-Asymm[5], K-RCPS [8], and our method QUTCC on five imaging inverse problems: accelerated MRI [5], quantitative phase imaging (QPI) [28], and denoising under real-noise, synthetic Poisson, and Gaussian noise [29]. MC-Dropout and Deep Ensembles are not conformal methods and provide no formal coverage guarantees, but they yield approximate uncertainty bounds via posterior sampling. Both are computationally expensive at inference: Deep Ensembles require training n separate models and performing n forward passes per input, while MC-Dropout uses a single model but still requires n stochastic forward passes with active dropout to estimate bounds. In contrast, the conformal methods (Im2Im, Im2Im-Asymm, K-RCPS, and QUTCC) use a single model and require only one or a few forward passes. For fair comparison, all methods are conformally calibrated to the same target risk with $\alpha = 0.1$. Full training details are provided in Appendix A.6. We evaluate predicted interval lengths stratified by pixel intensity across all tasks, and show interval size as a function of noise and intensity. Additionally, we show an example of a region with high uncertainty in which a hallucination is present, and examples of a predicted pixel-wise conditional PDFs.

4.1 Uncertainty interval length and risk

We compare predicted uncertainty interval lengths stratified by pixel intensity for each task in Table 1. QUTCC consistently achieves the best or second-best interval lengths across most tasks; when it ranks second, it is typically outperformed by MC-Dropout or Deep Ensembles. However, these gains come at substantially higher computational cost: MC-Dropout incurs a 28×–47× increase in inference time (Table 9), while Deep Ensembles require roughly 10× more training time (Table 10). Mean interval lengths and risk values are reported in Table 8. QUTCC achieves the smallest mean interval length on every task among conformal methods, and remains competitive with the posterior-sampling baselines despite using a single model and only two forward passes to construct each bound.

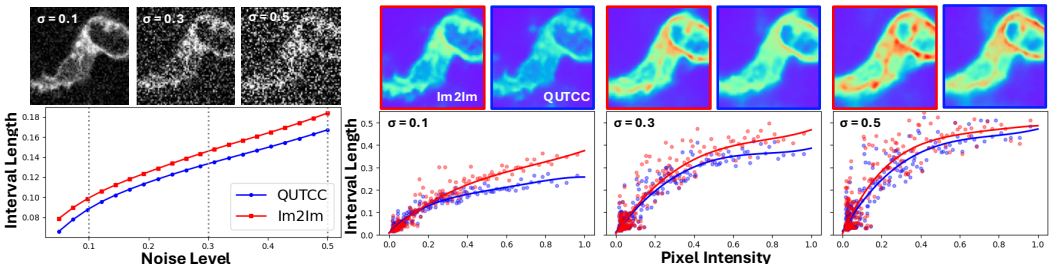


Figure 3: **QUTCC exhibits smaller uncertainty intervals in regions of high pixel intensity:** In a Gaussian denoising setting, we analyze how uncertainty interval lengths vary with pixel intensity across increasing noise levels ($\sigma = 0.1,\ 0.3,\ 0.5$). Under low-noise conditions, QUTCC exhibits narrower uncertainty intervals compared to the baseline. As noise increases to $\sigma = 0.5$, this advantage becomes less pronounced overall. However, QUTCC continues to produce shorter uncertainty intervals for high-intensity pixels (intensity > 0.8), even at higher noise levels. All confidence intervals were estimated over a set of 10 samples.

Interestingly, QUTCC's performance gains lie largely in high intensity pixels. For low-intensity background pixels, where the images contain minimal signal, QUTCC exhibits performance comparable to the baselines, reflecting high model confidence in these regions. In contrast, in high-intensity, signal-rich regions corresponding to biological structure, QUTCC produces narrower prediction intervals. In Figure 3, we compare QUTCC with Im2Im on Gaussian denoising under progressively increasing noise, reporting average uncertainty interval lengths as a function of the noise level. QUTCC consistently has smaller interval lengths at differet noise levels. Additionally, at $\sigma \in 0.1, 0.3, 0.5$, we examine a localized region containing signal and observe that QUTCC yields narrower uncertainty intervals in high-intensity pixels. This advantage becomes less pronounced as noise increases.

[4]To ensure that our performance improvements come from our uncertainty quantification technique and not network improvements, we upgrade Im2Im to use the same architecture and depth as QUTCC

[5]This is Im2Im with 2 calibrated lambdas- one for each predicted bound

Table 1: Uncertainty interval length stratified by pixel intensity. Blue entries denote the smallest uncertainty intervals within each bin, while bold black entries indicate the second smallest.

Task	Interval	Posterior-Based		Quantile-Based			
		MC-Drop.	Deep Ens.	Im2Im	Im2Im-Asym.	K-RCPS	QUTCC
Poisson	0.00–0.20	0.123 ± 0.107	$\mathbf{0.037 \pm 0.036}$	0.040 ± 0.013	0.039 ± 0.013	0.041 ± 0.011	0.035 ± 0.011
	0.20–0.40	0.371 ± 0.148	0.104 ± 0.097	0.112 ± 0.018	0.109 ± 0.018	$\mathbf{0.100 \pm 0.015}$	0.090 ± 0.015
	0.40–0.60	0.643 ± 0.213	0.240 ± 0.309	0.211 ± 0.026	0.206 ± 0.026	$\mathbf{0.181 \pm 0.021}$	0.160 ± 0.022
	0.60–0.80	0.663 ± 0.215	0.306 ± 0.438	0.244 ± 0.030	0.238 ± 0.030	$\mathbf{0.207 \pm 0.024}$	0.158 ± 0.028
	0.80–1.00	0.771 ± 0.250	0.653 ± 0.933	0.302 ± 0.042	0.294 ± 0.042	$\mathbf{0.225 \pm 0.028}$	0.187 ± 0.034
Gaussian	0.00–0.20	0.073 ± 0.070	0.050 ± 0.019	0.056 ± 0.020	0.055 ± 0.020	$\mathbf{0.054 \pm 0.019}$	0.057 ± 0.022
	0.20–0.40	0.198 ± 0.121	0.121 ± 0.027	0.135 ± 0.027	0.132 ± 0.026	$\mathbf{0.126 \pm 0.025}$	0.131 ± 0.030
	0.40–0.60	0.318 ± 0.169	0.216 ± 0.043	0.227 ± 0.041	0.222 ± 0.040	$\mathbf{0.211 \pm 0.037}$	0.206 ± 0.046
	0.60–0.80	0.318 ± 0.129	$\mathbf{0.198 \pm 0.050}$	0.239 ± 0.040	0.234 ± 0.039	0.222 ± 0.036	0.190 ± 0.048
	0.80–1.00	0.355 ± 0.126	$\mathbf{0.230 \pm 0.061}$	0.276 ± 0.045	0.270 ± 0.044	0.246 ± 0.038	0.205 ± 0.051
Real-Noise	0.00–0.20	$\mathbf{0.025 \pm 0.015}$	0.021 ± 0.006	0.030 ± 0.013	0.030 ± 0.013	0.031 ± 0.012	0.033 ± 0.012
	0.20–0.40	$\mathbf{0.104 \pm 0.035}$	0.088 ± 0.013	0.166 ± 0.017	0.165 ± 0.017	0.151 ± 0.015	0.140 ± 0.016
	0.40–0.60	$\mathbf{0.189 \pm 0.047}$	0.152 ± 0.021	0.232 ± 0.018	0.230 ± 0.018	0.208 ± 0.016	0.202 ± 0.018
	0.60–0.80	0.273 ± 0.058	0.233 ± 0.037	0.266 ± 0.017	0.263 ± 0.017	$\mathbf{0.236 \pm 0.015}$	0.248 ± 0.021
	0.80–1.00	0.342 ± 0.055	0.318 ± 0.063	0.278 ± 0.014	$\mathbf{0.274 \pm 0.014}$	0.197 ± 0.014	0.286 ± 0.025
MRI	0.00–0.20	0.132 ± 0.051	0.088 ± 0.026	$\mathbf{0.092 \pm 0.025}$	0.092 ± 0.025	0.091 ± 0.025	0.092 ± 0.026
	0.20–0.40	0.206 ± 0.073	0.127 ± 0.031	0.132 ± 0.029	0.132 ± 0.029	0.130 ± 0.028	$\mathbf{0.130 \pm 0.031}$
	0.40–0.60	0.304 ± 0.093	$\mathbf{0.132 \pm 0.029}$	0.135 ± 0.028	0.135 ± 0.028	0.133 ± 0.027	0.130 ± 0.030
	0.60–0.80	0.454 ± 0.131	$\mathbf{0.138 \pm 0.026}$	0.141 ± 0.024	0.140 ± 0.024	0.139 ± 0.024	0.133 ± 0.027
	0.80–1.00	0.680 ± 0.223	$\mathbf{0.140 \pm 0.020}$	0.140 ± 0.019	0.140 ± 0.019	0.138 ± 0.018	0.136 ± 0.022
QPI	0.00–0.20	0.055 ± 0.012	0.037 ± 0.005	0.059 ± 0.005	0.058 ± 0.004	0.061 ± 0.005	$\mathbf{0.057 \pm 0.005}$
	0.20–0.40	0.059 ± 0.144	0.042 ± 0.006	0.067 ± 0.005	0.066 ± 0.005	0.067 ± 0.006	$\mathbf{0.063 \pm 0.006}$
	0.40–0.60	0.088 ± 0.018	0.054 ± 0.006	0.085 ± 0.006	0.083 ± 0.006	0.088 ± 0.006	$\mathbf{0.081 \pm 0.006}$
	0.60–0.80	0.105 ± 0.017	0.059 ± 0.006	0.093 ± 0.006	0.091 ± 0.006	0.096 ± 0.006	$\mathbf{0.090 \pm 0.006}$
	0.80–1.00	0.116 ± 0.016	0.054 ± 0.006	0.087 ± 0.007	$\mathbf{0.085 \pm 0.007}$	0.091 ± 0.006	0.087 ± 0.007

Table 2: Interval length across MRI accelerations.

	4×	8×	12×	16×
Im2Im	0.1121	0.1371	0.1530	0.1690
QUTCC	0.1116	0.1348	0.1511	0.1643
QUTCC gain (%)	0.45	1.68	1.24	2.78

When is QUTCC more beneficial than scalar alternatives such as Im2Im and K-RCPS? QUTCC's advantage over global or partition-based scalar calibration depends on how much the true per-pixel uncertainty intervals vary across the image, i.e., the degree of heteroscedasticity in the ground-truth conditional distribution. When the ideal interval width is roughly uniform, a single global scalar, λ, (Im2Im) or a small set of partition-wise scalars, λ_k (K-RCPS), are sufficient. However, as the uncertainty becomes more spatially variable, scalar calibration becomes increasingly wasteful since it must inflate intervals everywhere to cover the most uncertain regions. For example, Poisson noise is signal-dependent with higher variance at larger intensities, so the optimal interval width differs between dark and bright regions. QUTCC has higher gains in this regime (Table 1). Additive Gaussian noise is approximately homoscedastic by construction, so a global scalar is already close to optimal and the gain is moderate. In addition, we note that this gain can grow for more ill-conditioned problems. We trained Im2Im and QUTCC across four MRI acceleration factors, Table 2. As the reconstruction problem becomes more ill-conditioned at the higher acceleration factors, QUTCC's gain increases.

4.2 Uncertainty Visualizations

Next, we visualize the predicted pixel-wise uncertainty for Im2Im and QUTCC for an undersampled MRI image (Fig. 4), which we compare against the true error. For both methods, high predicted uncertainty regions correspond to regions with high reconstruction error. Both models hallucinate a feature in the image reconstruction that is not present in the ground truth, as shown by arrows. For both methods, this hallucination is co-located with a high-uncertainty region in the uncertainty prediction. QUTCC produces a slightly narrower uncertainty interval around this region, suggesting better spatial pinpointing of high uncertainty regions. See Appendix A.3.9 for additional visualizations.

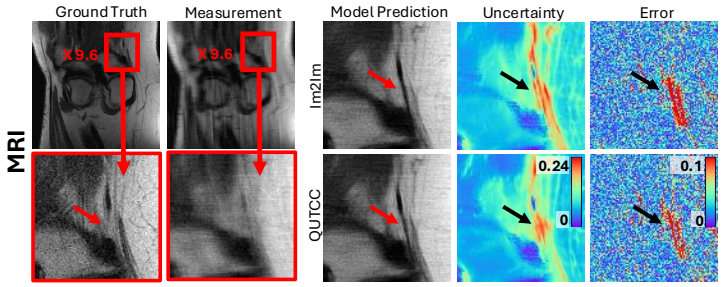


Figure 4: **Hallucination in a high uncertainty region.** The predicted image and corresponding pixel-wise uncertainty map for QUTCC and Im2Im for accelerated MRI. The uncertainty prediction is correlated with the true error (right). Both models hallucinate a feature that is not present in the ground truth (arrow). This hallucination is co-located with a high uncertainty region.

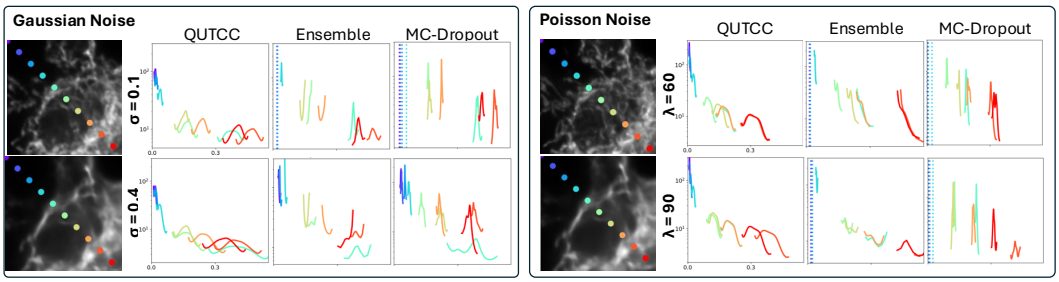


Figure 5: **QUTCC predicts diverse pixel-wise distributions under different noise distributions.** We compare the predicted conditional probability density functions (PDFs) of 10 representative pixels recovered from a measurement with different levels of Gaussian and Poisson noise across QUTCC, Ensemble, and MC-Dropout. Consistent with a Gaussian distribution, as σ increases, the conditional PDF at each pixel becomes broader, reflecting the increased variance of the distribution. In contrast, in the Poisson setting, increasing λ shifts the distribution upward, while the spread increases in proportion to the mean.

4.3 Estimating the Conditional Distribution

Finally, we demonstrate QUTCC's ability to predict a conformalized pixel-wise conditional PDF, first on a simple toy-distribution, and then on microscopy images.

First, we consider a toy problem in which the posterior is available in closed form, as described in ShuffleFlow [30]. We simulate a measurement of an MNIST '3' digit through a compressive sensing matrix with additive Gaussian noise. We fit a per-pixel Gaussian prior from all the '3' digits in the MNIST dataset. Together, these give us the exact posterior marginal PDF. We compare our method against MC-Dropout (dropout kept active at inference, $S = 50$ samples) and a 100-member Deep Ensemble trained with an identical, lighter-weight U-Net backbone on this toy problem. Table 3 shows the KL divergence and Wasserstein distance between each method's estimated conditional PDF and the true Gaussian marginal. QUTCC's predicted conditional distribution is closest to the true distribution across the methods. Moreover, it does so at a lower training time than Deep Ensemble, since it requires training only a single model. For visualizations, please see Fig. 10 in the supplement.

Table 3: Conditional distribution metrics.

	KL Divergence ↓	Wasserstein ↓
MC-Dropout	6.928 ± 3.077	0.050 ± 0.040
Deep Ensemble	0.307 ± 0.046	0.054 ± 0.035
QUTCC	$\mathbf{0.293 \pm 0.240}$	$\mathbf{0.048 \pm 0.042}$

Next, we simulate measurements with varying amounts of known Gaussian and Poisson noise to test how the predicted conditional PDF varies with measurement noise. We compare our predicted conditional PDF to those predicted via Deep Ensembles and MC-Dropout. A key advantage of QUTCC is that it constructs these distributions by querying an arbitrary number of quantiles, enabling a flexible estimate of the predictive density. In contrast, ensemble-based and MC-Dropout methods

are limited by the number of ensemble members or stochastic forward passes, which constrains the resolution of the resulting distributions. As a result, certain pixel distributions appear as dotted lines (Fig. 5), indicating point predictions due to insufficient variability in the sampled outputs. Furthermore, these approaches typically rely on implicit Bayesian assumptions about the underlying predictive distribution (e.g., Gaussian likelihoods or approximate posterior sampling), whereas QUTCC is distribution-free.

In QUTCC, the recovered PDFs reflect the underlying noise characteristics. Under Poisson noise, the predictive distributions exhibit a pronounced skew consistent with Poisson behavior, whereas under Gaussian noise they remain approximately symmetric at higher intensity values, with asymmetry introduced near lower intensities due to clipping effects. Additionally, the predicted PDFs broaden as the noise standard deviation increases, indicating that QUTCC appropriately captures changes in noise magnitude. Joint reconstruction and PDF estimation could be useful to assess the underlying noise distribution present in an image, and estimating the PDF has been useful for a number of tasks in microscopy [31]. See Appendix A.5 for additional results.

5 Limitations and Conclusion

QUTCC improves existing conformal calibration methods for high-dimensional image-to-image regression problems by leveraging a learned, non-linear, and spatially-adaptive conformal calibration procedure. QUTCC is most beneficial over existing scalar and partition-based conformal calibration methods when the per-pixel uncertainty varies across the image, such as for signal-dependent Poisson noise. Compared to K-RCPS, which performs conformal calibration over a small, discrete number of heuristic partitions for an image (Fig 2), QUTCC produces the equivalent of continuous, adaptive partitions which are a function of the local image content. This is made possible through simultaneous quantile regression with a quantile-embedded U-Net, enabling the conformal calibration procedure to be a function of a neural network that has learned to estimate not just a mean image, but any conditional quantile of the reconstruction. Our method achieves smaller uncertainty intervals on five imaging inverse problems than prior conformal methods and competitive performance with MC-Dropout and Deep Ensembles at a fraction of the time and compute cost, all while maintaining the same statistical coverage.

In addition, we demonstrated a method to estimate a conformalized conditional PDF; something previous conformal methods for image-to-image regression could not do. Finally, we show several exciting examples in which regions of high predicted uncertainty are co-located with regions of high reconstruction error and structural hallucinations. We hope this line of work will help practitioners identify harmful incorrect predictions from machine learning models for scientific and medical imaging applications.

Some limitations are that QUTCC requires paired data for both training and calibration, and offers only pixel-wise marginal coverage guarantees. Recent work attempts to relax the paired-data requirement for conformal prediction, offering a promising path forward in data-limited regimes [32, 33]. The marginal coverage guarantee is not unique to QUTCC but reflects a broader gap in conformal image-regression methods, which typically lack joint or spatially structured guarantees. Risks include the model incorrectly predicting low uncertainty in areas with potential model mistakes, such as for rare events or out of distribution data. However, as machine learning models improve, we hope that advances in uncertainty quantification can mitigate risks to enable more trustworthy imaging.

Ethical Considerations: Using machine learning methods for scientific and medical applications has inherent risk - producing realistic artifacts that are not truly present in the image can be catastrophic for discovery and medical diagnostics. As research in uncertainty quantification matures, we hope that some of these risks can be mitigated to enable more trustworthy imaging. We acknowledge and fully comply with the NeurIPS Code of Ethics.

Reproducibility: We describe the different imaging inverse tasks in more detail in A.6. In this section, we also include the model epochs that we used to evaluate in this paper. Additional explanation of QUTCC's model architecture can be found in A.2, where we describe the quantile embedding in more detail. Code to train our model and reproduce the experiments shown here can be found at the code repo.

6 Acknowledgement

The authors would like to thank Anastasios Angelopoulous for reviewing and providing feedback on our manuscript. The authors would like to thank Nora Gera, Benjamin Li, Gabriel Morete, and Adrian Hayler for their help running preliminary experiments and for valuable feedback on the statistical analysis. Research reported in this publication was supported by an Amazon Research Award, Fall 2024. We also thank the NeurIPS 2026 Financial Assistance Program for travel assistance.

References

[1] Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging, 2020.

[2] Amal Alshardan, Hany Mahgoub, Nuha Alruwais, Abdulbasit A Darem, Wafa Sulaiman Almukadi, and Abdullah Mohamed. Deep learning solutions for inverse problems in advanced biomedical image analysis on disease detection. *Scientific Reports*, 14(1):18478, 2024.

[3] George Barbastathis, Aydogan Ozcan, and Guohai Situ. On the use of deep learning for computational imaging. *Optica*, 6(8):921–943, 2019.

[4] Martin Weigert, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Akanksha Jain, Benjamin Wilhelm, Deborah Schmidt, Coleman Broaddus, Siân Culley, et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nature methods*, 15(12):1090–1097, 2018.

[5] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

[6] Berthy T Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L Bouman, and William T Freeman. Score-based diffusion models as principled priors for inverse imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10520–10531, 2023.

[7] Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pages 717–730. PMLR, 2022.

[8] Jacopo Teneggi, Matthew Tivnan, Web Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *International Conference on Machine Learning*, pages 33940–33960. PMLR, 2023.

[9] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[10] Alvaro HC Correia, Fabio Valerio Massoli, Christos Louizos, and Arash Behboodi. An information theoretic perspective on conformal prediction. *arXiv preprint arXiv:2405.02140*, 2024.

[11] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[12] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

[13] Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

[14] Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

[15] Margarida Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024.

[16] Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 8(8):5116–5123, 2023.

[17] Jordan Lekeufack, Anastasios N Angelopoulos, Andrea Bajcsy, Michael I Jordan, and Jitendra Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11668–11675. IEEE, 2024.

[18] Clara Fannjiang, Stephen Bates, Anastasios N Angelopoulos, Jennifer Listgarten, and Michael I Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

[19] Cassandra Tong Ye, Jiashu Han, Kunzan Liu, Anastasios Angelopoulos, Linda Griffith, Kristina Monakhova, and Sixian You. Learned, uncertainty-driven adaptive acquisition for photon-efficient scanning microscopy. *Optics Express*, 33(6):12269–12287, 2025.

[20] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

[21] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[23] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

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

[25] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

[26] 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. PMLR, 2016.

[27] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[28] Henry Pinkard, Cherry Liu, Fanice Nyatigo, Daniel A Fletcher, and Laura Waller. The berkeley single cell computational microscopy (bsccm) dataset. *arXiv preprint arXiv:2402.06191*, 2024.

[29] Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. A poisson-gaussian denoising dataset with real fluorescence microscopy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11710–11718, 2019.

[30] Tianao Li, Tjitske Starkenburg, Yu Sun, and Emma Alexander. Shuffleflow: Scalable posterior inference for bayesian inverse imaging. In *2026 IEEE International Conference on Computational Photography (ICCP)*, pages 1–11. IEEE, 2026.

[31] Alexander Krull, Tomáš Vičar, Mangal Prakash, Manan Lalit, and Florian Jug. Probabilistic noise2void: Unsupervised content-aware denoising. *Frontiers in Computer Science*, 2:5, 2020.

[32] Jasper Marijn Everink, Bernardin Tamo Amougou, and Marcelo Pereyra. Self-supervised conformal prediction for uncertainty quantification in imaging problems. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 108–118. Springer, 2025.

[33] Bernardin Tamo Amougou, Marcelo Pereyra, and Barbara Pascal. Self-supervised conformal prediction for uncertainty quantification in poisson imaging problems. In *2025 IEEE Statistical Signal Processing Workshop (SSP)*, pages 76–80. IEEE, 2025.

[34] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[35] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

[36] Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. 2011.

[37] Filipe Rodrigues and Francisco C Pereira. Beyond expectation: Deep joint mean and quantile regression for spatiotemporal problems. *IEEE transactions on neural networks and learning systems*, 31(12):5377–5389, 2020.

[38] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

[39] Georg Ostrovski, Will Dabney, and Rémi Munos. Autoregressive quantile networks for generative modeling. In *International Conference on Machine Learning*, pages 3936–3945. PMLR, 2018.

[40] Ran Xie, Rina Barber, and Emmanuel Candes. Boosted conformal prediction intervals. *Advances in Neural Information Processing Systems*, 37:71868–71899, 2024.

[41] Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff Schneider. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Advances in Neural Information Processing Systems*, 34:10971–10984, 2021.

[42] Elias Nehme, Omer Yair, and Tomer Michaeli. Uncertainty quantification via neural posterior principal components. *Advances in Neural Information Processing Systems*, 36:37128–37141, 2023.

[43] Omer Yair, Elias Nehme, and Tomer Michaeli. Uncertainty visualization via low-dimensional posterior projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11041–11051, 2024.

[44] Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, and Michael Elad. Principal uncertainty quantification with spatial correlation for image restoration problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3321–3333, 2023.

[45] Saumya Gupta, Yikai Zhang, Xiaoling Hu, Prateek Prasanna, and Chao Chen. Topology-aware uncertainty for image segmentation. *Advances in Neural Information Processing Systems*, 36:8186–8207, 2023.

[46] Jacopo Teneggi, J Webster Stayman, and Jeremias Sulam. Conformal risk control for semantic uncertainty quantification in computed tomography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 45–55. Springer, 2025.

[47] Jeffrey Wen, Rizwan Ahmad, and Philip Schniter. Task-driven uncertainty quantification in inverse problems via conformal prediction. In *European Conference on Computer Vision*, pages 182–199. Springer, 2024.

[48] 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.

[49] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 652–667, 2018.

[50] Elias Nehme, Rotem Mulayoff, and Tomer Michaeli. Hierarchical uncertainty exploration via feedforward posterior trees. *Advances in Neural Information Processing Systems*, 37:125142–125191, 2024.

[51] Hila Manor and Tomer Michaeli. On the posterior distribution in denoising: Application to uncertainty quantification. *arXiv preprint arXiv:2309.13598*, 2023.

[52] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[53] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

[54] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

[55] Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie. Nonparametric predictive distributions based on conformal prediction. In *Conformal and probabilistic prediction and applications*, pages 82–102. PMLR, 2017.

[56] Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alex Gammerman. Conformal predictive distributions with kernels. In *Braverman Readings in Machine Learning. Key Ideas from Inception to Current State: International Conference Commemorating the 40th Anniversary of Emmanuil Braverman's Decease, Boston, MA, USA, April 28-30, 2017, Invited Talks*, pages 103–121. Springer, 2018.

[57] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[58] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[59] Yi Zhang, Xiaoyu Shi, Dasong Li, Xiaogang Wang, Jian Wang, and Hongsheng Li. A unified conditional framework for diffusion-based image restoration. *Advances in Neural Information Processing Systems*, 36:49703–49714, 2023.

[60] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al. fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020.

A Appendix

These supplementary materials contain additional information to improve the reproducibility of the paper, as well as several additional analyses and experimental results. Code for training and analysis of QUTCC can be found at our repository. Section A.1 details related work to our method. Section A.2 details the network architecture used for QUTCC. Section A.3 provides additional comparisons against the original formulation of Im2Im-UQ without our network modifications and analysis on quantile crossing. Section A.3 also includes analysis on the size-stratified risk, more visualizations of our uncertainty predictions across the different inverse problems, and a visualization of how QUTCC produces narrower uncertainty intervals through asymmetry. Section A.4 provides a comparison of the scaling used in Im2Im, K-RCPS, and QUTCC. Sec. A.5 provides several visualizations of QUTCC's conditional distribution predictions, including an example showing how we conformalize the quantiles and the PDF prediction as a function of measurement noise. Finally, Section A.6 provides more details about the different inverse problems used throughout the paper and their associated datasets.

A.1 Related Work

A.1.1 Quantile Regression

Quantile regression is a general approach to estimate the conditional quantiles of a target distribution [34, 35]. This is often accomplished by leveraging an asymmetric loss function, called pinball loss (Fig. 1a , Eq. 1), tailored to the specified quantile level [36]. The estimated intervals obtained by quantile regression do not have formal guarantees on their own, but can be paired with conformal prediction to obtain coverage guarantees [11]. Learning quantiles during neural network training can improve predictive performance through regularization while enabling uncertainty estimation [37]. In our work, we leverage a single-network with shared parameters for simultaneous quantile prediction. We embed the quantile level as an explicit input parameter into a U-Net, which is well-suited for a variety of image regression and imaging inverse tasks. Embedding a notion of quantiles into deep learning architectures has been explored in reinforcement learning and generative modeling [38, 39], but its application to predicting uncertainty bounds for image regression tasks remains largely unexplored. We are the first to demonstrate that a single network trained for simultaneous quantile prediction can predict conformally calibrated uncertainty intervals, while mitigating quantile crossing, for imaging inverse problems.

A.1.2 Predicting Tighter or More Interpretable Bounds

Achieving smaller interval lengths without sacrificing coverage guarantees reflects increased confidence in the model's predictions. Producing smaller intervals is a common objective across uncertainty estimation methods, not just conformal prediction [40]. Several approaches have been proposed to enhance conformal prediction by targeting user-specified properties such as reduced interval length or improved conditional coverage [40, 41]; however, to date, none of these techniques have been applied to imaging tasks. On the other hand, several methods aim to improve the interpretability of uncertainty prediction for imaging tasks by moving away from per-pixel uncertainty estimates. These methods leverage principal components, posterior projected distribution, and spatial/topological relationships [42–45] to predict uncertainty in a more interpretable way. A few exciting directions include using conformal prediction for semantic uncertainty quantification in applications such as computed tomography [46], and task-driven uncertainty quantification [47], but this relies on labeled data which is not always available for imaging problems. In addition, studies in multi-hypothesis uncertainty estimation have investigated multi-head and mixture-based networks that predict multiple candidate hypotheses to explore the space of potential solutions instead of presenting a single candidate reconstruction [48–50]. However, without incorporating conformal prediction, these methods lack statistical guarantees. We present a more general method that can predict uncertainty for any imaging inverse problem, without task-specific labels, while achieving smaller uncertainty interval lengths than previous image-to-image regression methods.

A.1.3 Predicting the Conditional Distribution

In some scenarios, such as probabilistic image denoising [31, 51], predicting a full conditional probability distribution, $p(\mathbf{x}|\mathbf{y})$, could have value over predicting a confidence interval. This can be

accomplished with classic Bayesian approaches like variational inference [52, 53, 6] and Markov chain Monte Carlo [54], but these either require strong distributional assumptions or are impractically slow. There has been some theoretical work in extending conformal prediction methods to construct full nonparametric predictive distributions with guaranteed coverage properties [55, 56]. In addition, conditional distributions have been estimated using a quantile loss for problems in distributional reinforcement learning for Atari games [57, 38]. Our work extends the ideas from conformal predictive distributions to multi-dimensional imaging data and provides a practical way to apply these ideas to image regression and imaging inverse problems for the first time. As opposed to Bayesian methods, this approach is distribution-free, quick, and maintains conformal guarantees for the predicted conditional distribution.

A.2 QUTCC Model Architecture

Our method, QUTCC, is based on a U-Net backbone augmented with self-attention mechanisms, where quantile embeddings are propagated through the self-attention layers to guide the network's quantile predictions. This is inspired by the architecture of U-Nets used for diffusion models, which include a time-embedding [58, 59]. In our design, the target quantile level is embedded as a continuous scalar, analogous to the time-step embeddings in diffusion models. A core part of this architecture is the integration of self-attention layers within the U-Net, implemented as `AttentionBlock` modules. These blocks allow the model to capture global dependencies across spatial dimensions, enabling each pixel or feature location to attend to all other locations within its feature map. Specifically, attention layers are incorporated at various downsampling resolutions within both the encoder and decoder paths, as well as at the bottleneck of the U-Net.

The ability to condition the model's output on a given quantile is achieved through the quantile embedding mechanism. A quantile value between (0, 1) is chosen for each image sample at each iteration and then transformed into a high-dimensional vector representation using a positional encoding scheme, employing sinusoidal functions to create generalizable embeddings. This initial embedding is then processed by a small multi-layer perceptron, which is then used throughout to condition the network. By introducing this conditioning, the model learns to generate outputs that are responsive to the entire range of quantile levels. This mechanism works in tandem with the self-attention layers, which are specified using the `attention_resolutions` parameter.

In our experiments, QUTCC was trained using `attention_resolutions` configured as $[16, 8, 4, 2, 1]$. This means for 512 x 512 images, using the specified configuration results in attention layers within the encoder path (at resolutions 512x512, 256x256, 128x128, 64x64, and 32x32), one central attention layer in the middle block (at 8x8 resolution), and additional attention layers distributed across the decoder path (at resolutions 32x32, 64x64, 128x128, 256x256, and 512x512). The `attention_resolutions` parameter dictates the spatial scales at which these attention mechanisms are introduced. Further model configurations can be found in `models/model_config.yaml`.

A.2.1 Deep Ensemble Network Training

We train a Deep Ensemble [27] baseline with $M = 10$ independently initialized U-Net predictors. Each member f_{θ_m} uses the same backbone as the image-to-image and QUTCC models, with a two-channel output head. Each model predicts a Gaussian distribution over the output, producing per-pixel mean μ_m and variance σ_m^2 via a two-channel output head, with positivity enforced using a softplus activation and a minimum floor of 10^{-6}. We train each member for 50 epochs.

Each member is optimized independently with the Gaussian negative log-likelihood, using Adam with learning rate 10^{-4}, $\beta = (0.9, 0.99)$, and the same one-cycle schedule used for the other models. At inference, predictions from all M members are aggregated to compute the ensemble predictive mean and variance via the law of total expectation and variance. Then, we apply the same conformal calibration protocol used in all other baselines.

A.2.2 MC-Dropout Network Training

We train a single U-Net with dropout layers inserted throughout the network at rate p, following [26]. During training, dropout is active and the model is optimized with mean squared error using Adam with learning rate 10^{-4}, $\beta = (0.9, 0.99)$, and the same one-cycle learning-rate schedule as the other learned baselines. At inference, dropout remains active and each stochastic forward pass is treated as

Table 4: Image reconstruction performance. Arrows indicate the direction of better performance.

Metric	Model	MRI	QPI	Gaussian	Poisson	Real Noise
MSE (↓)	MC-Drop.	0.002 ± 0.002	0.001 ± 0.001	0.001 ± 0.001	0.0002 ± 0.0002	0.0002 ± 0.0001
	Deep Ens.	0.001 ± 0.002	0.0002 ± 0.0001	0.001 ± 0.001	0.0003 ± 0.0003	0.0001 ± 0.0001
	Im2Im-UQ	0.003 ± 0.002	0.0006 ± 0.0005	0.0007 ± 0.0006	0.0003 ± 0.0003	0.0030 ± 0.0004
	Im2Im-Deep	0.001 ± 0.002	0.0004 ± 0.0003	0.0006 ± 0.0006	0.0003 ± 0.0002	0.0004 ± 0.0002
	Im2Im-Median	0.001 ± 0.002	0.0003 ± 0.0002	0.0006 ± 0.0005	0.0003 ± 0.0002	0.0006 ± 0.0004
	QUTCC	0.001 ± 0.002	0.0004 ± 0.0003	0.0006 ± 0.0005	0.0003 ± 0.0003	0.0002 ± 0.0001
SSIM (↑)	MC-Drop.	0.684 ± 0.129	0.941 ± 0.017	0.810 ± 0.124	0.941 ± 0.033	0.957 ± 0.004
	Deep Ens.	0.717 ± 0.141	0.965 ± 0.008	0.806 ± 0.126	0.933 ± 0.038	0.967 ± 0.006
	Im2Im-UQ	0.668 ± 0.127	0.949 ± 0.017	0.852 ± 0.102	0.931 ± 0.038	0.803 ± 0.016
	Im2Im-Deep	0.707 ± 0.139	0.961 ± 0.010	0.856 ± 0.107	0.937 ± 0.035	0.959 ± 0.006
	Im2Im-Median	0.708 ± 0.139	0.961 ± 0.010	0.865 ± 0.101	0.932 ± 0.039	0.952 ± 0.013
	QUTCC	0.708 ± 0.139	0.959 ± 0.010	0.865 ± 0.102	0.941 ± 0.036	0.957 ± 0.008
PSNR (↑)	MC-Drop.	28.381 ± 2.858	30.672 ± 2.551	30.423 ± 2.871	37.546 ± 2.881	37.446 ± 1.595
	Deep Ens.	30.090 ± 3.266	38.032 ± 1.804	30.051 ± 2.834	35.965 ± 3.217	38.866 ± 1.732
	Im2Im-UQ	26.867 ± 2.923	33.135 ± 2.543	32.739 ± 3.535	36.163 ± 3.059	25.833 ± 0.734
	Im2Im-Deep	29.711 ± 3.138	34.565 ± 2.393	33.557 ± 4.018	37.062 ± 2.968	34.038 ± 1.837
	Im2Im-Median	29.744 ± 3.130	35.191 ± 2.398	33.708 ± 4.244	36.576 ± 2.824	33.388 ± 3.646
	QUTCC	29.833 ± 3.156	34.948 ± 2.436	33.660 ± 4.143	37.498 ± 3.797	37.350 ± 1.936
LPIPS (↓)	MC-Drop.	0.345 ± 0.034	0.125 ± 0.018	0.392 ± 0.122	0.191 ± 0.087	0.153 ± 0.014
	Deep Ens.	0.313 ± 0.053	0.095 ± 0.011	0.407 ± 0.129	0.192 ± 0.088	0.159 ± 0.015
	Im2Im-UQ	0.343 ± 0.033	0.153 ± 0.025	0.420 ± 0.092	0.294 ± 0.071	0.360 ± 0.033
	Im2Im-Deep	0.324 ± 0.043	0.125 ± 0.015	0.414 ± 0.103	0.299 ± 0.071	0.297 ± 0.029
	Im2Im-Median	0.322 ± 0.040	0.121 ± 0.015	0.405 ± 0.097	0.304 ± 0.071	0.324 ± 0.027
	QUTCC	0.323 ± 0.040	0.121 ± 0.015	0.408 ± 0.102	0.284 ± 0.072	0.312 ± 0.026

a sample from an approximate posterior. We draw $S = 50$ Monte Carlo samples $\{\hat{y}_s(x)\}_{s=1}^{S}$ and estimate

$$\hat{\mu}(x) = \frac{1}{S}\sum_{s=1}^{S}\hat{y}_s(x), \qquad \hat{\sigma}^2(x) = \frac{1}{S}\sum_{s=1}^{S}(\hat{y}_s(x) - \hat{\mu}(x))^2 + \tau^{-1}.$$

We then conformally calibrate the resulting bounds. This procedure yields a practical approximation to Bayesian inference under a Gaussian process prior.

A.3 Model Analysis

In Fig. 4, we restricted our comparisons to the Im2Im-Deep (which we simply refer to as Im2Im) and QUTCC models. The original Im2Im-UQ model was excluded due to its comparatively shallow architecture, resulting in decreased performance. However, for the subsequent analysis, we reintroduce Im2Im-UQ for completeness. In this section, we assess the model's mean predictive performance and the quantile crossing occurrences.

A.3.1 Prediction Performance

Does QUTCC produce tighter intervals because it is simply a better image prediction network? To investigate this, we compare the predictive performance of MC-Dropout, Deep Ensemble, Im2Im-UQ, Im2Im-Deep, Im2Im-Deep-Median, and QUTCC. Im2Im-Deep-Median is a variant of Im2Im-Deep that predicts the median rather than the mean. Since QUTCC's estimates are centered on the median, we train this median-predicting version of Im2Im-Deep to ensure a fair comparison. In Table 4, we compare model performance using standard image reconstruction metrics: MSE, SSIM, PSNR, and LPIPS. For QUTCC, the mean prediction was obtained by setting the quantile level to $q = 0.5$. The results indicate that all models achieve nearly identical performance in terms of MSE, with only minor differences observed in SSIM, PSNR, and LPIPS. These variations are not substantial enough to suggest that QUTCC provides a significantly better mean prediction. These findings suggest that QUTCC 's improved uncertainty quantification predictions are not attributed to better mean prediction performance. Rather, its ability to more effectively characterize uncertainty appears to come from the explicit learning of quantiles during training.

Table 5: Quantile crossing analysis.

Task	Crossed Pix.	Total Pix.	Ratio
MRI	2.20e1	1.64e9	1.34e−8
QPI	5.00	2.62e8	1.91e−8
Gaussian	1.10e6	3.36e9	3.29e−4
Poisson	3.49e3	3.36e9	1.04e−6
Real Noise	3.38e4	1.05e9	3.23e−5

(a) QUTCC crossings for quantiles $0.1, 0.2, \ldots, 0.9$.

Task	Im2Im	QUTCC	Total Pix.
MRI	0	0	2.05e8
QPI	0	0	3.28e7
Gaussian	3.91e3	3.30e3	4.19e8
Poisson	7.00	2.47e2	4.19e8
Real Noise	0	0	1.31e8

(b) Crossings for upper and lower bounds of Im2Im versus QUTCC.

A.3.2 Quantile-Conditioned Model vs Quantile Calibration Scheme

To what extent do the improvements come from the quantile-conditioned model itself versus the proposed quantile-space calibration procedure? To address this question, we take the 5th and 95th quantile from our quantile model and calibrate it using the same procedure as Im2Im. We observe the following mean interval lengths for Gaussian, Poisson, Real Noise, MRI, and QPI - 0.0604, 0.0482, 0.0375, 0.1071, and 0.0631. Our quantile-conditioned, quantile-calibrated model achieves smaller interval widths on Gaussian, Poisson, and Real Noise. Performance on QPI is comparable between the two approaches, while MRI shows slight improvements with Im2Im's calibration. Because the results are mixed, we conclude that both the quantile conditioning of the model and the quantile-space calibration procedure contribute value and are complementary components of our approach.

A.3.3 Quantile Crossing Performance

In section 3.2 we describe the conformal calibration step, which is dependent on the quantile function being monotonic. To ensure the validity of the predicted quantiles, specifically to avoid the issue of quantile crossing, we quantified the number of quantile crossing occurrences between $q = [0.1, 0.2, 0.3, ..., 0.9]$ in QUTCC (Tbl 5a). Quantile crossing can undermine the interpretability of our uncertainty estimates, as it contradicts the notion that quantile functions should be non-decreasing/non-overlapping. The results indicate that across all imaging tasks, the ratio of quantile crossing occurrences is minimal.

We additionally compare the quantile crossing occurrences between Im2Im-UQ and QUTCC, utilizing each model's respective bounds, which can be visualized in Tbl 5b, observing comparable performance between both models.

A.3.4 Post-Processing Monotonicity Constraint

Because QUTCC calibrates prediction sets through a binary search over quantile levels, the procedure is justified when the learned quantile function is monotonic. When this assumption holds, the calibrated bounds returned by the binary search procedure are guaranteed to satisfy the α coverage level. However, if quantile crossings do occur during calibration, the validity of the resulting calibration can be compromised. We therefore include a monotonic post-processing variant that guarantees valid intervals. For each candidate pair $(q_{\text{lo}}, q_{\text{hi}})$ evaluated during calibration, we compute the raw predictions $Q_\theta(x, q_{\text{lo}})$ and $Q_\theta(x, q_{\text{hi}})$ and form the deployed interval using the pixelwise repair:

$$L(x) = \min\{Q_\theta(x, q_{\text{lo}}), Q_\theta(x, q_{\text{hi}})\}, \qquad U(x) = \max\{Q_\theta(x, q_{\text{lo}}), Q_\theta(x, q_{\text{hi}})\}.$$

The empirical lower- and upper-tail risks used for calibration are then computed from these repaired endpoints, and the same repair is applied at test time. Because the repair is included inside the evaluation loop, the resulting bounds are guaranteed to satisfy the per-side coverage budget for the repaired intervals. We additionally monitor empirical risk monotonicity along the calibration path; if crossings or risk-direction violations indicate that binary search is unstable, calibration can fall back to a repaired grid search over candidate quantile pairs. In practice, as shown in Table 6, we find that quantile crossings occur sufficiently rarely that this refinement step has a negligible effect on the calibrated bounds, and the simpler calibration scheme without the monotonicity post-processing yields identical coverage and interval widths.

Table 6: Comparison of calibrated quantiles with and without monotonic post-processing across all tasks.

Task	Normal Calibration		Monotonic Post-Processing	
	Lower q	**Upper q**	**Lower q**	**Upper q**
Gaussian	0.0078125	0.9794922	0.0078125	0.9794922
Poisson	4.77×10^{-7}	0.9999995	4.77×10^{-7}	0.9999995
Real-Noise	0.2177734	0.9580078	0.2177734	0.9580078
MRI	0.0380859	0.9570313	0.0380859	0.9570313
QPI	0.0205078	0.9921875	0.0205078	0.9921875

Table 7: Quantile embedding ablation on Gaussian denoising.

Model	Interval Length	Total Risk
Embedding in all blocks (ours)	**0.0592**	**0.0908**
Embedding in first block only	0.0619	0.1243
Embedding in encoder blocks only	0.0645	0.1199
Embedding in middle block only	0.0695	0.0829
Embedding in decoder blocks only	0.0679	0.0895
Embedding in final block only	0.0648	0.0987
Embedding in first and final blocks	0.0676	0.1012

A.3.5 Quantile Embedding Ablation

To validate our design choice of conditioning every residual block of our model with quantile embeddings, we evaluate the impact of restricting these embeddings to specific network components. We train six variants of our model on the Gaussian denoising task, where each variant keep the same architecture and optimization but restricts conditioning to a subset of the residual blocks. We test four competing hypotheses regarding where quantile information is more critical: (i) that the model only requires embedded feature extraction in the encoder; (ii) that the model only requires embeddings in the decoder; (iii) that the model only requires semantic conditioning at the bottleneck; and (iv) that the model only requires conditioning at the highest-resolution layers (input and/or output).

We report the mean interval length and risk coverage of each calibrated model, targeting $\alpha = 0.1$, in Table 7. The baseline configuration achieves the tightest intervals while maintaining valid risk coverage. Notably, restricting conditioning to the early layers or exclusively to the high-resolution layers results in poor calibration, suggesting that the model is unable to propagate quantile-specific information to the final output. Conversely, the model calibrates to a valid α but produces overly conservative intervals when conditioning on only the bottleneck or decoder. This suggests that the network learns more efficient features and can use them more effectively when quantile information is injected at multiple scales throughout the architecture.

A.3.6 Mean Interval Length and Corresponding Risk

We show QUTCC's performance gains stratified by pixel intensity in Table 1. We demonstrate the mean performance in Table 8. For all five tasks, QUTCC predicts smaller mean interval lengths while controlling the risk amongst the conformal methods. To provide additional reassurance that the calibration behaves as expected across different target risk levels, we evaluated one of our models, MRI, with $\alpha = 0.1$, 0.2, 0.3, and 0.4. After conformal calibration, the empirical test risks were 0.10 ± 0.03, 0.20 ± 0.04, 0.30 ± 0.04, and 0.40 ± 0.04, respectively, closely matching the desired target risks.

A.3.7 Computational Cost

Tables 9 and 10 summarize the inference and training costs of each uncertainty method. Inference times are measured as the mean and standard deviation over 50 timed runs after 5 warmup runs on a single NVIDIA RTX A6000 GPU. Input sizes are $1 \times 512 \times 512$ for Poisson, Gaussian, and Real-Noise, $1 \times 320 \times 320$ for MRI, and $2 \times 128 \times 128$ for QPI. Because MC-Dropout uses 50 stochastic forward passes and Deep Ensemble uses 10 members, their inference and training costs are much higher than other methods. Im2Im-UQ is fastest because it uses the lightest network architecture, while Im2Im-Deep and QUTCC use the same deeper backbone. QUTCC is slightly

Table 8: Mean interval length and total risk after calibration.

Metric	Method	MRI	QPI	Gaussian	Poisson	Real Noise
Interval Length	MC-Dropout	0.185 ± 0.022	0.062 ± 0.003	0.084 ± 0.022	0.147 ± 0.012	0.030 ± 0.004
	Deep Ensemble	0.1055 ± 0.059	0.041 ± 0.012	0.055 ± 0.046	0.044 ± 0.070	0.025 ± 0.026
	Im2Im-Deep	0.1096 ± 0.056	0.065 ± 0.007	0.063 ± 0.049	0.047 ± 0.038	0.038 ± 0.045
	Im2Im-Asym.	0.1096 ± 0.048	0.064 ± 0.007	0.062 ± 0.028	0.046 ± 0.016	0.037 ± 0.007
	K-RCPS (k = 2)	0.1084 ± 0.047	0.067 ± 0.006	0.060 ± 0.026	0.047 ± 0.014	0.038 ± 0.006
	QUTCC	0.1083 ± 0.057	0.063 ± 0.008	0.059 ± 0.048	0.040 ± 0.029	0.036 ± 0.035
Total-Risk	MC-Dropout	0.046 ± 0.081	0.093 ± 0.089	0.092 ± 0.064	0.087 ± 0.161	0.046 ± 0.081
	Deep Ensemble	0.099 ± 0.034	0.097 ± 0.036	0.092 ± 0.034	0.073 ± 0.074	0.096 ± 0.018
	Im2Im-Deep	0.098 ± 0.048	0.100 ± 0.100	0.094 ± 0.065	0.049 ± 0.042	0.096 ± 0.040
	Im2Im-Asym.	0.099 ± 0.046	0.097 ± 0.097	0.099 ± 0.065	0.056 ± 0.045	0.098 ± 0.044
	K-RCPS (k = 2)	0.103 ± 0.050	0.092 ± 0.094	0.103 ± 0.063	0.045 ± 0.041	0.074 ± 0.036
	QUTCC	0.099 ± 0.035	0.098 ± 0.099	0.090 ± 0.046	0.093 ± 0.091	0.098 ± 0.029

Table 9: Inference time for uncertainty-bound prediction. We report mean $\pm$ standard deviation in milliseconds.

Task	**MC-Drop.**	**Deep Ens.**	**Im2Im-UQ**	**Im2Im-Deep**	**QUTCC**
Poisson	4335.34 ± 18.54	840.11 ± 0.83	16.64 ± 0.08	84.00 ± 0.14	152.40 ± 0.23
Gaussian	4346.23 ± 1.07	836.58 ± 1.17	16.52 ± 0.09	83.45 ± 0.13	151.72 ± 0.24
Real-Noise	4343.69 ± 3.15	837.94 ± 1.14	16.58 ± 0.21	83.68 ± 0.16	152.16 ± 0.35
MRI	1864.37 ± 9.18	359.14 ± 4.15	7.09 ± 0.07	35.30 ± 0.13	61.67 ± 0.17
QPI	949.32 ± 28.99	185.23 ± 10.77	2.30 ± 0.15	14.69 ± 0.65	20.05 ± 0.71

slower at inference than Im2Im-Deep because estimating the lower, mean and upper quantiles are done in three separate forward passes for QUTCC, whereas Im2Im-Deep evaluates them as a single pass. we report training costs as GPU-hour estimates for 50 epochs on the same A6000 hardware in Table 10. QUTCC has training cost comparable to Im2Im-Deep across tasks, while achieving performance similar to or exceeding Deep Ensemble, which is consistently the most expensive method because it trains 10 independent predictors.

Additionally, it takes approximately 2000 seconds ($\approx$30 minutes) to calibrate the quantile bounds of a PDF with 11 quantiles (q = 0.05, , 0.15, ..., 0.85, 0.95). The calibration time for constructing a PDF is largely independent of the number of specified quantiles. Although multiple quantile bounds at different risk levels must be calibrated to form the full distribution, each calibration can be performed as an independent job. As a result, these calibrations can be executed in parallel, and the overall calibration time is determined by the longest individual calibration rather than by the total number of quantiles. However, increasing the number of quantiles yields a more detailed representation of the underlying distribution (Fig. 13). The PDF can then be reconstructed by performing one forward pass per quantile. Once calibrated for a given dataset, the PDF quantile bounds are fixed and can be consistently applied to all images within the dataset.

A.3.8 Size-Stratified Risk

We observed the size-stratified risk of all inverse tasks between Im2Im-Deep and QUTCC (Fig. 6). To calculate size-stratified risk, the prediction intervals are first binned into different sizes, ranging from smallest to largest. Then the risk is calculated across all the bins to ensure that the model's uncertainty estimates are well-calibrated across different levels of confidence. While both Im2Im-Deep and QUTCC have bins that exceed the α, generally, most bins fall under the chosen risk.

A.3.9 Additional Visualizations and Hallucinations

We also provide visualizations of the remaining imaging tasks not included in the main results (Fig. 7). For all sample tasks, both QUTCC and Im2Im-Deep effectively highlight regions with high reconstruction error. However, QUTCC has smaller interval lengths and produces slightly more localized uncertainty estimates, which may be more useful for downstream applications. This trend is consistent across all five imaging inverse problems.

Table 10: Training cost for each uncertainty method. We report the GPU hours for 50 training epochs.

Task	MC-Drop.	Deep Ens.	Im2Im-UQ	Im2Im-Deep	QUTCC
Poisson	52.11	403.92	4.28	38.42	39.96
Gaussian	52.92	401.11	4.32	38.48	39.50
Real-Noise	53.68	533.66	6.36	55.63	53.34
MRI	56.37	411.80	5.49	41.17	41.08
QPI	87.79	590.31	11.65	56.71	59.33

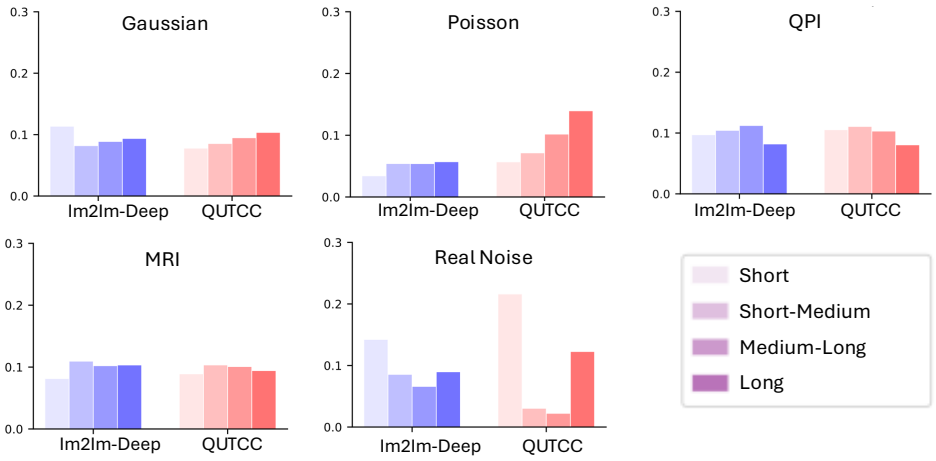


Figure 6: **Size-Stratified Risk of Im2Im-Deep vs. QUTCC**: We evaluate the size-stratified risk of Im2Im-Deep and QUTCC across all tasks. Overall, neither model exhibits a strong relationship between interval width and empirical risk, suggesting limited sensitivity to interval size. However, in the Gaussian and Poisson settings, both models display a mild trend toward improved calibration, or lower risk, for narrower prediction intervals.

Figure 8 provides additional MRI examples in which the reconstruction contains hallucinations that are not present in the ground truth. These errors are difficult to identify from the reconstructed image alone because the hallucinated features appear visually plausible. However, the QUTCC uncertainty predictions can help us locate suspicious regions with high uncertainty, which are often co-located with regions with high reconstruction error where hallucinations are more common.

A.4 Bound Asymmetry and non-uniform scaling

How is QUTCC able to predict tighter uncertainty intervals while maintaining the same level of risk? In this section, we investigate the effects of asymmetry and non-uniform pixel-wise scaling. In Fig. 9 we investigate the effect of asymmetry. QUTCC produces asymmetric predictive intervals— its upper and lower bounds are adjusted independently. In contrast, Im2Im-UQ applies a single global scaling factor λ uniformly to both bounds, which can be suboptimal in cases where only one side of the interval requires adjustment. In the red boxed region of Fig. 9, both QUTCC and Im2Im-UQ share a similar lower bound, yet QUTCC predicts a significantly tighter upper bound. Similarly, in the green boxed region, both methods align on the upper bound, but QUTCC yields a tighter lower bound. These examples highlight QUTCC's ability to adaptively adjust its interval predictions, leading to more precise interval estimates.

Next, we examine the effects of non-uniform, pixel-wise scaling. To do this, we demonstrate how scaling is applied during conformal calibration for Im2Im-UQ, Im2Im-UQ with K-RCPS calibration, and QUTCC. To ensure valid upper and lower bounds with statistical guarantees, the Im2Im calibration procedure measures miscoverage of the prediction intervals on a held-out calibration dataset. The bounds are then uniformly scaled by a constant factor λ, and the smallest λ that controls the risk is selected. In the K-RCPS procedure, instead of calibrating a single global λ, pixels in the calibration dataset are grouped into membership classes (using Otsu's method or any chosen heuristic), and a separate λ is calibrated for each class. This λ map is then applied on a dataset level. This behavior is

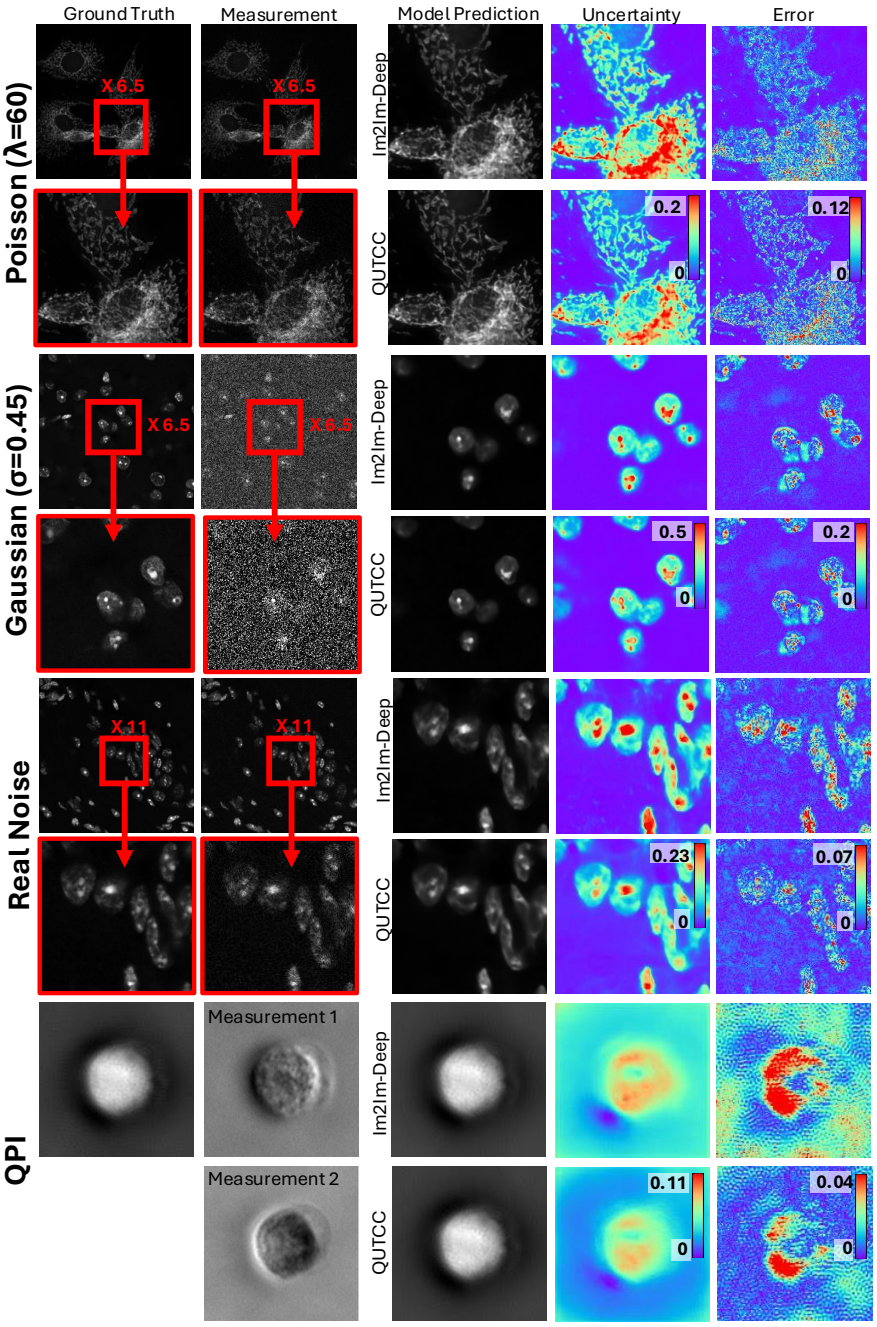


Figure 7: **Additional Uncertainty Visualizations:** We visualize both the full and zoomed-in regions of image reconstructions for QPI and denoising with Poisson, Gaussian and Real Noise. Consistent with observations presented in the results section, QUTCC produces more precise uncertainty estimates that closely align with localized regions of high reconstruction error. In contrast, Im2Im-Deep tends to highlight broader regions of uncertainty and lacks specificity, making it hard to distinguish areas of importance. This highlights QUTCC 's ability predict more informative uncertainty maps.

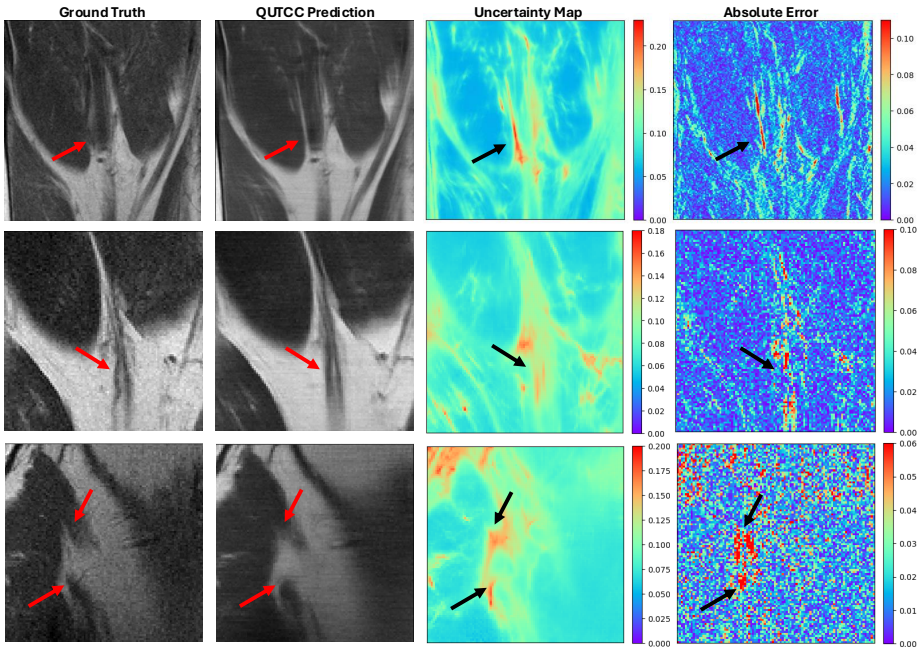


Figure 8: **QUTCC localizes MRI hallucinations.** We show additional MRI reconstruction results along with QUTCC's predicted uncertainty map and absolute reconstruction error. Red arrows mark structures that are hallucinated in the reconstruction relative to the ground truth. Black arrows show that QUTCC can produce precise uncertainty maps where high uncertainty is correlated with areas of high reconstruction error.

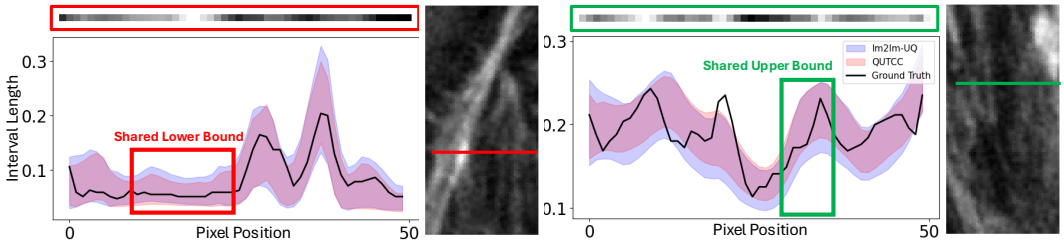


Figure 9: **QUTCC produces narrower intervals through asymmetric bounds**: We analyze the pixel-wise uncertainty bounds predicted by Im2Im-UQ and QUTCC and observe that QUTCC exhibits asymmetric behavior in its interval estimates. In Im2Im-UQ, both the upper and lower bounds are uniformly scaled by a global factor λ to satisfy coverage constraints, which limits flexibility in adapting to signal-specific uncertainty. In contrast, QUTCC learns to predict quantiles directly, enabling it to independently modulate upper and lower bounds based on signal characteristics. This results in more adaptive and efficient uncertainty intervals. For instance, in the red boxed region, QUTCC matches Im2Im-UQ's lower bound but produces a significantly tighter upper bound. Conversely, in the green boxed region, both models share an upper bound, yet QUTCC yields a tighter lower bound. Samples shown are Gaussian images with $\sigma = 0.1$.

illustrated in Fig. 2: the Im2Im λ map is constant, whereas the K-RCPS λ map contains two distinct values (for $k = 2$ classes).

In contrast, QUTCC does not explicitly calibrate λ. Instead, the calibration is a function of a neural network with input parameters, q_{hi} and q_{lo}. This results in an implicit, pixel-wise scaling factor. While we do not predict a per-pixel λ scaling, we can calculate the effective per-pixel scaling by comparing the initial and final calibrated bounds. The calibrated upper and lower bounds, $f_\theta(\mathbf{y}, \hat{q}_{\text{up}}), f_\theta(\mathbf{y}, \hat{q}_{\text{lo}})$, are related to the initial upper and lower bounds, $f_\theta(\mathbf{y}, \tilde{q}_{\text{up}}), f_\theta(\mathbf{y}, \tilde{q}_{\text{lo}})$, and the effective scaling, λ_{hi},

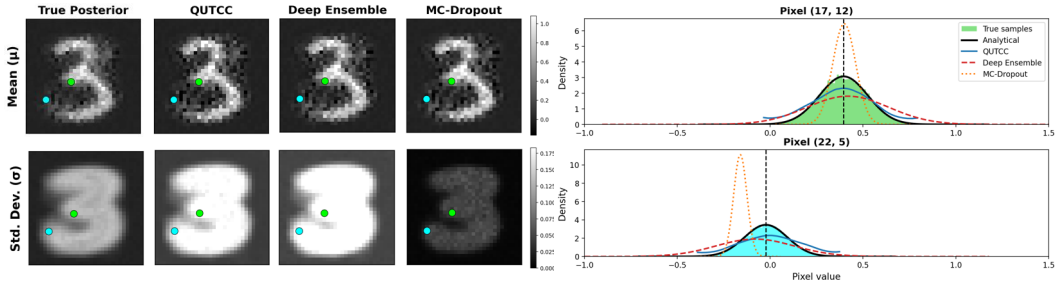


Figure 10: **QUTCC predicts more accurate pixel-wise PDFs of a known toy-distribution than compared to MC-Dropout and Deep Ensemble**

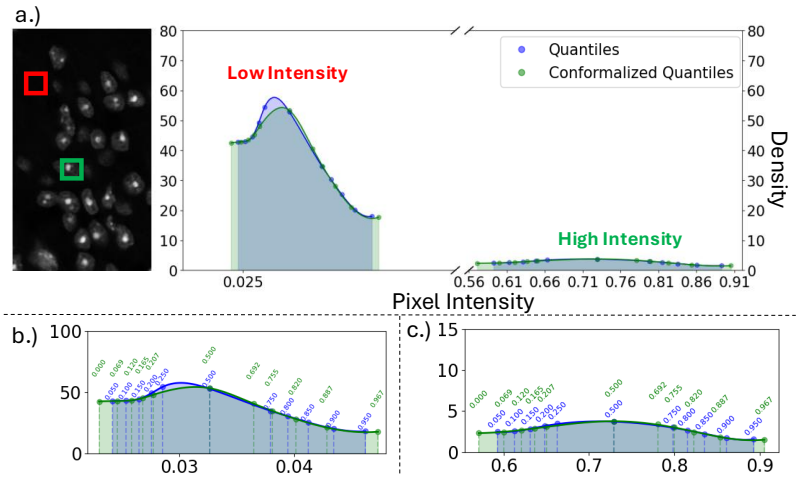


Figure 11: **QUTCC predicts different pixel-wise PDFs based on different signal intensity a)** Comparison of pixel-wise PDFs for representative low-intensity and high-intensity pixels in a Gaussian measurement ($\sigma = 0.4$). **b)** Detailed view of the low-intensity pixel PDF, exhibiting a narrow, high-density distribution concentrated around few intensity values, indicating low predictive uncertainty. **c)** Detailed view of the high-intensity pixel PDF, showing a broader, lower-density distribution with increased spread, reflecting higher predictive uncertainty in bright image regions. The blue lines show the uncalibrated model, while the green lines show the conformalized quantiles after calibration.

λ_{lo}, by the equations:

$$f_\theta(\mathbf{y}, \hat{q}_{\text{lo}}) = \hat{\mathbf{x}} - \lambda_{\text{lo}}(f_\theta(\mathbf{y}, \tilde{q}_{\text{lo}}) - \hat{\mathbf{x}}), \quad f_\theta(\mathbf{y}, \hat{q}_{\text{hi}}) = \hat{\mathbf{x}} + \lambda_{\text{hi}}(f_\theta(\mathbf{y}, \tilde{q}_{\text{hi}}) - \hat{\mathbf{x}}). \tag{10}$$

Rearranging these equations to solve for the effective λ as a function of the initial and final predicted bounds, we obtain:

$$\lambda_{\text{hi}}(\mathbf{y})[k] = \frac{f_\theta(\mathbf{y}, \hat{q}_{\text{hi}})[k] - f_\theta(\mathbf{y}, q = 0.5)[k]}{f_\theta(\mathbf{y}, \tilde{q}_{\text{hi}})[k] - f_\theta(\mathbf{y}, q = 0.5)[k]}, \quad \lambda_{\text{lo}}(\mathbf{y})[k] = \frac{f_\theta(\mathbf{y}, q = 0.5)[k] - f_\theta(\mathbf{y}, \hat{q}_{\text{lo}})[k]}{f_\theta(\mathbf{y}, \tilde{q}_{\text{lo}}[k]) - f_\theta(\mathbf{y}, q = 0.5)[k]}, \tag{11}$$

resulting in a per-pixel effective scaling that is adapted to each measurement. The corresponding upper and lower effective scaling maps, λ_{hi} and λ_{lo}, for QUTCC are shown in Fig. 2. Here we can see that QUTCC automatically learns a pixel-wise scaling that depends on the features within the image, unlike Im2Im, which uses a constant, and K-RCPS, which uses a membership map based on a heuristic measure (e.g., intensity) on the dataset level. This per-pixel adaptivity enables QUTCC to produce more tailored uncertainty estimates, leading to tighter uncertainty intervals.

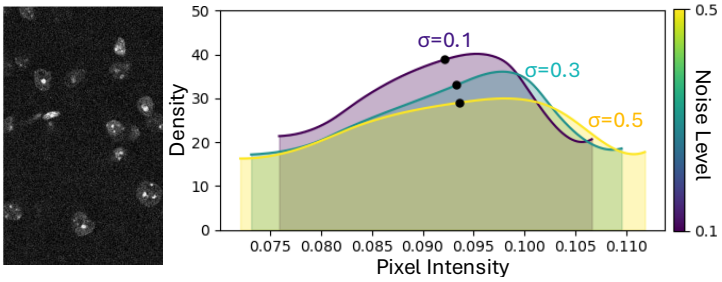


Figure 12: **PDF broadens as noise increases:** We observe the PDF of a single pixel under varying noise levels. At $\sigma = 0.1$, the noise is low, and the PDF is compact. As the noise increases to $\sigma = 0.3$ and $\sigma = 0.5$, the PDF gradually broadens, while the mean prediction value remains relatively unchanged. This broadening occurs due to the increased uncertainty introduced by higher noise levels. QUTCC successfully predicts this increased uncertainty as the noise increases.

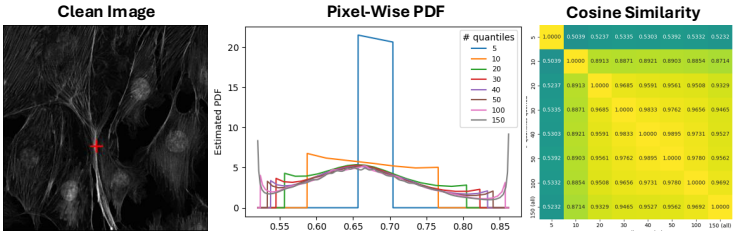


Figure 13: **Pixelwise PDF as a function of queried quantiles**: We demonstrate the changes in PDF shape as the number of queried quantiles increase. We compare PDF similarity in a cosine similarity map.

A.5 Additional PDF results

In this section, we present additional results highlighting QUTCC's ability to estimate a conditional probability density function. In Fig. 10, we show the visual results of MC-dropout, Deep Ensemble, and QUTCC on a toy-distribution as described in the main text. In Fig. 11a, we show several of the predicted PDFs for image denoising with Gaussian noise with $\sigma = 0.4$. Detailed views of the corresponding pixel-wise PDFs are presented in Fig. 11b for the low-intensity pixel and Fig. 11c for the high-intensity pixel. Each graph displays two PDFs: the blue PDF represents the quantile predictions from the uncalibrated model, while the green PDF represents the conformally calibrated quantiles that provide finite-sample statistical coverage guarantees. The conformal calibration procedure adjusts the quantile levels to ensure valid coverage properties. For instance, while the 25th and 75th percentiles theoretically provide 50% coverage, conformal calibration determines that the 20.7th and 69.2nd percentiles are required to achieve statistically guaranteed 50% coverage for this specific dataset and model. Additionally, Fig. 12 illustrates the evolution of the pixel-wise PDF for a fixed pixel coordinate under varying Gaussian noise levels $\sigma \in \{0.1, 0.3, 0.5\}$. As the noise standard deviation increases, the PDF widens. This widening corresponds to increased epistemic uncertainty in the model's predictions, as higher noise levels reduce the information content available for accurate pixel intensity estimation. This then directly widens prediction intervals to maintain the desired coverage guarantees.

A.5.1 Impact of Quantile number on estimated PDF

We investigate how the number of quantiles influences the predicted PDF. The PDF is estimated from the numerical derivative of the quantile function. This estimate depends on the spacing between adjacent quantiles. Figure 13 illustrates this effect by showing PDFs reconstructed using progressively more quantile samples. The overall shape stays relatively similar as the number of quantile queries increases. This is quantitatively confirmed by measuring cosine similarity between the queried PDFs.

A.6 Experiment Details

In this section, we provide additional experimental details about our training process and datasets used for our five different tasks: denoising (real, Gaussian, Poisson), MRI, and QPI. For Im2Im-Deep and QUTCC, we conducted a model selection sweep to identify the epoch that yielded the narrowest uncertainty intervals while satisfying the target risk level ($\alpha = 0.1$). The selected epochs are reported below for each task. For all models, an initial learning rate of 1e-4 and weight decay of 0 was used. Batch size was adjusted depending on the task, with 4 used for denoising tasks, 12 for MRI, and 72 for QPI for U-Net backbones and 32, 512, and 16 (respectively) for ResNet-18 backbones. Images were normalized before training.

A.6.1 Real Noise Task

For the real noise task, we used the Fluorescence Microscopy Denoising (FMD) dataset [29], which contains experimentally acquired fluorescence microscopy images spanning 12 wide-field, confocal, and two-photon modalities. The model was trained on 10,000 images, with 500 confocal mouse images used for calibration and an additional 500 for validation. The epochs used for Im2Im-Deep and QUTCC are 5 and 10, respectively.

A.6.2 Gaussian and Poisson Noise Task

For both the Gaussian and Poisson tasks, we synthetically introduced varying levels of noise to the FMD ground truth images. The dataset was split into 180 ground truth images for training, 40 for calibration, and 20 for validation. The pseudocode for generating gaussian and poisson noise are shown below in Algorithm 2 and Algorithm 3. For Gaussian noise, we set $\sigma_{\max} = 0.5$. For Poisson

Algorithm 2 Add Gaussian Noise

Require: Clean image $\mathbf{x}$, max noise level σ_{max}
Ensure: Noisy image $\mathbf{y}$
1: Sample noise std: $\sigma \sim \mathcal{U}(0, \sigma_{\text{max}})$
2: Sample Gaussian noise: $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2)$
3: Add noise: $\mathbf{y} \leftarrow \mathbf{x} + \boldsymbol{\eta}$
4: Clamp: $\mathbf{y} \leftarrow \text{Clamp}(\mathbf{y}, 0, 1)$
5: **return** $\mathbf{y}$

Algorithm 3 Add Poisson Noise

Require: Clean image $\mathbf{x}$, min/max noise levels $\sigma_{\text{min}}, \sigma_{\text{max}}$
Ensure: Noisy image $\mathbf{y}$
1: Sample scale: $\lambda \sim \mathcal{U}(\sigma_{\text{min}}, \sigma_{\text{max}})$
2: Scale image: $\mathbf{x}_{\text{scaled}} \leftarrow \lambda \cdot \mathbf{x}$
3: Sample noise: $\boldsymbol{\eta} \sim \text{Poisson}(\mathbf{x}_{\text{scaled}})$
4: Clamp: $\mathbf{y} \leftarrow \text{Clamp}(\boldsymbol{\eta}, 0, 1)$
5: **return** $\mathbf{y}$

noise, the noise level range $(\lambda_{\min}, \lambda_{\max})$ was set to $(50, 100)$. In each iteration, with a batch size of 16, we generated 25 random noise levels uniformly sampled between the specified minimum and maximum values. The number of training epochs for Im2Im-Deep and QUTCC with Gaussian noise were 15 and 20, respectively. For Poisson noise, Im2Im-Deep was trained for 10 epochs, while QUTCC was trained for 35 epochs.

A.6.3 Magnetic Resonance Imaging (MRI) Task

Data used for the MRI task was obtained from the NYU fastMRI initiative database (fastmri.med.nyu.edu) [60, 5]. The primary goal of fastMRI is to test whether machine learning can aid in the reconstruction of medical MRI images. To train our models, we split the fastMRI dataset into 700 volumes for training, 200 volumes for calibration, and 200 for validation. It is important to note that a single volume contains multiple MRI images. The epochs used for Im2Im-Deep and QUTCC are 25 and 40, respectively.

To simulate the forward model in MRI, we start with a fully-sampled 3D volume composed of multiple 2D image slices. Each 2D image slice is transformed into its frequency domain representation using the 2D Fourier Transform, producing its k-space data. To simulate undersampled acquisition, we apply a 4× undersampling mask to the k-space. The resulting undersampled k-space is then transformed back into the image domain using the inverse Fourier Transform, yielding an aliased or artifact-corrupted image that serves as the input for the models.

A.6.4 Quantitative Phase Imaging (QPI) Task

Data used for the QPI task were obtained from the Berkeley Single Cell Computational Microscopy (BSCCM) dataset [28]. The BSCCM dataset contains image samples of individual white blood cells that have been captured with several illumination patterns on an LED array microscope. For this task, we used 289,059 images for training, 82,588 images for calibration, and 41,294 images for validation. The number of training epochs for Im2Im-Deep and QUTCC were 15 and 20, respectively.

In QPI, the goal is to image the structure of transparent cells by recovering the phase delay of the light that passes through the cells. Several intensity-only images are used to computationally estimate the phase of the light, since this cannot be measured directly. The input measurement $\mathbf{y}$ is the concatenation of two cell intensity images acquired at different illumination angles. The corresponding ground truth is the quantitative phase image recovered from four or more illumination angles. The model is trained to map these two intensity images into a phase image.

A.7 LLM Usage

LLMs were used to edit and correct grammar during writing, but were not involved in the research ideation process.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: In the abstract we claim that our model produces tighter uncertainty intervals than prior methods, predicts plausible conditional distributions, and that high regions of uncertainty can be co-located with regions of hallucination. We support this claim in **Table 1**, **Figure 4**, and **Figure 5** of the main paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide limitations of our work in **Section 5: Limitations and Conclusions** of the main paper. We point out that our method is limited by a need for paired data and that the calibration procedure only provides pixel-wise marginal coverage guarantees.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification:

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We fully disclose all the information needed to reproduce the main experimental results of the paper.

   - **Training/Testing Image Data:** We use only open-source data and these have been cited accordingly in the paper. We also elaborate more on experiment details in **Sec. A.6**.
   - **Model Architecture:** We describe our model architecture in the **Methodology Section**. We elaborate more in **Sec. A.2**.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide an anonymous open source repository in the appendix: textbfhttps://anonymous.4open.science/r/QUTCC-UQ-CB0F. Additionally, we cite all datasets used in **Sec. A.6: Experiment Details**.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We fully disclose all the information needed to reproduce the main experimental results of the paper. We also elaborate more on experiment details in **Sec. A.6**.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, when reporting mean uncertainty length, MSE, SSIM, PSNR, and LPIPS, all values are reported with standard deviation values.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the training cost for each uncertainty method on each image tasks in **Table 8** of the Appendix. We also disclose computational cost in **Sec. A.3.6**.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not have human subjects or participants. We use publicly available datasets and respect the terms of the datasets. We do not anticipate harmful consequences from this research.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Both positive and negative societal impacts are discussed in the Conclusion.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: [N/A]

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: [N/A]

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: [N/A]