Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging

Anastasios N. Angelopoulos [*1] **Amit Kohli** [*1] **Stephen Bates** [1] **Michael I. Jordan** [1] **Jitendra Malik** [1]
Thayer Alshaabi [2] **Srigokul Upadhyayula** [2 3] **Yaniv Romano** [4]

Abstract

Image-to-image regression is an important learning task, used frequently in biological imaging. Current algorithms, however, do not generally offer statistical guarantees that protect against a model's mistakes and hallucinations. To address this, we develop uncertainty quantification techniques with rigorous statistical guarantees for image-to-image regression problems. In particular, we show how to derive uncertainty intervals around each pixel that are guaranteed to contain the true value with a user-specified confidence probability. Our methods work in conjunction with any base machine learning model, such as a neural network, and endow it with formal mathematical guarantees, regardless of the true unknown data distribution or choice of model. Furthermore, they are simple to implement and computationally inexpensive. We evaluate our procedure on three image-to-image regression tasks: quantitative phase microscopy, accelerated magnetic resonance imaging, and super-resolution transmission electron microscopy of a *Drosophila melanogaster* brain, and provide accompanying open source code.

1. Introduction

The deployment of image-to-image regression in scientific imaging has generated enormous excitement, promising a future where the resolution of an imaging system can be improved algorithmically (Weigert et al., 2018). For example, research developments in machine learning have accelerated MRI scans by an order of magnitude (Zbontar et al., 2018b). But, to this day, there remains an elephant in the room, obstructing the deployment of these systems: *how can we know when the model has produced an incorrect prediction?*

In most cases, we cannot. Indeed, the expressive power of modern machine learning is also its torment. Deep learning models have revolutionized predictive accuracy, but they fail in silent, unknown, and even unknowable ways. For scientific imaging settings, where learning will be used for inference and discovery, we need ways to understand when and how a model's predictions might be wrong. Nonetheless, image-to-image regression algorithms, such as those for denoising and super-resolution, are normally deployed without any notion of statistical reliability. The scientist is therefore left worrying that their new discovery is simply the model's hallucination. The purpose of this paper is to introduce a technique which rigorously quantifies the uncertainty in an image-valued point prediction, thereby alerting the scientist of potential hallucinations (see Figure 1).

We will develop a method for endowing any image-to-image regression model with per-pixel *uncertainty intervals*. At a particular pixel, an uncertainty interval is a range of values guaranteed to contain the true value of that pixel with high probability. Our contributions are the following:

1. We introduce *distribution-free* uncertainty quantification to image-to-image regression; this means the uncertainty intervals will have a rigorous guarantee for any image dataset and any regression model, regardless of the number of data points used to construct the interval.
2. We introduce and evaluate several practical algorithms for constructing these sets, including *pixelwise quantile regression*, an extension of quantile regression (Koenker & Bassett Jr, 1978) to this setting. In experiments, pixelwise quantile regression consistently leads to the best performance of any uncertainty quantification algorithm, often by a large amount.
3. We apply our methods to three challenging imag-

[*]Equal contribution [1]Department of Electrical Engineering and Computer Science, University of California, Berkeley [2]Advanced Bioimaging Center, Department of Molecular and Cell Biology, University of California, Berkeley [3]Chan Zuckerberg Biohub, San Francisco, CA [4]Departments of Electrical and Computer Engineering and of Computer Science, Technion - Israel Institute of Technology. Correspondence to: Anastasios N. Angelopoulos <angelopoulos@berkeley.edu>, Amit Kohli <apkohli@berkeley.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

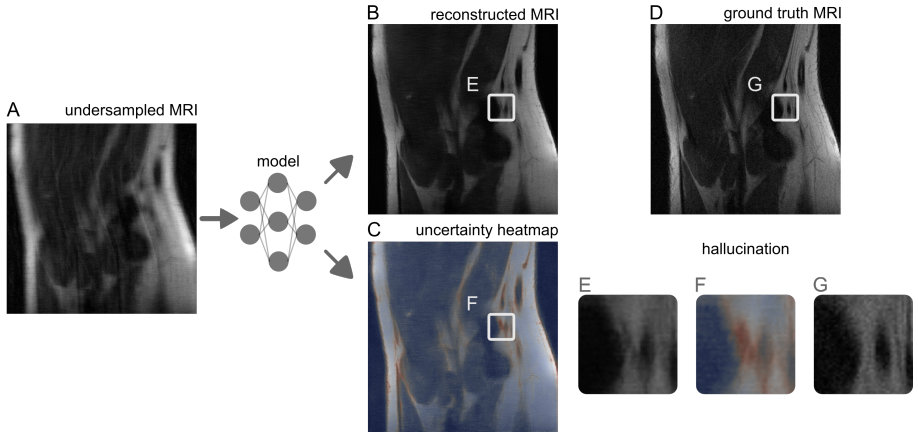

Figure 1. **An algorithmic MRI reconstruction with uncertainty.** A rapidly acquired but undersampled MR image of a knee (A) is fed into a model that predicts a sharp reconstruction (B) with calibrated uncertainty (C). In (C), red means high uncertainty and blue means low uncertainty. Wherever the reconstruction contains hallucinations, the uncertainty is high; see the hallucination in the image patch (E), which has high uncertainty in (F), and does not exist in the ground truth (G). For experimental details, see Section 3.4.

ing problems: quantitative phase microscopy, accelerated magnetic resonance imaging (MRI), and super-resolution transmission electron microscopy of a *Drosophila melanogaster* brain. Our accompanying codebase allows for easy application of these methods to any imaging problem, and the exact reproduction of the aforementioned examples. The proposed calibrated pixelwise quantile regression approach offers state-of-the-art results on these tasks, in the sense that its uncertainty intervals are smaller than those from other methods.

1.1. Notation and Goal

The inputs X and outputs Y are both images in $\mathcal{X} = [0,1]^{M \times N}$ (for simplicity of notation, we discuss the case where X and Y are the same size). We also assume access to an *underlying predictor* $\hat{f}(X)$ mapping from X to a point prediction of Y. The reader can imagine X to be a downsampled version of Y, and $\hat{f}(X) \in \mathbb{R}^{M \times N}$ to be the output of a neural network trained to upsample X and reconstruct Y (this is the *super-resolution* task).

Our task is to create uncertainty intervals around each pixel of the predicted image $\hat{f}(X)$ that contain the true pixel values with a user-specified probability. Formally, we will construct the following interval-valued function for each pixel,

$$\mathcal{T}(X)_{(m,n)} = \Big[\hat{f}(X)_{(m,n)} - \hat{l}(X)_{(m,n)}, \; \hat{f}(X)_{(m,n)} + \hat{u}(X)_{(m,n)}\Big], \tag{1}$$

which takes an image and outputs the uncertainty interval for each pixel (m,n). Notice that the intervals always include the prediction $\hat{f}(X)$, and have width $\hat{l}(X)$ in the lower direction and $\hat{u}(x)$ in the upper direction. Intuitively, a large value in $\hat{u}(X)$ indicates a pixel that could have a much higher value than the prediction (undershooting). Likewise, a large pixel value in $\hat{l}(X)$ indicates a pixel that could have a much lower value than the prediction (overshooting). We will form the uncertainty intervals by using a held-out set of calibration data, $\{(X_i, Y_i)\}_{i=1}^n$, to assess the model's performance. The uncertainty intervals will be statistically valid in the following sense. The user selects a risk level $\alpha \in (0,1)$, and an error level $\delta \in (0,1)$, such as $\alpha = \delta = 0.1$. Then, we construct intervals that contain at least $1-\alpha$ of the ground-truth pixel values with probability $1-\delta$. That is, with probability at least $1-\delta$,

$$\mathbb{E}\left[\frac{1}{MN}\Big|\{(m,n) : Y^{\text{test}}_{(m,n)} \in \mathcal{T}(X^{\text{test}})_{(m,n)}\}\Big|\right] \geq 1-\alpha,$$

where $X^{\text{test}}, Y^{\text{test}}$ is a fresh test point from the same distribution as the calibration data.

In the next section, we will describe in detail the algorithm for generating $\hat{l}(X)$ and $\hat{u}(X)$ as well as its statistical properties. Importantly, this algorithm is modular, allowing the user to use the most complex, cutting-edge methods for learning $\hat{f}$ (i.e., the best neural network methods), all the while having uncertainty intervals that reliably communicate the quality of the predictions.

2. Methods

We now formally describe the method for constructing uncertainty intervals. Each pixel in the image will get its own uncertainty interval, as in (1), that is statistically guaranteed to contain the true value with high probability.

The procedure that yields these intervals, visualized in Figure 2, has two phases. First, we train a model to output a heuristic notion of uncertainty. In practice, this amounts to training a machine learning system to output a point prediction $\hat{f}$, a heuristic lower interval length $\tilde{l}$, and a heuristic upper interval length $\tilde{u}$ using any method, such as a neural network. In Section 2.1 we introduce and benchmark four

possible methods of learning these heuristics. The uncalibrated intervals $(\hat{f}(X) - \tilde{l}(X), \hat{f}(X) + \tilde{u}(X))$ are heuristic in the sense that they do not contain the ground truth with the desired probability—we made no assumptions about the algorithm used to train $\tilde{l}$ and $\tilde{u}$. To remedy this, in the second phase we calibrate the heuristic notions of uncertainty by scaling them until they contain the right fraction of the ground truth pixels. That is, we multiply the upper and lower lengths by a value $\hat{\lambda}$ that is chosen using the procedure that we will describe in Section 2.2. The final intervals are exactly those in (1), with the upper and lower widths

$$\hat{l}(x) = \hat{\lambda}\tilde{l}(x) \quad \text{and} \quad \hat{u}(x) = \hat{\lambda}\tilde{u}(x).$$

Algorithm 1 summarizes this process.

Following the above strategy will give us uncertainty intervals that satisfy the desired statistical guarantee from Section 1.1. We call a set of these rigorous uncertainty intervals—one for each pixel in an image—an image-valued *Risk-Controlling Prediction Set*.

Definition 2.1 (Risk-Controlling Prediction Set (RCPS), modified from (Bates et al., 2021))**.** We call a random set-valued function $\mathcal{T} : \mathcal{X} \to \left(2^{[0,1]}\right)^{M \times N}$ an *(α,δ)-Risk-Controlling Prediction Set* if

$$\mathbb{P}\left(\mathbb{E}[L(\mathcal{T}(X), Y)] > \alpha\right) \leq \delta, \tag{2}$$

where

$$L(\mathcal{T}(X), Y) = 1 - \frac{\left|\{(m,n) : Y_{(m,n)} \in \mathcal{T}(X)_{(m,n)}\}\right|}{MN}.$$

Remark 2.2. The inner expectation in (2) is over a new test point, (X, Y). The outer probability is over the calibration data, $\{(X_i, Y_i)\}_{i=1}^n$. In other words, $\mathcal{T}$ is constructed based on the calibration data, which makes it a random function. We will only fail to control the risk if we are unlucky with the sample of calibration data, with probability δ.

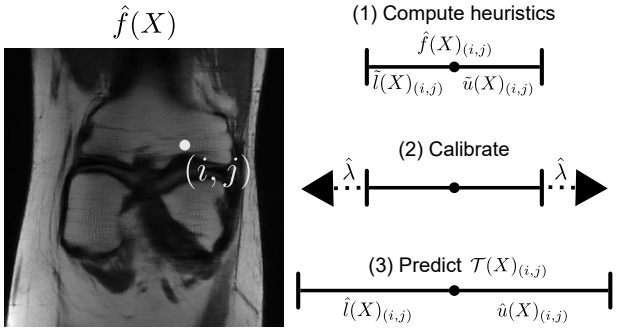

Figure 2. **An explanation of image-valued risk-controlling prediction sets.** We visualize the process of constructing an uncertainty interval for a single pixel (i, j) of the model's prediction $\hat{f}(X)$. In the first step, we compute the heuristic upper and lower interval lengths. Second, we choose $\hat{\lambda}$ via the RCPS calibration procedure in Section 2.2. Finally, we form the risk-controlling prediction set $\mathcal{T}_{\hat{\lambda}}(X)_{(i,j)}$ as in (1).

Algorithm 1 Generating Image-Valued RCPS

1: Train model that outputs point prediction $\hat{f}$ and heuristic lower and upper interval lengths $\tilde{l}$ and $\tilde{u}$.
2: Compute the calibrated parameter $\hat{\lambda}$ using the calibration data and Algorithm 2.
3: Construct $\mathcal{T}$ as in (1).
4: For a new image X, output the risk-controlling prediction set $\mathcal{T}(X)$.

Parsing the above equation, we define a level α, which tells us what fraction of pixels in the image we allow to fall outside of the intervals. If we set $\alpha = 0.1$, for example, it means no more than 10% of the true pixel values will lie outside of $\mathcal{T}$ except with probability δ.

Having laid out the goal and general algorithm, we now discuss how to train the model to output heuristic notions of uncertainty for eventual calibration.

2.1. Picking a Heuristic Notion of Uncertainty

The selections of $\tilde{l}$ and $\tilde{u}$ will ultimately determine the properties of the prediction sets, such as their size and shape. We will learn these heuristics from the same training dataset used to train $\hat{f}$. Here, we develop four different heuristic notions of uncertainty, which we will evaluate and compare in later experiments (Section 3). These heuristics are (1) regression to the magnitude of the residual, (2) parameterizing each pixel as a Gaussian and reporting its standard deviation, (3) outputting a softmax distribution at each pixel, and (4) pixelwise quantile regression.

Although each of these methods is trained to predict some form of uncertainty, they may not do it well—hence the need for calibration via Algorithm 2 after training. Each heuristic requires the use of a different loss function when training the neural network via gradient descent. The remainder of this subsection describes each loss function precisely. For notational simplicity, we omit subscripts and sums indexing different pixels; in the experiments, we train our models by averaging the loss function applied to each pixel separately.

2.1.1. Magnitude of the Residual

In this flavor of uncertainty quantification, we set $\tilde{u} = \tilde{l}$, referring to both the upper and lower interval lengths as $\tilde{u}$ (the letter 'u' is a mnemonic for the 'uncertainty' of the model). We then optimize $\tilde{u}$ for the following loss function:

$$\mathcal{L}(x, y) = \left(\tilde{u}(x) - |\hat{f}(x) - y|\right)^2.$$

The loss function encourages each pixel of $\tilde{u}$ to be equal to the model's error at that pixel. Notice that $\mathcal{L}(x, y) = 0$ in the ideal case when the heuristic is exactly equal to the magnitude of the residual, i.e., $\tilde{u}(x) = |\hat{f}(x) - y|$. Estimating

the magnitude of the residual is a straightforward way of quantifying a model's error, although it has two downsides. Firstly, it can only construct symmetric intervals, which makes the pixelwise intervals less informative and can inflate the set size. Second, unlike quantile regression, there is no guarantee that the residual estimate results in a valid prediction set without RCPS. Third, estimating the residual's magnitude is challenging since the training residuals are likely to be smaller than the test ones due to overfitting, unless an extra data split is used.

2.1.2. One Gaussian Per Pixel

We will now explain another common heuristic, which involves modeling each pixel as a sample from a Gaussian distribution with a particular mean and standard deviation (Nix & Weigend, 1994). Translating into our notation, $\hat{f}$ will be the mean function, and $\tilde{u} = \tilde{l}$ will be the standard deviation. We proceed by minimizing the negative log-likelihood of the Gaussian distribution,

$$\mathcal{L}(x, y) = \log\left(\tilde{u}(x)\right) + \frac{\left(\hat{f}(x) - y\right)^2}{\tilde{u}(x)}.$$

Like the residual magnitude method from Section 2.1.1, this heuristic is only suited to symmetric intervals and provides no guarantees of coverage without strong assumptions. Additionally, unlike the residual magnitude and quantile regression methods, one cannot use data splitting to avoid overconfidence due to overfitting.

2.1.3. Softmax Outputs

This next heuristic is most common in classification; indeed, it involves reframing image-to-image regression as a classification problem over a discrete set of K possible pixel values. Then, we train the image-to-image regression with a cross-entropy loss as if it were doing K-class classification for each pixel. This has two main downsides: first, the method is limited to resolutions of $1/K$, which makes it conservative; second, it requires a factor of $\mathcal{O}(K)$ more memory than the other heuristics due to the extra dimension. For the sake of space, we defer the precise details of this method to Appendix B.

2.1.4. Pixelwise Quantile Regression

This final heuristic is a multi-dimensional version of conformalized quantile regression (Romano et al., 2019; Koenker & Bassett Jr, 1978). If we want a 90% uncertainty interval, then reporting the interval between the estimated 95% and 5% quantiles for each pixel is a valid approach. Thus, we set $\tilde{u}$ to be an estimate of the $1 - \alpha/2$ conditional quantile and $\tilde{l}$ to be an estimate of the $\alpha/2$ conditional quantile. We estimate these pixelwise quantiles with a special loss function called a *quantile loss* (sometimes informally referred to as a *pinball loss*), shown below in its general form for the α quantile and its quantile estimator $\hat{q}_\alpha(x)$,

$$\mathcal{L}_\alpha(\hat{q}_\alpha(x), y) = (y - \hat{q}_\alpha(x))\alpha \mathbb{1}\left\{y > \hat{q}_\alpha(x)\right\} + (\hat{q}_\alpha(x) - y)(1 - \alpha)\mathbb{1}\left\{y \leq \hat{q}_\alpha(x)\right\}.$$

Omitting some algebra, we can see that the minimizer of this loss is the conditional quantile, i.e., $\text{Quantile}_{Y|X}(\alpha) = \min\{q : \mathbb{P}[Y < q \mid X] \leq \alpha\}$. Estimating $\hat{q}$ via empirical risk minimization should therefore approximate the conditional quantile. This can be made rigorous—under some regularity conditions, quantile regression converges asymptotically to the conditional quantile (Koenker & Bassett Jr, 1978; Chaudhuri, 1991; Steinwart & Christmann, 2011; Takeuchi et al., 2006; Zhou et al., 1996; Zhou & Portnoy, 1998). This analysis suggests that quantile regression could be practically effective.

Note that in this case, $\tilde{u}$ and $\tilde{l}$ must be trained with different loss functions, since they estimate different quantiles. Ultimately, we collapse these into one global loss for the heuristic,

$$\mathcal{L}(x, y) = \mathcal{L}_{\alpha/2}(\tilde{l}(x), y) + \mathcal{L}_{1-\alpha/2}(\tilde{u}(x), y).$$

After training, we expect $\tilde{l}$ and $\tilde{u}$ to approximate the $\alpha/2$ and $1 - \alpha/2$ quantiles respectively.

2.2. Calibrating Heuristic Notions of Uncertainty

As earlier discussed, we seek to form the RCPS in (1), which we can compute using any of the heuristics from Section 2.1. The function $\mathcal{T}$ will vary based on the heuristic notion of uncertainty used; however, the algorithm for selecting $\hat{\lambda}$ will provide the guarantee in Definition 2.1 regardless.

The calibration algorithm upper bounds the fraction of pixels falling outside the intervals, and then picks the smallest uncertainty intervals where the upper bound falls below α. Making this more concrete, we index the size of the intervals

Algorithm 2 Pseudocode for computing $\hat{\lambda}$

Input: Calibration data, (X_i, Y_i), $i = 1, \ldots, n$; risk level α; error rate δ; underlying predictor $\hat{f}$; heuristic lower and upper interval lengths $\tilde{l}$ and $\tilde{u}$; maximum value $\lambda_{\max}$; step size $d\lambda > 0$.
Output: Parameter $\hat{\lambda}$ for computing RCPS.

$\lambda \leftarrow \lambda_{\max}$
UCB $\leftarrow 1$
while UCB $\leq \alpha$ **do**
  $\lambda \leftarrow \lambda - d\lambda$
  for $i = 1, ..., n$ **do**
    $L_i \leftarrow L(\mathcal{T}_\lambda(X_i))$
  UCB $\leftarrow \frac{1}{n}\sum_{i=1}^{n} L_i + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$
$\hat{\lambda} \leftarrow \lambda + d\lambda$ # Backtrack by one (we overshot).

with a free multiplicative factor λ,

$$\mathcal{T}_\lambda(X)_{(m,n)} = \Big[\hat{f}(X)_{(m,n)} - \lambda\tilde{l}(X)_{(m,n)}, \\ \hat{f}(X)_{(m,n)} + \lambda\tilde{u}(X)_{(m,n)}\Big].$$

For a particular input image, when λ grows, the intervals grow; for a sufficiently large λ, the intervals will contain all of the ground truth pixel values. Our job is to pick $\hat{\lambda}$ to be the smallest value such that $\mathcal{T}_{\hat{\lambda}}(X)$ satisfies Definition 2.1 (note that $\mathcal{T}(X) = \mathcal{T}_{\hat{\lambda}}(X)$). Using the calibration dataset, we form Hoeffding's upper-confidence bound,

$$\hat{R}^+(\lambda) = \frac{1}{n}\sum_{i=1}^{n} L(\mathcal{T}_\lambda(X_i), Y_i) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}.$$

It is shown in (Hoeffding, 1963) that the Hoeffding bound is valid, that is, $\mathbb{P}\left[\hat{R}^+(\lambda) < R(\lambda)\right] < \delta$. Knowing this, we can use $\hat{R}^+(\lambda)$ to pick the smallest λ satisfying Definition 2.1. There is a closed-form expression for this process,

$$\hat{\lambda} = \min\left\{\lambda : \hat{R}^+(\lambda') \leq \alpha, \ \ \forall \alpha' \geq \alpha\right\}. \tag{3}$$

Proposition 2.3 ($\mathcal{T}_{\hat{\lambda}}$ is an RCPS (Bates et al., 2021))**.** *With $\hat{\lambda}$ selected as in (3), $\mathcal{T}_{\hat{\lambda}}$ satisfies Definition 2.1.*

For the proof of this fact, along with a discussion of the tighter confidence bounds used in the experiments, see previous work (Bates et al., 2021; Angelopoulos et al., 2021a). This calibration procedure is easy to implement in code, and we summarize it in Algorithm 2.

3. Experiments

The following sequence of experiments applies our methods to several challenging settings in biological imaging. The goal of these experiments is twofold. First, we demonstrate the utility of the procedures in practical experiments. Second, we evaluate the comparative effectiveness of the different heuristics qualitatively, as well as with a series of quantitative metrics. We will briefly discuss these metrics before providing the details of each experiment.

3.1. Evaluation Metrics

Empirical risk. The first quantity to notice is the risk, which should fall below α with probability $(1 - \delta)$. This is guaranteed in general by Proposition 2.3. For each dataset and heuristic, we make a histogram of the risk over several runs of the RCPS calibration, showing it is indeed controlled at the desired level.

Prediction set size. If the underlying heuristic notion of uncertainty is poor, then, in order to control the risk, the sets may need to be large. Generally, such an output is not informative to a practitioner, and all else equal, smaller intervals give more actionable assessments of the regression's quality. Thus, we report histograms of the interval size for each metric—smaller is better. As another view onto the regression's quality, ; we provide code in our GitHub for running Spearman correlations between the set size and the indicator of miscoverage (similar to measuring AUSE (Ilg et al., 2018)).

Size-stratified risk. Next, we seek prediction sets that do not systematically make mistakes in difficult parts of the image. Our risk control requirement in Definition 2.1 may be satisfied even if the prediction sets systematically fail to contain the most difficult pixels. For example, if $\alpha = 0.1$ and 90% of pixels are covered by fixed-width intervals of size 0.01, then the requirement is satisfied—however, the sets no longer serve as useful notions of uncertainty. To investigate such behavior, we evaluate the *size-stratified risk* (Angelopoulos & Bates, 2021)—i.e., we stratify pixels by the quartile of their interval sizes, and report the empirical risk within each of these quartiles. The desire is to have the risk be at approximately the same level for all strata, i.e., the risk should be as similar as possible between pixels with different set sizes. In other words, when we make a barplot of the stratified risk, the bars should all be the same height. Achieving this balance means the algorithm is not over-including easier-to-estimate pixels in order to excuse poor performance on difficult ones.

MSE of point prediction. Finally, we want to pick a heuristic notion of uncertainty that does not harm the accuracy of the point prediction during the joint training process. To measure this, we plot the *mean-squared error* (MSE) on the validation set for the point prediction which was jointly trained with each heuristic measure of uncertainty. A lower mean-squared error means that the joint training of the point prediction and heuristic uncertainty worked nicely, and did not degrade the point prediction. For certain heuristics, such as the Gaussian and softmax versions, this measure is particularly important because these methods do not directly optimize for the MSE and instead require a different procedure for supervising the point prediction (maximizing the Gaussian log likelihood and minimizing the cross-entropy loss, respectively).

Visualizations. In addition to the quantitative metrics, there is no substitute for seeing visualizations of the uncertainty intervals. For each example, we show the input, the output, the prediction sets generated by quantile regression, the absolute difference between the prediction and the ground truth, and the ground truth target. We represent the prediction sets by passing the pixelwise interval lengths through a colormap, where small sets render a pixel blue and large sets render it red. The interpretation, then, is that the redder a region is, the more uncertain it is and conversely, the bluer a region is, the more confident it is. Consequently, we expect the colormap to be red where the model is missing biological features and around fine structures such as edges

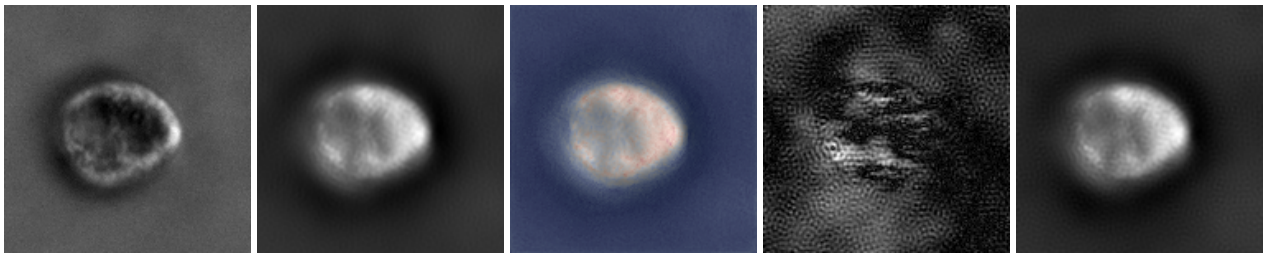

Figure 3. **Examples of quantitative phase reconstructions of leukocytes with uncertainty** shown in the following order: input (we only show one of the two illuminations), prediction, uncertainty visualization (produced with quantile regression), absolute difference between prediction and ground truth (renormalized for visualization), ground truth.

which are difficult to reconstruct from partial data.

Spatial miscoverage. As another way of evaluating the conditional coverage of our method, we look at spatial variations in prediction set errors. Specifically, we report the average miscoverage per-pixel as a spatial heatmaps in Appendix A.

3.2. Experimental Details

We use a standard experimental pipeline for all of the forthcoming experiments. In all experiments, we fit the predictor $\hat{f}$ and the heuristic notions of uncertainty $\tilde{u}$ and $\tilde{l}$ jointly. To ensure a level playing field and to promote reproducibility, the code used to define, train, and evaluate the model is shared among all heuristics and datasets. In order to run a new experiment (e.g., on a new dataset or with a new heuristic), minimal additional code is needed. See the code at this Github link: `https://github.com/aangelopoulos/im2im-uq`.

In each experiment, an 8-layer U-Net (Ronneberger et al., 2015) is used as the base model architecture and trained with an Adam optimizer for 10 epochs. We swept over two learning rates, $\{0.001, 0.0001\}$, and chose the learning rate that minimized the point prediction's MSE for each method in each experiment. All images get normalized to the interval $[0, 1]$. For the softmax heuristic, we discretized the prediction space with $K = 50$ because larger choices of K become too computationally expensive due to the amount of memory needed to store the extra dimension—a major practical limitation of this heuristic. We choose $\alpha = \delta = 0.1$ for the RCPS procedure in all cases, and adaptively select a grid of 1000 values of λ for each experiment. We evaluate each method by plotting its risk, average set size, size-stratified risk, and mean-squared error of the jointly trained prediction, as well as displaying an example. Further experimental details are available in the codebase.

We now discuss each experiment in detail. For each experiment, we include a brief background of the imaging problem, a description of the inputs X and outputs Y, and the aforementioned evaluation metrics for each heuristic.

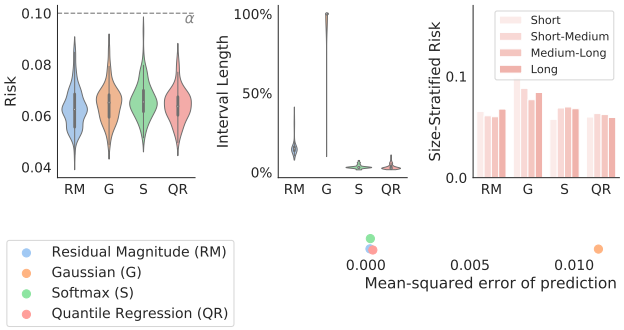

Figure 4. **A quantitative summary of all four heuristics** after RCPS calibration in the quantitative phase example. All methods control the risk, and quantile regression has the smallest set size. The Gaussian method has poor MSE, interval size, and size-stratified risk because it did not converge in training for either of the learning rates we chose.

3.3. Quantitative Phase Microscopy of Leukocytes

Background. In order to image the structure of cells—which are essentially transparent bags of water—one must measure their local refractive index, or equivalently, the phase delay incurred by light passing through each region of the cell. This task, known as quantitative phase imaging or QPI (Mir et al., 2012; Park et al., 2018; Jo et al., 2018), requires an algorithm to map intensity-only images to the phase value at each pixel, since it is impossible to directly measure the phase of light. Generally, as input to the algorithm, these methods take in a diverse set of intensity images captured under different imaging conditions, such as the illumination angle. Their performance improves with more input images.

Dataset description. Framing QPI as image-to-image regression, we take the input X to be the concatenation of two obliquely illuminated cell intensity images (from opposite angles) and the target Y to be a reference phase image. Y is obtained using an analytic phase recovery technique known as differential phase contrast (DPC), which takes four or more images as input (Tian & Waller, 2015). The utility, then, of the regression model is to reduce the number of input images needed for high quality phase prediction, thereby improving the tradeoff between acquisition speed and prediction accuracy. Furthermore, the quantitative phase values

have an intrinsic meaning, so by adding uncertainty intervals which are on the scale of the phase values, we provide an important inferential tool for analyzing cell morphology.

For the experiment, we use the Berkeley Single Cell Computational Microscopy (BSCCM) dataset (Pinkard, 2021), which contains 2,000 single leukocyte (white blood cell) images with 150x150 pixels taken using several imaging modalities. Of particular interest to us, this dataset includes images taken under a variety of different angles of illumination co-registered with quantitative phase maps obtained via 4-image DPC. As input to the U-Net, we concatenate two obliquely illuminated cell images along the channel dimension. We use 1800 randomly selected data points with a batch size of 64 to train the model, 100 points for calibration, and 100 points for validation. Our results are visualized in Figure 3.

Results. We report our results in Figure 5. As promised by the calibration procedure, risk-control holds for all choices of heuristic uncertainties. In terms of statistical power, we see that quantile regression outcompetes the other heuristics in the trifecta of evaluations—it has the smallest average set size and best size-stratified coverage while remaining competitive with other methods in mean-squared error. Altogether, these metrics express that the uncertainty intervals produced by calibrated quantile regression are tight and adaptive to the model's performance, even among different pixels within a single prediction. The softmax heuristic, though seemingly competitive in these evaluations, gives nearly fixed-width intervals, most of which have exactly the same size because of the discretization.

3.4. Fast Magnetic Resonance Imaging

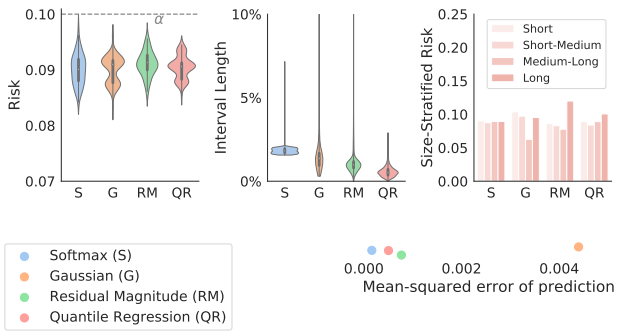

Figure 5. **A quantitative summary of all four heuristics** after RCPS calibration in the FastMRI example. All three methods control the risk. Quantile regression has the smallest interval size and best size-stratified risk.

Background. Much like our previous example, in MRI there exists a tradeoff between imaging speed and quality. MRI directly samples an object's spatial frequency (k-space) over time; so it is possible to reduce the scan time by lowering the effective sampling rate in k-space. Although fast imaging is more comfortable for human subjects and also critical for certain fast movements like the beating of the heart, insufficient sampling results in low quality, aliased MR images. However, with deep learning, we can try to fill in the information lost by undersampling to emulate fully sampled images, thereby getting the joint benefits of fast scan times and high quality reconstructions.

Dataset description. The inputs X are the undersampled images formed by downsampling k-space by a factor of four along a single dimension (the phase encoding direction), and then taking an inverse Fourier transform. Our outputs Y are the fully sampled MR images. Successfully regressing X to Y essentially accelerates the MRI scan time by a factor of four.

We use the FastMRI dataset for this example (Zbontar et al., 2018b). The dataset includes 10,000 clinical knee MR volumes taken with 3T or 1.5T magnets which are algorithmically undersampled with k-space masks that emulate fast sampling strategies. We dissect the volumes into 27,993 randomly selected 320x320 pixel coronal knee slices for training the model, 3,474 for the RCPS calibration, and 3,474 for validation. We use a batch size of 78. For the Gaussian method, we standardized the output space to be mean zero and unit variance, since it failed to properly train when normalized to fall in the interval $[0, 1]$.

Results. Qualitatively, Figure 1 shows an example of an MRI reconstruction using calibrated quantile regression. The predictions are slightly blurred versions of the ground truth, likely due to the network's bias toward low frequency outputs (Rahaman et al., 2019). The uncertainty intervals have large values in areas with high contrast, expressing the intrinsic uncertainty in localizing edges using incomplete information. The quantitative results of this experiment, visualized in Figure 5, are in line with those of the QPI experiment; we achieve the desired risk level and quantile regression performs best on all metrics. Although the softmax heuristic has near-even size-stratified risk, this is because it outputs quantized sets of nearly fixed size, and the strata are therefore decided by random tie-breaking.

3.5. Fly Brain Transmission Electron Microscopy

Background. Finally, we perform algorithmic super-resolution transmission electron microscopy (TEM) of the brain of a Drosophila melanogaster (fruit fly). TEM uses focused electron beams rather than visible light to produce images, and due to the small de Broglie wavelength of electrons, it can achieve significantly higher resolution than visible light microscopy, on the order of single nanometers. However, TEM sequentially scans over the sample volume, imaging point-by-point; thus, its scan time scales cubically with the desired resolution. For large volumes like the fly brain shown in Figure 6, imaging can take years. Upsampling lower resolution (say 16nm) TEM data to high

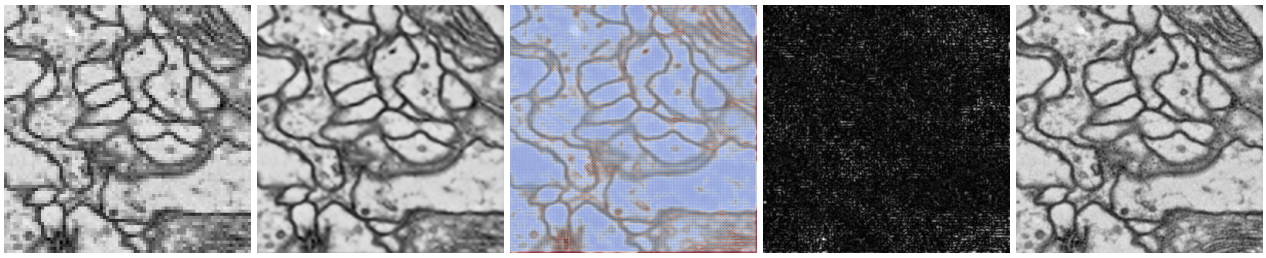

Figure 6. **Examples of Drosophila brain reconstructions with uncertainty** shown in the following order: input, prediction, uncertainty visualization, absolute difference between prediction and ground truth, ground truth. We use the pixelwise quantile regression version of the procedure.

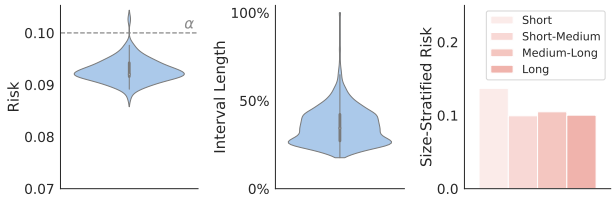

Figure 7. **Quantitative results of pixelwise quantile regression** on the TEMCA fly brain dataset. The risk is controlled, the intervals have reasonable lengths, and the size-stratified risk is roughly balanced, although slightly more permissive with small intervals.

resolution (4nm) images could therefore save months of time.

Dataset description. We consider super-resolution as image-to-image regression, where X is simply a manually downsampled version of a 4nm TEM image Y. In particular, we consider a factor of 4x nearest-neighbor downsampling along both image dimensions to emulate the acquisition of a 16nm TEM image.

We use the Janelia Transmission Electron Microscopy Camera Array (TEMCA2) dataset of the Full Adult Fly Brain (Zheng et al., 2018). This dataset contains a 26 TB brain volume at four nanometer resolution isotropically along all dimensions. As a consequence of this data burden, we did not run the full suite of procedures on the dataset, and instead ran only the consistently best performing method—pixelwise quantile regression. The dataset size is unknown in advance because we randomly sample subregions of a TEM slice *on-the-fly*, then throw the sample out if it has sufficiently many black pixels (i.e., does not contain tissue of interest). The fraction of images that are thrown out is random and unknown before runtime because the full 16TB dataset cannot be stored in random access memory. This procedure is reproducible via our GitHub. We used roughly 2M images of size 320x320 for training, 25K images for calibration, and 25K images for validation. We use a batch size of 16. Each image is only seen once.

Results. The results of this procedure are depicted in Figure 6. The quantitative measures in Figure 7 are similar to those of past experiments. Qualitatively, the prediction sets identify regions of high contrast as more uncertain, perhaps due to the spectral bias of CNNs. The sets are periodically zero-length, i.e., fully confident, every four pixels. This is because the input is a downsampled version of the target, so the model perfectly knows every fourth pixel. Consequently, at those pixels, the model does not have any uncertainty. This highlights the adaptivity and tightness of the prediction sets; they are not only useful in understanding where the model is poor, but also where the model performs reliably.

4. Related Work

Image-to-image regression. Methods for interpolating between samples of a two dimensional digital function, which we now call *image super-resolution*, have existed since antiquity (Generations of Chinese mathematicians, 200BC-100AD; Needham & Gwei-Djen, 1959; Ptolemy, 200AD). Such methods, along with classical spline approximations (Hahn, 1918; Schoenberg & Whitney, 1953; Walsh et al., 1962), have been commonly deployed since at least the 1980s (Keys, 1981). In the 21st century, learning-based approaches (Freeman et al., 2002; Chang et al., 2004; Yang et al., 2010) have dominated the research conversation, particularly those using convolutional neural networks (Dong et al., 2014; 2015; Johnson et al., 2016) and generative adversarial networks (Goodfellow et al., 2014; Ledig et al., 2017). Beyond interpolation, image-to-image regression also encompasses denoising (Buades et al., 2005b;a; Burger et al., 2012; Tian et al., 2020; Goyal et al., 2020), style transfer (Gatys et al., 2016; Isola et al., 2017; Zhu et al., 2017; Jing et al., 2020), image colorization (Zhang et al., 2016), and so on. A line of work based on the U-Net (Ronneberger et al., 2015) adapts the above techniques for biomedical imaging problems, achieving strong results (Zbontar et al., 2018a; Zhou et al., 2018). We build directly on this work.

Heuristic notions of uncertainty. Parameterizing a Gaussian distribution and maximizing its log-likelihood with gradient descent has been employed since at least 1994 (Nix & Weigend, 1994). The idea of the cross-entropy loss leading to the softmax distributional estimate has its roots in the Kraft-McMillan theorem (Kraft, 1949; McMillan, 1956) and related information-theoretic concepts (Thomas & Cover, 1999). Quantile regression was proposed in the mid-1970s by Koenker & Bassett Jr (1978). Many analyses

and variations of the technique have been proposed, such as for local polynomials (Chaudhuri, 1991), in additive models (Koenker, 2011) or for conditional coverage (Feldman et al., 2021). Quantile regression comes with an asymptotic guarantee of conditional coverage under certain regularity conditions (Chaudhuri, 1991; Steinwart & Christmann, 2011; Takeuchi et al., 2006; Zhou et al., 1996; Zhou & Portnoy, 1998). Many papers have used the technique, applying it to economics (Chaney et al., 2011; McKenzie & Rapoport, 2007; Machado & Mata, 2005), machine learning (Hwang & Shim, 2005; Meinshausen & Ridgeway, 2006; Natekin & Knoll, 2013), medical research (Armitage et al., 2008), and more. Accessible references to the topic of quantile regression are provided (Koenker, 2005; Koenker et al., 2018). Ensemble methods (Hansen & Salamon, 1990; Lakshminarayanan et al., 2017; Fort et al., 2019) and Bayesian methods such as MC-Dropout (Gal & Ghahramani, 2016) are also common in deep learning. Prediction intervals can also be formed with interval neural networks (Oala et al., 2020); our codebase includes the capability to calibrate and evaluate this heuristic as well. These heuristics do not have finite-sample guarantees and fall outside the scope of our discussion; see (Gal, 2016) for an introduction.

Distribution-free uncertainty quantification. Conformal prediction is a lightweight procedure for creating prediction sets with finite-sample coverage while requiring no model retraining (Vovk et al., 1999; 2005; Lei et al., 2015; 2013; Lei, 2014; Sadinle et al., 2019). We build directly on conformalized quantile regression (CQR) (Romano et al., 2019). We replace the conformal subroutine in CQR with the fixed-sequence testing procedure from (Bates et al., 2021; Angelopoulos et al., 2021a). Other works have applied distribution-free uncertainty quantification to biological and medical computer vision tasks (Hechtlinger et al., 2018; Cauchois et al., 2021; Romano et al., 2020; Angelopoulos et al., 2021b;c; Lu et al., 2021). However, we are not aware of any that have studied image-to-image regression. An introduction to these topics is available in (Angelopoulos & Bates, 2021).

5. Conclusion

The methods described herein allow for rigorous per-pixel uncertainty estimation in image-to-image regression problems. In our view, the major limitation of this work is the inability to express uncertainty at a conceptual level. The uncertainty maps produced by the methods herein often co-occur with edges and other high-frequency objects. Future work may focus on producing uncertainties that are disentangled at a conceptual level or attempt to disaggregate error due to the presence of high frequencies from other factors.

Acknowledgements

We would like to acknowledge Dr. Henry Pinkard for early access to the BSCCM dataset. A.A. was supported by the NSFGRFP and a Berkeley Fellowship. A.K. is funded by the Berkeley Fellowship for Graduate Study. T.A. and S.U. are funded by the Philomathia Foundation, Chan Zuckerberg Initiative Imaging Scientist program. S.U. is a Chan Zuckerberg Biohub Investigator. Y.R. was supported by the Israel Science Foundation (grant 729/21). Y.R. also thanks the Career Advancement Fellowship, Technion, for providing research support. This work is partially supported by NSF Grant IIS-1901252.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021a.

Angelopoulos, A. N., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2021b.

Angelopoulos, A. N., Bates, S., Zrnic, T., and Jordan, M. I. Private prediction sets. *arXiv preprint arXiv:2102.06202*, 2021c.

Armitage, P., Berry, G., and Matthews, J. N. S. *Statistical methods in medical research*. John Wiley & Sons, 2008.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the Association for Computing Machinery*, 68 (6), 9 2021.

Buades, A., Coll, B., and Morel, J.-M. A non-local algorithm for image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 60–65, 2005a.

Buades, A., Coll, B., and Morel, J.-M. A non-local algorithm for image denoising. *Multiscale Modeling and Simulation*, 4(2):490–530, 2005b.

Burger, H. C., Schuler, C. J., and Harmeling, S. Image denoising: Can plain neural networks compete with BM3D? In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2392–2399. IEEE, 2012.

Cauchois, M., Gupta, S., and Duchi, J. C. Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81):1–42, 2021.

Chaney, P. K., Faccio, M., and Parsley, D. The quality of accounting information in politically connected firms. *Journal of Accounting and Economics*, 51(1-2):58–76, 2011.

Chang, H., Yeung, D.-Y., and Xiong, Y. Super-resolution through neighbor embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pp. I–I. IEEE, 2004.

Chaudhuri, P. Global nonparametric estimation of conditional quantile functions and their derivatives. *Journal of Multivariate Analysis*, 39(2):246–269, 1991.

Dong, C., Loy, C. C., He, K., and Tang, X. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pp. 184–199. Springer, 2014.

Dong, C., Loy, C. C., He, K., and Tang, X. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2015.

Feldman, S., Bates, S., and Romano, Y. Improving conditional coverage via orthogonal quantile regression. In *Advances in Neural Information Processing Systems*, 2021.

Fort, S., Hu, H., and Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Freeman, W. T., Jones, T. R., and Pasztor, E. C. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65, 2002.

Gal, Y. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Gatys, L. A., Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, 2016.

Generations of Chinese mathematicians. *The Nine Chapters on the Mathematical Art*. 200BC-100AD.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

Goyal, B., Dogra, A., Agrawal, S., Sohi, B. S., and Sharma, A. Image denoising review: From classical to state-of-the-art approaches. *Information Fusion*, 55:220–244, 2020.

Hahn, H. Über das interpolationsproblems. *Mathematische Zeitschrift*, 1:115–142, 1918.

Hansen, L. K. and Salamon, P. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

Hechtlinger, Y., Póczos, B., and Wasserman, L. Cautious deep learning. *arXiv preprint arXiv:1805.09460*, 2018.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Hwang, C. and Shim, J. A simple quantile regression via support vector machine. In *International Conference on Natural Computation*, pp. 512–520. Springer, 2005.

Ilg, E., Cicek, O., Galesso, S., Klein, A., Makansi, O., Hutter, F., and Brox, T. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 652–667, 2018.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., and Song, M. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics*, 26(11):3365–3385, 2020.

Jo, Y., Cho, H., Lee, S. Y., Choi, G., Kim, G., Min, H., and Park, Y. Quantitative phase imaging and artificial intelligence: a review. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(1):1–14, 2018.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pp. 694–711. Springer, 2016.

Keys, R. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160, 1981.

Koenker, R. *Quantile Regression*. Cambridge University Press, 2005.

Koenker, R. Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239–262, 2011.

Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46(1):33–50, 1978.

Koenker, R., Chernozhukov, V., He, X., and Peng, L. Handbook of quantile regression. 2018.

Kraft, L. G. *A device for quantizing, grouping, and coding amplitude-modulated pulses*. PhD thesis, Massachusetts Institute of Technology, 1949.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, 2017.

Lei, J. Classification with confidence. *Biometrika*, 101(4): 755–769, 10 2014.

Lei, J., Robins, J., and Wasserman, L. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Lei, J., Rinaldo, A., and Wasserman, L. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43, 2015.

Lu, C., Lemay, A., Chang, K., Hoebel, K., and Kalpathy-Cramer, J. Fair conformal predictors for applications in medical imaging. *arXiv preprint arXiv:2109.04392*, 2021.

Machado, J. A. and Mata, J. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*, 20(4):445–465, 2005.

McKenzie, D. and Rapoport, H. Network effects and the dynamics of migration and inequality: Theory and evidence from Mexico. *Journal of Development Economics*, 84(1): 1–24, 2007.

McMillan, B. Two inequalities implied by unique decipherability. *IRE Transactions on Information Theory*, 2(4): 115–116, 1956.

Meinshausen, N. and Ridgeway, G. Quantile regression forests. *Journal of Machine Learning Research*, 7(6), 2006.

Mir, M., Bhaduri, B., Wang, R., Zhu, R., and Popescu, G. Quantitative phase imaging. *Progress in Optics*, 57 (133-37):217, 2012.

Natekin, A. and Knoll, A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7:21, 2013.

Needham, J. and Gwei-Djen, L. *Science and Civilisation in China: Volume 3, Mathematics and the Sciences of the Heavens and the Earth*. Cambridge University Press, 1959.

Nix, D. and Weigend, A. Estimating the mean and variance of the target probability distribution. In *IEEE International Conference on Neural Networks*, volume 1, pp. 55–60, 1994.

Oala, L., Heiß, C., Macdonald, J., März, M., Samek, W., and Kutyniok, G. Interval neural networks: Uncertainty scores. *arXiv preprint arXiv:2003.11566*, 2020.

Park, Y., Depeursinge, C., and Popescu, G. Quantitative phase imaging in biomedicine. *Nature Photonics*, 12(10): 578–589, 2018.

Pinkard, H. Berkeley single cell computational microscopy dataset, 2021. URL https://henrypinkard.github.io/.

Ptolemy. *Amagest*. 200AD.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In *International Conference on Machine Learning*, volume 97, pp. 5301–5310. PMLR, 2019.

Romano, Y., Patterson, E., and Candès, E. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3543–3553. 2019.

Romano, Y., Sesia, M., and Candès, E. Classification with valid and adaptive coverage. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591. Curran Associates, Inc., 2020.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Sadinle, M., Lei, J., and Wasserman, L. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114:223 – 234, 2019.

Schoenberg, I. J. and Whitney, A. On Pólya frequency functions III. the positivity of translation determinants whith an application to the interpolation problem by spline curves. *Transactions of the American Mathematical Society*, 74:246–259, 1953.

Steinwart, I. and Christmann, A. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17 (1):211–225, 2011.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.

Thomas, J. A. and Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., and Lin, C. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, 2020.

Tian, L. and Waller, L. Quantitative differential phase contrast imaging in an led array microscope. *Optics Express*, 23(9):11394–11403, 2015.

Vovk, V., Gammerman, A., and Saunders, C. Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning*, pp. 444–453, 1999.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, New York, NY, USA, 2005.

Walsh, J., Ahlberg, J., and Nilson, E. Best approximation properties of the spline fit. *Journal of Mathematics and Mechanics*, 11(2):225–234, 1962.

Weigert, M., Schmidt, U., Boothe, T., Müller, A., Dibrov, A., Jain, A., Wilhelm, B., Schmidt, D., Broaddus, C., Culley, S., et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nature Methods*, 15 (12):1090–1097, 2018.

Yang, J., Wright, J., Huang, T. S., and Ma, Y. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010.

Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M. J., Defazio, A., Stern, R., Johnson, P., Bruno, M., et al. Attention U-Net: Learning where to look for the pancreas. *Conference on Medical Imaging with Deep Learning*, 2018a.

Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M. J., Defazio, A., Stern, R., Johnson, P., Bruno, M., et al. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018b.

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *European Conference on Computer Vision*, pp. 649–666. Springer, 2016.

Zheng, Z., Lauritzen, J. S., Perlman, E., Robinson, C. G., Nichols, M., Milkie, D., Torrens, O., Price, J., Fisher, C. B., Sharifi, N., et al. A complete electron microscopy volume of the brain of adult Drosophila melanogaster. *Cell*, 174(3):730–743, 2018.

Zhou, K. Q. and Portnoy, S. L. Statistical inference on heteroscedastic models based on regression quantiles. *Journal of Nonparametric Statistics*, 9(3):239–260, 1998.

Zhou, K. Q., Portnoy, S. L., et al. Direct use of regression quantiles to construct confidence sets in linear models. *The Annals of Statistics*, 24(1):287–306, 1996.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. U-Net++: A nested U-Net architecture for medical image segmentation. In Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J. M. R., Bradley, A., Papa, J. P., Belagiannis, V., Nascimento, J. C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., and Madabhushi, A. (eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, Cham, 2018. Springer International Publishing.

Zhu, J., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.

A. Spatial Miscoverage

A.1. Quantitative Phase Microscopy of Leukocytes

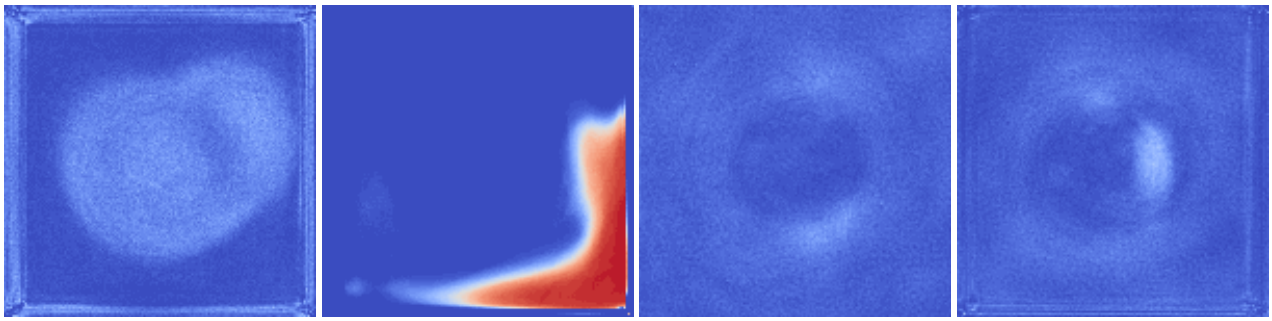

Figure 8. **Spatial variations in miscoverage** in the BSCCM dataset are shown for each of the four methods as a heatmap. Blue represents 0% miscoverage and red represents 100%. The methods are, in order, residual magnutude, gaussian, softmax, and quantile regression.

A.2. Fast Magnetic Resonance Imaging

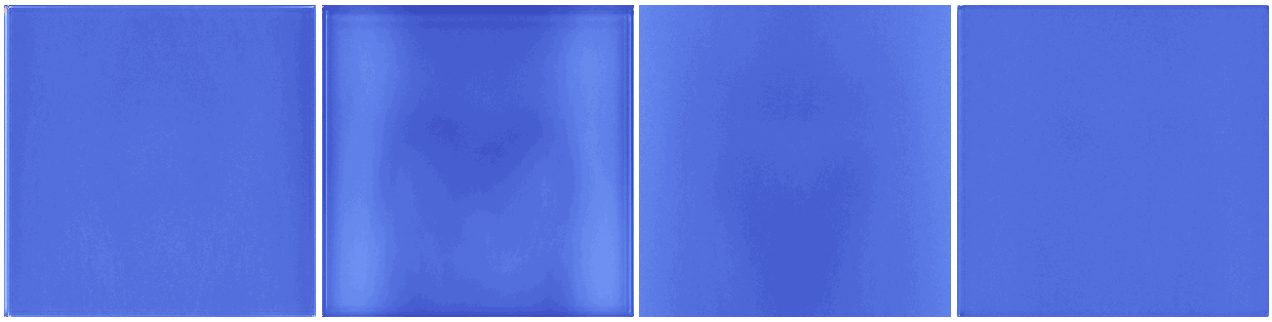

Figure 9. **Spatial variations in miscoverage** in the fast MRI dataset are shown for each of the four methods as a heatmap. Blue represents 0% miscoverage and red represents 100%. The methods are, in order, residual magnutude, gaussian, softmax, and quantile regression.

B. Mathematical Description of the Softmax Heuristic Notion of Uncertainty

The softmax heuristic is different from the others described in Section 2; the functions $\tilde{u}$ and $\tilde{l}$ are not equal, and they are not learned directly. Instead, we train the network to produce an entire probability distribution, and directly extract all three of $\hat{f}$, $\tilde{u}$, and $\tilde{l}$.

Let us first discretize the possible pixel values into K categories: $\{0, \frac{1}{K-1}, ..., \frac{K-1}{K-1}\}$. We then associate a discrete label with an otherwise continuous label via the function

$$D(y) = \left| \left\{ i : i \in 0, 1, ..., K-1 \text{ and } \frac{i}{K-1} \geq y \right\} \right|.$$

Intuitively, the function $D(y)$ discretizes $[0, 1]$, then bins the pixel accordingly.

This allows us to train the neural network to output distributions over pixel values $\hat{\pi}_y(x)$ estimating the conditional probabilities $\mathbb{P}[Y = y \mid X = x]$ via the cross-entropy loss,

$$\mathcal{L}(x, y) = \frac{1}{MN} \sum_{\substack{1 \leq i \leq M \\ 1 \leq j \leq N}} -\hat{\pi}_{D(y)}(x) + \log \left(\sum_{k=1}^{K} \exp\left(\hat{\pi}_{D(k)}(x)\right) \right).$$

Finally, we can extract the prediction and heuristic uncertainties,

$$\begin{aligned}
\hat{f}(x) &= \frac{1}{K-1}\underset{k}{\text{argmax}}\ \hat{\pi}_k(x); \\
\tilde{u}(x) &= \frac{1}{K-1}\text{Quantile}\left(1-\frac{\alpha}{2}, \hat{\pi}(x)\right); \\
\tilde{l}(x) &= \frac{1}{K-1}\text{Quantile}\left(\frac{\alpha}{2}, \hat{\pi}(x)\right),
\end{aligned}$$

where

$$\text{Quantile}\left(\beta, \hat{\pi}(x)\right) = \min\left\{K' : \sum_{k=1}^{K'} \hat{\pi}_k(x) \geq \beta\right\}.$$

The softmax approach requires discretizing $\mathcal{Y}$ into K bins, which can severely limit its utility. The heuristic can only create prediction sets whose endpoints are multiples of $1/K$, which may make it too conservative. Furthermore, the output representation can be enormous, making the memory constraints infeasible for large images (e.g., for $K = 256$, the model produces an output of size $M \times N \times 256$).