This is a repository copy of *Uncertainty estimation for heatmap-based landmark localization*.

Article:

Uncertainty Estimation for Heatmap-based Landmark Localization

Lawrence Schöbs, Andrew J. Swift, and Haiping Lu, *Senior Member, IEEE*

Abstract— **Automatic anatomical landmark localization has made great strides by leveraging deep learning methods in recent years. The ability to quantify the uncertainty of these predictions is a vital component needed for these methods to be adopted in clinical settings, where it is imperative that erroneous predictions are caught and corrected. We propose Quantile Binning, a data-driven method to categorize predictions by uncertainty with estimated error bounds. Our framework can be applied to any continuous uncertainty measure, allowing straightforward identification of the best subset of predictions with accompanying estimated error bounds. We facilitate easy comparison between uncertainty measures by constructing two evaluation metrics derived from Quantile Binning. We compare and contrast three epistemic uncertainty measures (two baselines, and a proposed method combining aspects of the two), derived from two heatmap-based landmark localization model paradigms (U-Net and patch-based). We show results across three datasets, including a publicly available Cephalometric dataset. We illustrate how filtering out gross mispredictions caught in our Quantile Bins significantly improves the proportion of predictions under an acceptable error threshold. Finally, we demonstrate that Quantile Binning remains effective on landmarks with high aleatoric uncertainty caused by inherent landmark ambiguity, and offer recommendations on which uncertainty measure to use and how to use it. The code and data are available at `https://github.com/schobs/qbin`.**

Index Terms— **Uncertainty estimation, landmark localization, confidence, heatmaps, U-Net**

I. INTRODUCTION

AUTOMATIC landmark localization is an important step in many medical image analysis methods, such as image segmentation [1] and image registration [2], [3]. An erroneous landmark prediction at an early stage of analysis will flow downstream and compromise the validity of final conclusions. Therefore, the ability to quantify the uncertainty of a prediction is a vital requirement in a clinical setting where explainability is crucial and there is a human in-the-loop to correct highly uncertain predictions [4].

In this study we propose Quantile Binning, a data-driven framework to estimate a prediction's quality by learning the relationship between any continuous uncertainty measure and

localization error. Using the framework, we place predictions into bins of increasing subject-level, epistemic uncertainty and assign each bin a pair of estimated localization error bounds. These bins can be used to identify the subsets of predictions with expected high or low localization errors, allowing the user to make a choice of which subset of predictions to review and reannotate based on their expected error bounds. Our method is agnostic to the particular uncertainty metric used, as long as it is continuous and the true function between the uncertainty metric and localization error is monotonically increasing. We showcase our method using three uncertainty measures: a baseline derived from predicted heatmap activations, a strong baseline of ensemble model prediction variance [5], as well as introducing our own measure based on ensemble heatmap activations. Furthermore, we introduce two uncertainty evaluation methods, measuring how well an uncertainty measure truly predicts localization error and the accuracy of our predicted error bounds.

We explore the efficacy of our three uncertainty metrics on two paradigms of localization models: an encoder-decoder U-Net that regresses Gaussian heatmaps [6], and a patch-based network that generates a heatmap from patch voting, PHD-Net [7]. We compare how the same uncertainty measures perform under the two different approaches to landmark localization on two Cardiac Magnetic Resonance Imaging (CMR) datasets of varying difficulty and a publicly available Cephalometric dataset [8], finding promising results for both paradigms. Furthermore, we explore the effect of aleatoric uncertainty caused by landmark ambiguity on Quantile Binning and our uncertainty measures. Our Quantile Binning method is generalizable to any continuous uncertainty measure, and the examples we investigate in this study can be applied as a post-processing step to any heatmap-based landmark localization method. We aspire that this work can be used as a framework to build, evaluate and compare uncertainty metrics in landmark localization beyond those demonstrated in this paper.

II. RELATED WORK

A. Landmark Localization

The recent advancement in machine learning has led to convolutional neural networks (CNNs) dominating the task of landmark localization. Encoder-decoder methods, originally proposed for the task of image segmentation [6], have cemented themselves as one of the leading approaches for landmark localization in both the medical domain [9]–[12] and computer vision [13], [14]. The architecture of these methods allow the analysis of images at multiple resolutions,

learning to predict a Gaussian heatmap centred around the predicted landmark location. The activation of each pixel in the heatmap can be interpreted as the pseudo-probability of the pixel being the target landmark. The network learns to generate a high response near the landmark, smoothly attenuating the responses in a small, predefined radius around it. Regressing heatmaps proves more effective than regressing coordinates [15], as the heatmap image offers smoother supervision than direct coordinate values, and also models some uncertainty in the prediction.

However, in medical imaging the number of available training samples is often small so the encoder-decoder network is forced to be shallow, compromising its performance [15]. One method to overcome this is via a *patch-based* approach; alleviating the problem by sampling many small "patches' from an image, learning the relationship between each patch and the target landmark [16], [17]. This approach can generate orders of magnitude more training samples from a single image compared to the encoder-decoder style methods. Furthermore, patch-based models that use Fully Convolutional Networks (FCN) have fewer parameters than encoder-decoder architectures, decreasing computational requirements and training times [7].

Noothout *et al.* [18] implemented a patch-based network using an FCN to jointly perform classification and regression on each patch. The coarse binary classification task determines whether a patch contains the landmark, and the more precise regression task estimates the displacement from the patch to the landmark. This multi-task, joint learning leads to a lightweight network and enhanced localization performance, with the two tasks sharing a feature representation that improves the performance of both [19]. However, the resulting network has a strong local focus and is also susceptible to failure if the predicted patch from the classification task is incorrect. In a follow-up work, Noothout *et al.* [20] extended their work [18] into a two-stage method: they first train a CNN to provide global estimates for the landmarks, then employ specialised CNNs for each landmark for the final prediction. This method improves upon the first in terms of localization error, but has the drawback of requiring multiple training stages.

To mitigate the inherent local focus of the patch-based methods, we extended the patch-based network [18] by borrowing heatmap regression from the encoder-decoder networks; reforming the binary classification task as a Gaussian heatmap regression task [7]. Named PHD-Net (Patch Heatmap & Displacement regression), this smoother supervision improved performance, reducing misidentifications compared to using the classification branch from the prior work [18]. Furthermore, we introduced the method Candidate Smoothing, combining the features from the two branches to output more accurate predictions along with an uncertainty measure.

B. Uncertainty Estimation

Estimating the uncertainty of machine learning predictions is a topic of growing interest, particularly relevant in the domain of medical imaging where there is often a human in the loop to manually correct flagged predictions. A concentrated effort in uncertainty estimation has been applied to image segmentation by the community, a task similar to landmark localization that instead aims to predict a mask for an entire structure rather than a single point.

Segmentation aims to produce a binary map, with pixels activated on the structure of interest and inactive elsewhere. In traditional heatmap-based landmark localization, only the landmark coordinate has a magnitude of 1, smoothly attenuating to 0 in a set radius. The loss function used by landmark localization is the Mean Squared Error (MSE) between the target and predicted heatmap, whereas segmentation uses pixel-wise classification-based losses [21]. Nevertheless, the tasks are similar in that the magnitude of the activation of any given pixel in each image can be leveraged for information on the epistemic "confidence" (inverse of uncertainty) of the model. Jungo *et al.* [21] use pixel activation to measure the uncertainty of each pixel segmentation class, including using the average activation of an ensemble. They found that this naive approach was surprisingly well calibrated and that ensembling $5+$ identical but randomly initialized models significantly improved calibration. These results have been corroborated by Mehrtash *et al.* who also used pixel activation for confidence as well as an ensemble of 5 identical, randomly initialized networks to convincing effect [22].

Other successful approaches for epistemic uncertainty estimation use Bayesian Neural Networks [23] or Bayesian approximation methods like Monte-Carlo dropout [24]. However, the prevailing approach is an ensemble of identical, randomly initialized networks. This method affords better performance [22] and a more accurate mechanism for Bayesian marginalization [25] compared to a single model using Monte-Carlo dropout. Fort *et al.*'s study suggests this behaviour is due to random initializations exploring entirely different modes of the loss landscape, facilitating a powerful decorrelation effect between models [26]. We see extensive use of ensembles of identical models with random initializations in the domain of medical image segmentation to improve accuracy and estimate uncertainty [21], [22], [27], [28].

In landmark localization we are ultimately predicting a single coordinate point rather than a mask, but similar uncertainty estimation approaches can be utilized. However, there are limited works exploring uncertainty in landmark localization. Payer *et al.* [9] directly modeled aleatoric uncertainty during training by learning the isotropic Gaussian covariances of target heatmaps, and predicting the distribution of likely locations of the landmark at test time. Thaler *et al.* took this approach further, learning anisotropic (directionally skewed) Gaussian heatmaps for each landmark, demonstrating that the learned heatmap shapes correspond to inter-observer variability from multiple annotators [12]. In terms of epistemic uncertainty, Lee *et al.* [29] borrowed from image segmentation approaches by proposing a Bayesian CNN that utilized Monte-Carlo dropout to predict the location and subject-level uncertainty of cephalometric landmarks.

Another method to measure the subject-level, epistemic uncertainty of a heatmap-based landmark prediction is to measure the maximum heatmap activation (MHA) of the predicted heatmap. Since the activation of a Gaussian heatmap at a particular pixel represents the pseudo-probability of the
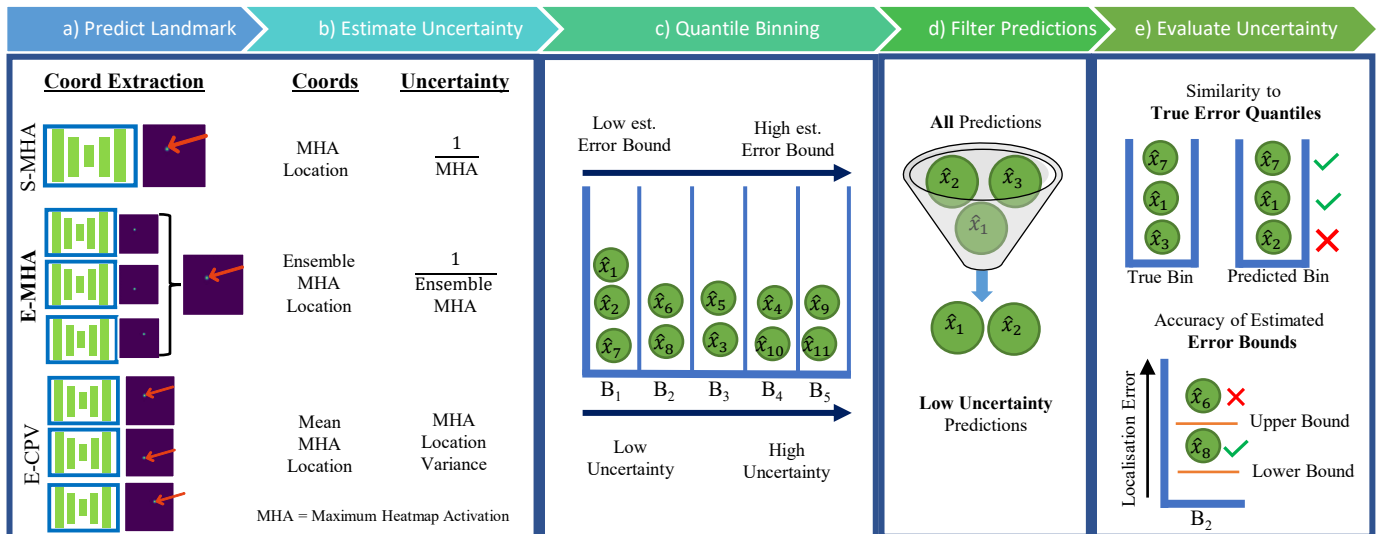
Fig. 1: Overview of our general Quantile Binning framework. **a)** We make a prediction using a heatmap-based landmark localization model, and **b)** extract a continuous uncertainty measure. **c)** We learn thresholds to categorize predictions into bins of increasing uncertainty, estimating error bounds for each bin. **d)** We filter out predictions from high uncertainty bins to improve the proportion of acceptable predictions. **e)** Finally, we evaluate each uncertainty measure's ability to capture the true error quantiles and the accuracy of the estimated error bounds.

pixel being the landmark, we can use this pseudo-probability as an uncertainty measure: the higher the activation, the more certain the prediction. Drevicky *et al.* [5] compared MHA with ensemble and Monte-Carlo dropout methods, finding MHA surprisingly effective given its simplicity. However, similarly to image segmentation, they found using an ensemble of models was best at predicting uncertainty. They calculated the coordinate prediction variance between an ensemble of models, and found this method performed best at estimating prediction uncertainty.

In our earlier work utilising the patch-based model PHD-Net, MHA was also used as the uncertainty metric [7]. However, the heatmap analysed is distinctly different from the heatmaps predicted by encoder-decoder networks. Rather than explicitly learning a Gaussian function centred around the landmark, the approach combined patch-wise heatmap and displacement predictions. We produced a new non-Gaussian heatmap, where the activation of each pixel is defined by the number of patches that voted for it, regularized by the coarse global likelihood prediction. Therefore, the resulting heatmap represents patch-wise ensemble votes rather than a Gaussian function, where the MHA is the pixel with the most "patch votes".

To the best of our knowledge, no study has investigated how heatmap-based uncertainty estimation measures can be used to filter out poor predictions in landmark localization. Furthermore, no general framework has been proposed to compare how well uncertainty measures can predict localization error - an important practical application in clinical settings.

III. CONTRIBUTIONS

In this paper, we propose a general framework to compare and evaluate uncertainty measures for landmark localization. This work extends the analysis of MHA in [7], with more

in depth experiments and comparisons. Our contributions, depicted in Fig. 1, are threefold:

- We propose Quantile Binning, a method to categorize predictions by any continuous uncertainty measure, and estimate error bounds for each bin (Fig. 1c, Sec. IV-C).
- We construct two evaluation metrics for uncertainty estimation methods from Quantile Binning: 1) Similarity between predicted bins and true error quantiles; 2) Accuracy of estimated error bounds (Fig. 1e, Sec. IV-D).
- We evaluate three heatmap-derived uncertainty measures and recommend our proposed method Ensemble Maximum Heatmap Activation (E-MHA) to extract landmark coordinates from an ensemble of heatmaps and estimate uncertainty (Fig. 1a, 1b, Sec. IV-B).

We demonstrate the impact of our contributions by using our proposed Quantile Binning to compare E-MHA to two existing coordinate extraction and uncertainty estimation methods: a weak baseline of Single Maximum Heatmap Activation (S-MHA), and a stronger baseline of Ensemble Coordinate Prediction Variance (E-CPV). In Sec. VI-B, we compare the baseline coordinate extraction performance of the three approaches, followed by the uncertainty estimation performance in Sec. VI-C. We explore the reach of heatmap-based uncertainty measures by demonstrating they are applicable to both U-Net regressed Gaussian heatmaps and patch-based voting heatmaps. We show each uncertainty measure can identify a subset of predictions with significantly lower mean error than the full set by filtering out predictions from high uncertainty bins (Fig. 1c). In Sec. VI-E we demonstrate the generalizability of our method by applying Quantile Binning to a publicly available Cephalometric dataset [8], with significantly more annotated landmarks and images containing some repetitive structures. We show the flexibility of our method by reporting results over a range of binning resolutions in Sec. VI-F. Furthermore, in Sec. VI-G we select a subset of landmarks

from the Cephalometric dataset with multiple annotations (provided by [12]) to explore the effect of aleatoric uncertainty caused by landmark ambiguity on Quantile Binning using our three uncertainty measures. Finally, in Sec. VII-B we make recommendations for which uncertainty measure to use, and how to use it.

We provide an open source implementation of this work and the tabular data obtained from the landmark localization models to reproduce our results, alongside extensive experimental results at `https://github.com/schobs/qbin`.

IV. METHODS

A. Landmark Localization Models

First, we briefly review the two models we use for landmark localization, allowing us to compare the generalizability of our uncertainty measures across different heatmap generation approaches. We implement a variation of the popular encoder-decoder networks that regresses Gaussian heatmaps, U-Net [6]. We also implement a patch-based method, PHD-Net [7], which produces a heatmap from patch votes.

1) Encoder-Decoder Model (U-Net): The vast majority of state-of-the-art landmark localization approaches are based on the foundation of a U-Net style encoder-decoder architecture. The architecture of U-Net follows a "U" shape, first extracting features at several downsampled resolutions, before rebuilding to the original dimensionality in a symmetrical upsampling path. Skip connections are employed between each level, preserving spatial information. The rationale behind the architecture design is to inject some inductive bias into the model architecture itself, helping it learn the local characteristics of each landmark, while preserving the global context.

Rather than regressing coordinates directly, the objective of the model is to learn a Gaussian heatmap image for each landmark, with the centre of the heatmap on the target landmark. For each landmark L_i with 2D coordinate position $\widetilde{\mathbf{x}}_i$, the 2D heatmap image is defined as the 2D Gaussian function:

$$g_i\left(\mathbf{x} \,||\, \boldsymbol{\mu} = \widetilde{\mathbf{x}}_i; \sigma\right) = \frac{1}{(2\pi)\sigma^2} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}\|_2^2}{2\sigma^2}\right), \quad (1)$$

where $\mathbf{x}$ is the 2D coordinate vector of each pixel and σ is a user-defined standard deviation. The network learns weights $\mathbf{w}$ and biases $\mathbf{b}$ to predict the heatmap $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$. During inference, we can interpret the activation of each pixel in the predicted heatmap as the pseudo-probability of that pixel being the landmark. We will exploit this in our uncertainty estimation methods.

2) Patch-based Model (PHD-Net): Patch-based models use a Fully Convolutional Network (FCN), with the architecture resembling the first half of an encoder-decoder architecture. Therefore, they are more light-weight than encoder-decoder networks, with significantly less parameters leading to faster training.

In our earlier work, we proposed PHD-Net: a multi-task patch-based network [7], building on the work by Noothout *et al.* [18]. We incorporate a variant of the heatmap objective function from encoder-decoder networks into the objective function, predicting the 2D displacement from each patch to the landmark alongside the coarse Gaussian pseudo-probability of each patch.

PHD-Net aggregates the patch-wise predictions to obtain a heatmap by plotting candidate predictions from the displacement branch as small Gaussian blobs, then regularising the map by the upsampled Gaussian from the heatmap branch.

Again, we can consider the activation of each pixel in heatmap as an indicator for uncertainty, where instead of the pseudo-probability, the activation represents the amount of "patch votes".

3) Ensemble Models: Using an ensemble of identical but randomly initialized models is more robust than using a single model, as it reduces the effect of a single model becoming stuck in a local minima during training. Furthermore, random initializations explore different modes of the loss landscape, facilitating a powerful decorrelation effect between models [26]. We use the variance in the predictions of each model to estimate the uncertainty of the prediction, using an use an ensemble of T models.

B. Estimating Uncertainty and Coordinate Extraction

Although generated differently, we hypothesize both U-Net and PHD-Net produce heatmaps containing useful information to quantify a prediction's uncertainty - but are they equally effective? To this end, we compare the performance of both models under three uncertainty estimation methods: two baseline approaches, and a proposed approach extending one of the baselines to an ensemble of networks. Each method extracts coordinate values from the predicted heatmap, and estimates the prediction's uncertainty.

1) Single Maximum Heatmap Activation (S-MHA): We introduce the baseline coordinate extraction and uncertainty measure. We use the standard method to obtain the predicted landmark's coordinates $\widehat{\mathbf{x}}_i$ from the predicted heatmap $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$, by finding the pixel with the highest activation:

$$\widehat{\mathbf{x}}_i = \arg\max_{\mathbf{x}} h_i(\mathbf{x}; \mathbf{w}, \mathbf{b}). \quad (2)$$

We hypothesize that the pixel activation at the coordinates $\widehat{\mathbf{x}}_i$ can describe the model's uncertainty: the higher the activation, the lower the uncertainty, and the lower the prediction error. However, due to this inverse relationship, this measures "confidence", not uncertainty.

We transform our confidence metric to an "uncertainty" metric $\widehat{y}_i$, by applying the following transformation to the pixel activation at the predicted landmark location:

$$\widehat{y}_i = \frac{1}{\max\limits_{\mathbf{x}} h_i(\mathbf{x}; \mathbf{w}, \mathbf{b}) + \epsilon}, \quad (3)$$

where ϵ is a small constant scalar that prevents $\frac{1}{0}$. Now, as the pixel activation at $\widehat{\mathbf{x}}_i$ increases, $\widehat{y}_i$ decreases.

We call the transformed activation of this peak pixel Single Maximum Heatmap Activation (S-MHA). This is a continuous value bounded between $[\frac{1}{\epsilon}, \frac{1}{1+\epsilon}]$ for U-Net, and bounded between $[\frac{1}{\epsilon}, \frac{1}{N+\epsilon}]$ for PHD-Net, where N is the number of patches. The lower the S-MHA, the lower the uncertainty.

2) Ensemble Maximum Heatmap Activation (E-MHA): In this work we extend the S-MHA uncertainty measure to ensemble models. We hypothesize that E-MHA should hold a stronger correlation with error than S-MHA due to the additional robustness an ensemble of models affords. We generate the mean heatmap of the T models in the ensemble, and obtain the predicted landmark coordinates as the pixel with the highest activation:

$$\widehat{\mathbf{x}}_i = \arg\max_{\mathbf{x}} \frac{1}{T} \sum_{t=1}^{T} h_i^t(\mathbf{x}; \mathbf{w}, \mathbf{b}). \tag{4}$$

Using the average prediction of an ensemble is the simplest, low-cost, standard form of ensemble fusion [21], [22], [27], [28]. Again, we hypothesize the activation of the pixel $\widehat{\mathbf{x}}_i$ correlates with model confidence. Similar to S-MHA, we inverse the pixel activation and add a small ϵ to the activation of $\widehat{\mathbf{x}}_i$ to give us our uncertainty measure, $\widehat{y}_i$:

$$\widehat{y}_i = \frac{1}{\left(\max_{\mathbf{x}} \frac{1}{T} \sum_{t=1}^{T} h_i^t(\mathbf{x}; \mathbf{w}, \mathbf{b})\right) + \epsilon}. \tag{5}$$

E-MHA is a continuous value constrained to the same bounds as S-MHA. This is a form of late feature fusion, combining features from all models before a decision is made.

3) Ensemble Coordinate Prediction Variance (E-CPV): We also implement an additional strong baseline for uncertainty estimation: Ensemble Coordinate Prediction Variance (E-CPV) [5]. The more the models disagree on where the landmark is, the higher the uncertainty.

To extract a landmark's coordinates we first use the traditional S-MHA coordinate extraction method on each of the T models' predicted heatmaps. Then, we use decision-level fusion to calculate the mean coordinate of the individual predictions to compute the final coordinate predictions $\widehat{\mathbf{x}}_i$:

$$\widehat{\mathbf{x}}_i = \frac{1}{T} \sum_{t=1}^{T} \arg\max_{\mathbf{x}} h_i^t(\mathbf{x}; \mathbf{w}, \mathbf{b}). \tag{6}$$

We generate the E-CPV by calculating the mean absolute difference between the T model predictions $\left(\widehat{\mathbf{x}}_i^1 \text{ to } \widehat{\mathbf{x}}_i^T\right)$ and $\widehat{\mathbf{x}}_i$:

$$\widehat{y}_i = \frac{1}{T} \sum_{t=1}^{T} \left\| \widehat{\mathbf{x}}_i^t - \widehat{\mathbf{x}}_i \right\|. \tag{7}$$

This is a continuous value bounded between $[0, \sqrt{H^2 + W^2}]$, where H and W are the height and width of the original image, respectively. The more the models disagree on the landmark location, the higher the coordinate prediction variance, and the higher the uncertainty.

Unlike S-MHA and E-MHA, this metric completely ignores the value of the heatmap activations. Therefore, it potentially loses useful uncertainty information but avoids possible bias caused by model miscalibration [30] or the Gaussian assumptions of the target heatmap.

C. Quantile Binning: Categorising Predictions by Uncertainty and Estimating Error Bounds

We leverage the described uncertainty measures to inform the epistemic, subject-level uncertainty of any given prediction, i.e. *is the model's prediction likely to be accurate,*

or inaccurate based on this uncertainty value? We propose a data-driven approach, Quantile Binning, using a hold-out validation set to establish thresholds delineating varying levels of uncertainty specific to each trained model. We use these learned thresholds to categorize our predictions into bins and estimate error bounds for each bin. We opt for a data-driven approach over using static, pre-defined thresholds to increase robustness. For example, two identical models with randomly initialized weights trained on the same training set will converge to different modes [26], with a different distribution of MHA on the same test set. Furthermore, the difficulty of the landmark will also influence the characteristics of the resulting localization model as well as the distribution of the uncertainty measures. Therefore, establishing a set of thresholds for each model is more invariant to training differences compared to using the same thresholds for all models.

Quantile Binning is application agnostic; applicable to any data as long as it consists of continuous tuples of `<Uncertainty Measure, Evaluation Metric>`. In this context, a continuous tuple is a pair of continuous variables output by the prediction model, relating to a single sample.

In this paper, we generate these pairings after the landmark localization model is trained. We use a hold-out validation set and make coordinate predictions and uncertainty estimates using each of our three uncertainty measures described in Section IV-B. Since we have the ground truth annotations of the validation set we can produce continuous `<Uncertainty Measure, Localization Error>` tuples for each uncertainty measure.

1) Establishing Quantile Thresholds: We aim to categorize predictions using our continuous uncertainty metrics into Q bins. We make the following assumption: *The true function between a good uncertainty measure and localization error is monotonically increasing (i.e. the higher the uncertainty, the higher the error).*

Quantile binning is a non-parametric method that fits well with these assumptions - a variant of histogram binning which is commonly used for calibration of predictive models [30], [31]. By considering the data in quantiles rather than intervals, we can better capture a skewed distribution as the outliers in the tail of the distribution can be grouped into the same group. In other words, *quantiles divide the probability distribution into areas of approximately equal probability.*

This property allows us to interrogate model-specific (epistemic) uncertainties. Rather than compute uncertainty thresholds based on predefined error thresholds for each bin, we use Quantile Binning to create thresholds that group our samples in relative terms. This enables the user to flag the worst $X\%$ of predictions. We describe the steps below.

First, for any given uncertainty measure we sort our validation set `<Uncertainty Measure, Localization Error>` tuples in ascending order of their uncertainty value and sequentially group them into Q equal-sized bins $B_1, ..., B_Q$. We assign each bin B_q a pair of boundaries defined by the uncertainty values of the tuples at the edges of the bin to create an interval: $[\alpha_{q-1}, \alpha_q)$. To capture all predictions at the tail ends of the distribution, we set $\alpha_0 = 0$, and $\alpha_Q = \infty$. During inference, we use these boundaries to bin our

predictions into Q bins ($B_1...B_Q$), with uncertainty increasing with each bin. For each predicted landmark $\widehat{\mathbf{x}}_i$ with uncertainty $\widehat{y}_i$ where $\alpha_{q-1} \leq \widehat{y}_i < \alpha_q$, $\widehat{\mathbf{x}}_i$ is binned into B_q. As long as the validation set is representative of the true distribution, the distribution of samples should be uniform across the bins due to the quantile method we used to obtain thresholds.

The higher the value of Q, the more fine-grained we can categorize our uncertainty estimates. However, as Q increases the method becomes more sensitive to any noise present in the uncertainty measure, leading to less accurate prediction binnings. We demonstrate this trade-off in Sec. VI-F.

Since the uncertainty boundaries are defined by the density of the validation set distribution, the method is agnostic to the absolute range of the uncertainty measure. Therefore it is applicable to any continuous uncertainty measure.

2) Estimating Error Bounds using Isotonic Regression: Establishing thresholds has allowed us to filter predictions by uncertainty in relative terms, but we lack a method to estimate absolute localization error for each bin. For example, for an easy landmark, the samples in B_1 may have a very low localization error in absolute terms, but for a more difficult landmark even the lowest relative uncertainty samples in B_1 may have a high error. Therefore, in order to offer users information about the expected error for each group, we present a data-driven approach to predict error bounds.

A simple approach would be to observe the localization error of the tuple at the quantile boundaries [α_{q-1} and α_q). However, observing a single sample from the validation set is subject to noise and may produce a poor estimate for an error bound. Therefore, on our hold-out validation set, we first use Isotonic Regression to approximate the function between uncertainty and error, constraining it to be monotonically increasing. Isotonic regression is a method to fit a free-form, non-decreasing line to a set of observations, also commonly used for predictive model calibration [30], [32]. It is non-parametric, so can learn the true distribution if given enough i.i.d. data. The regression seeks a weighted least squares fit $\widehat{\beta}_i \approx \beta_i$ subject to the constraint that $\widehat{\beta}_i \leq \widehat{\beta}_j$ whenever $\eta_i \leq \eta_j$:

$$\min \sum_{i=1}^{n} \left(\widehat{\beta}_i - \beta_i\right)^2 \text{ s.t. } \widehat{\beta}_i \leq \widehat{\beta}_j \text{ for all } (i,j) \in E, \quad (8)$$

where $E = \{(i,j) : \eta_i \leq \eta_j\}$ and n is the number of (η_i, β_i) pairs. In our case, the observations $(\eta_1, \beta_1), \ldots, (\eta_n, \beta_n)$, are the $(UncertaintyMeasure, LocalizationError)$ tuples.

Next, we use our isotonically regressed line to estimate error bounds for each of our quantile bins. We input each bin's threshold intervals [α_{q-1}, α_q) into our fitted Isotonic Regression function, obtaining error predictions for each threshold, [γ_{q-1}, γ_q). We use these values as the estimated lower and upper error bounds, respectively, of the predictions in bin B_q. Note, that for B_1 we only estimate an upper bound, and for B_Q we only estimate a lower bound.

In summary, we use a data-driven approach to learn thresholds to progressively filter predictions at inference into Q bins of increasing uncertainty, and assign each bin estimated error bounds.

| (a) SA CMR. | (b) 4CH CMR. | (c) Cephalometric. |
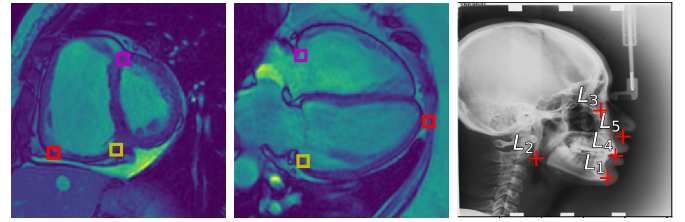
Fig. 2: **(a)** Landmarks for Short Axis (SA) CMR: Magenta = superior right ventricle insertion point valve; Yellow = inferior right ventricle insertion point; Red = inferior lateral reflection of right ventricle free wall. **(b)** Landmarks for 4 chamber (4CH) CMR: Magenta = tricuspid valve; Yellow = mitral valve; Red = apex of left ventricle. **(c)** Subset of Landmarks included in the Cephalometric dataset [8]. Displayed landmarks are used in the aleatoric uncertainty analysis (Sec. VI-G).

D. Evaluation Metrics for Uncertainty Measures

Next, we construct methods to evaluate how well an uncertainty measure's predicted bins represent the true error quantiles, and how accurate each bin's estimated error bounds are.

1) Evaluating the Predicted Bins: A good uncertainty measure will have a strong correlation with localization error. Therefore, it should provide quantile thresholds that correspond to the true error quantiles. For example, since Bin B_1 contains the predictions with the uncertainties at the lowest $\frac{1}{Q}$ quantile, the localization errors of the predictions in B_1 should be the lowest $\frac{1}{Q}$ quantile of the test set. This can be generalized to each group, until B_Q, which should contain the errors in the $\frac{Q-1}{Q}$ quantile.

To evaluate this desired property, we propose to measure the similarity between each predicted bin and its respective theoretically perfect bin.

We create the ground truth (GT) bins by ordering the test set samples in ascending order of error. Then, we sequentially bin them into Q equally sized bins: $\widehat{B}_1...\widehat{B}_Q$.

For each predicted and GT bin pair B_q & $\widehat{B}_q$, we calculate the Jaccard Index (JI) between them and report the mean measure of each bin across all folds:

$$J_q(B_q, \widehat{B}_q) = \frac{|B_q \cap \widehat{B}_q|}{|B_q \cup \widehat{B}_q|}. \quad (9)$$

The higher the JI, the better the uncertainty measure has binned predictions by localization error. Therefore, it follows that the higher the JI, the better the uncertainty measure predicts localization error.

2) Accuracy of Estimated Error bounds: A good uncertainty measure will have a monotonically increasing relationship with localization error. Therefore, estimating the true function using isotonic regression should provide accurate error bound estimations.

To measure this, for each bin B_q, we calculate the percentage of predictions whose error falls between the estimated error bound interval, [γ_{q-1}, γ_q). The higher the percentage, the higher the accuracy of our estimated upper error bounds.

V. DATASETS

We perform our experiments using three datasets. The first two datasets are from the ASPIRE Registry [33], containing Cardiac Magnetic Resonance Imaging (CMR) sequences, from a 1.5 Tesla GE HDx (GE Healthcare, Milwaukee, USA) system using an eight-channel cardiac coil. Images were acquired using a cardiac-gated multislice balanced steady-state free precession sequence (20 frames per cardiac cycle, slice thickness 10mm, 0mm inter-slice gap, field of view 480mm, acquisition matrix 256×200, flip angle 60°, BW 125 KHz/pixel, TR/TE 3.7/1.6 ms). Each subject has a four chamber (4CH) view and/or a short axis view (SA). Each CMR sequence has a spatial resolution of 512×512 pixels, where each pixel represents 0.9375mm of the organ, and the first frame was used for landmark localization in this study. There are 303 SA images, each with three annotated landmarks: the inferior right ventricle insertion point (infSA), the superior right ventricle insertion point (supSA), and the inferior lateral reflection of the right ventricle free wall (RVSA). There are 422 4CH images, each with three annotated landmarks: the apex of the left ventricle at end diastole (LVDEV Apex), the mitral valve (mitral), and tricuspid valve (tricuspid). The 4CH dataset represents a more challenging landmark localization task as the images have much higher variability than the SA dataset. The landmarks were decided and manually labelled by a radiologist, as shown in Figs. 2a & 2b. For this study, we consider the SA images the **EASY** dataset, and the 4CH images the **HARD** dataset.

To test generalizablity across imaging modalities, we use a third dataset consisting of Cephalometric Radiographs, in which the images contain repetitive structures [8]. The dataset has a total of 19 annotated landmarks, where we use the junior annotator as the ground truth (following the convention of [10], [12], [34]). For our study of aleatoric uncertainty in Sec. VI-G, we use subset of 5 landmarks which have a total of 11 annotations provided by [12]. The images have a spatial resolution of 512×512 pixels, where each pixel represents 0.1mm of the structure. Fig. 2c shows an example image annotated with the aleatoric uncertainty landmark subset.

VI. EXPERIMENTS AND RESULTS

First, in Sec. VI-B we present the baseline landmark localization performance of PHD-Net and U-Net over both SA and 4CH datasets using the S-MHA, E-CPV, and E-MHA methods to extract coordinates. This gives us a comparison of the coordinate extraction performance from each of our methods, and a baseline to measure the effectiveness of each method's uncertainty estimation. Second, in Sec. VI-C we interrogate how using Quantile Binning with our uncertainty measures delineates predictions in terms of their localization error, and compare the predicted bins to the ground truth error quantiles. We show a practical example of how filtering out highly uncertain predictions can dramatically increase the proportion of acceptable localization predictions. In Sec. VI-D we assess how well the uncertainty measures can predict error bounds for each bin. Next, we demonstrate the generalizability of Quantile Binning in Sec. VI-E on the more diverse Cephalometric dataset. In Sec. VI-F, we highlight the flexibility of the method

TABLE I: Localization errors (mm) for the uncertainty methods outlined. *All* indicates entire set of predictions; B_1 indicates subset with the *lowest uncertainties*. Mean error and standard deviation are reported across all folds & all landmarks. **Bold** indicates best results in row for the given dataset.

Method	4 Chamber Images		Short Axis Images	
	U-Net	PHD-Net	U-Net	PHD-Net
S-MHA All	**10.00 ± 18.99**	11.07 ± 21.33	5.86 ± 14.19	**3.58 ± 3.52**
S-MHA B_1	6.79 ± 6.09	**5.80 ± 9.03**	3.62 ± 2.45	**2.78 ± 1.99**
E-MHA All	**6.36 ± 8.01**	9.14 ± 18.11	4.37 ± 8.86	**3.36 ± 3.50**
E-MHA B_1	4.93 ± 2.85	**4.70 ± 3.21**	2.98 ± 2.09	**2.39 ± 1.90**
E-CPV All	**8.13 ± 10.16**	9.42 ± 13.07	4.97 ± 7.51	**3.22 ± 2.93**
E-CPV B_1	5.34 ± 3.00	**5.10 ± 6.76**	3.75 ± 2.13	**2.47 ± 2.08**

by quantifying the effects of varying the number of quantile bins (Q) used. Finally, in Sec. VI-G we explore aleatoric uncertainty, demonstrating Quantile Binning's effectiveness on landmarks with high ambiguity, as well as sharing insights on our studied uncertainty measure's relationship with aleatoric uncertainty. When comparing between B_1, B_{2-4}, B_5 we use an unpaired t-test ($p \leq 0.05$) to test for significance. When comparing uncertainty metrics among the same Bin category and model, we use a paired t-test ($p \leq 0.05$) to test for significance.

A. Experimental Setup

We split both CMR datasets into 8 folds, and perform 8-fold cross validation for both U-Net and PHD-Net. For each of the eight iterations, we select one fold as our testing set, one our hold-out validation set and the remaining 6 as our training set. For the Cephalometric dataset we follow previous work ([10], [12], [34]) and perform 4-fold cross validation using junior annotations, setting aside a random 20% of each fold's training set as our hold-out validation-set. We select $T = 5$ for the ensemble methods, training 5 identical, randomly initialized models at each iteration. We chose $T = 5$ to compromise with computational constraints, asserting that 5 models are representative to compare the uncertainty methods for our purposes. We randomly select a model from our trained ensemble for our S-MHA uncertainty measure. For our Quantile Binning method, we select $Q = 5$ for 5 bins, striking a balance between the resolution of separation of the data and the limited size of our hold-out validation set (∼30 samples for the CMR datasets, ∼60 samples for the Cephalometric dataset). We explore the effect of changing Q in Sec. VI-F.

We implement a vanilla U-Net model [6]. We design the architecture with 5 encoding-decoding levels, creating 1.63M learnable parameters. Each level contains 2 residual units, where each residual unit applies a 3×3 convolution, instance normalization, and ReLU to the input, before concatenating the resulting output with the unit input. As we descend down the five levels of the encoder we use (16, 32, 64, 128, 256) input channels respective to each layer, mirroring this in the decoder path. On the encoder path we use maxpooling after each level to reduce spatial dimensions, and on the decoder path we use transposed convolutions to upsample the spatial resolution. We modify the objective function from image segmentation to simultaneous landmark localization, minimising the mean squared error between the target and predicted heatmaps. We

use the full 512×512 pixel image as input, and learn heatmaps of the same size. We train for 1000 epochs with a batch size of 2, and a learning rate of 0.001 using the Adam Optimizer (settings from [7]). We generate target heatmaps using Eq. (1) with a standard deviation of 8 for our CMR datasets and 2 for our Cephalometric dataset (chosen experimentally using the first fold of each dataset). We do not use data augmentation.

We implement our PHD-Net model following [7], creating a model with 0.06M learnable parameters. For all experiments we trained PHD-Net for 1000 epochs using a batch size of 32 and a learning rate of 0.001, using the Adam Optimizer. We train one landmark at a time. Note, the only difference in setup from [7] in this work is different fold splits and training for an additional 500 epochs (same as U-Net) with no early stopping, since we now use our validation set for Quantile Binning. We do not use data augmentation.

B. Baseline Landmark Localization Performance

Table I show the baseline performance for U-Net and PHD-Net at localizing landmarks in our 4CH and SA datasets. We make the following observations:

- When considering localization error for the entire set of landmarks (*All*), performance is better on the SA dataset for both models, with PHD-Net outperforming U-Net. On the 4CH dataset, U-Net outperforms PHD-Net in terms of fewer gross mispredictions, suggesting the higher capacity model of U-Net is more robust to datasets with large variations.
- Simply using a single model with our S-MHA strategy is predictably less robust than ensemble approaches.
- E-MHA outperforms the previous strong baseline of E-CPV for coordinate extraction. However, does it outperform E-CPV in terms of uncertainty estimation? We explore this in Section VI-C.
- The standard deviation in the error for the baseline *All* results in Table I is high for all models. We aspire to catch these bad predictions using Quantile Binning in Sec. VI-C.

C. Analysis of the Predicted Quantile Bins

We apply quantile binning to each uncertainty measure: S-MHA, E-MHA and E-CPV. We compare results over U-Net and PHD-Net for both the SA and 4CH datasets.

First, we test our assumption that our uncertainty measures correlate with localization error. We report the Spearman's Rank Correlation Coefficient (ρ) since we are measuring a monotonic correlation which is not necessarily linear. All correlations are reported from the aggregated test sets across all eight folds of our CMR datasets, using a significance level of $p < 0.001$. For our 4CH dataset, S-MHA achieves correlations of 0.33 (weak-moderate) & 0.47 (moderate), E-MHA shows weak-moderate correlations of 0.39 & 0.39, and E-CPV shows moderate correlations of 0.42 & 0.53; for U-Net and PHD-Net respectively. For our SA dataset, S-MHA achieves correlations of 0.27 (weak) & 0.33 (weak-moderate), E-MHA weak-moderate correlations of 0.38 & 0.38, and E-CPV correlations of 0.27 (weak), 0.36 (weak-moderate); for U-Net and PHD-Net, respectively. The correlation strength
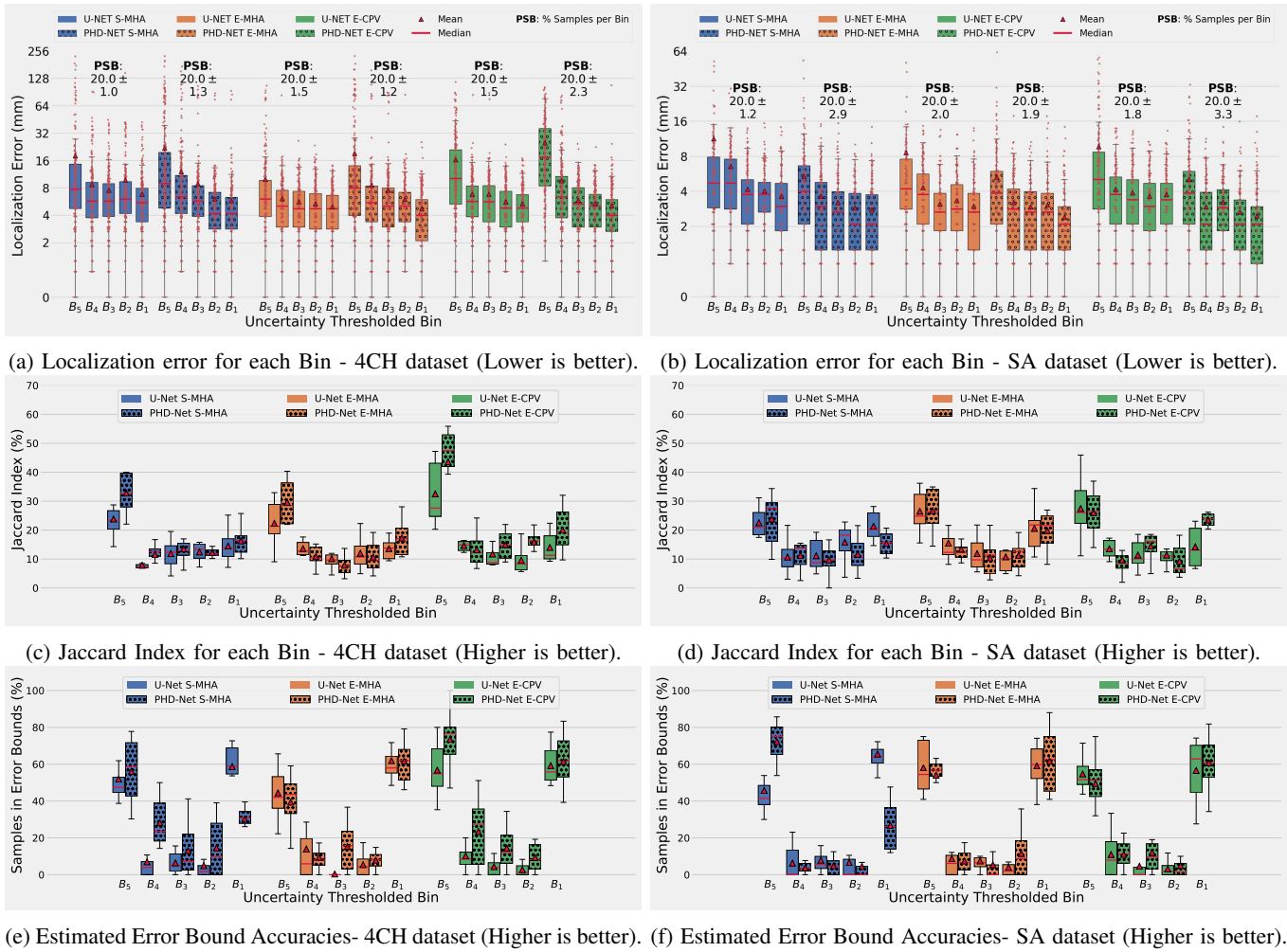
of S-MHA has high variance, whereas E-MHA shows a stable correlation across datasets and localization models. E-CPV achieves the strongest correlation with error across both models for our harder 4CH dataset, but a weaker correlation than E-MHA for our easier SA dataset. Overall, these results show that MHA and E-CPV contain information that can be exploited to estimate the uncertainty of our predictions.

Next, we compare how our uncertainty measures can predict the true error quantiles. We found the most useful information is at the tail ends of the uncertainty distributions. Figs. 3c & 3d plot the Jaccard Index between ground truth error quantiles and predicted error quantiles. We notice a parabolic trend, where the outer bins are closer to the true error quantiles than the middle bins. The highest uncertainty quantile bin (B_5) is significantly better at capturing the correct subset of predictions than the intermediate bins ($B_2 - B_4$). Similarly, in some cases the bin representing the lowest uncertainties (B_1) had a significantly higher Jaccard Index than the intermediate bins, but still lower than B_5. Figs. 3a & 3b show the mean error (▲) of the samples of each quantile bin over both datasets. The most significant reduction in localization error is from B_5 to B_4 for all uncertainty measures. The sample distribution over the bins, indicated by the red dots, confirms that B_5 captures more gross mispredictions than the remaining bins, particularly for the 4CH dataset. These findings suggest that most of the utility in the uncertainty measures investigated can be found at the tail ends of the scale. This is an intuitive finding, as the predictions in B_5 are *certainly uncertain*, and the predictions in B_1 are *certainly certain*. Figs. 3a & 3b show that each bin contains ~20% of the predictions, confirming our data-driven approach to setting uncertainty thresholds successfully approximates the true uncertainty distribution.

The worse trained the landmark localization model, the more useful the uncertainty measure. Table I shows the localization error of all methods, models and datasets for the entire set (*All*) and lowest uncertainty subset (B_1) of predictions. PHD-Net's baseline localization performance on the 4CH dataset was worse than U-Net. However, when we consider the lowest uncertainty subset of predictions (B_1), PHD-Net sees a 47% average reduction in error from all predictions (*All*), compared to U-Net's average reduction of 30%. Similarly, U-Net performed worse than PHD-Net for the SA dataset, but saw an average error reduction of 31% compared to PHD-Net's 25%. This suggests that all investigated uncertainty measures are more effective at identifying gross mispredictions when models are poorly trained.

Using heatmap-based uncertainty measures is generalizable across heatmap generation approaches. The bin similarities in Figs. 3c & 3d show that using S-MHA and E-MHA yields similar performance with PHD-Net and U-Net, despite their different heatmap derivations. Surprisingly using E-MHA does not give a significant increase in bin similarity compared to S-MHA, suggesting the thresholds remain relatively stable across models.

No investigated method is conclusively best for estimating uncertainty in all scenarios. For the more challenging 4CH data, Fig. 3c shows E-CPV is significantly better than S-MHA and E-MHA for both models at capturing the true

(a) Localization error for each Bin - 4CH dataset (Lower is better).

(b) Localization error for each Bin - SA dataset (Lower is better).

(c) Jaccard Index for each Bin - 4CH dataset (Higher is better).

(d) Jaccard Index for each Bin - SA dataset (Higher is better).

(e) Estimated Error Bound Accuracies- 4CH dataset (Higher is better). (f) Estimated Error Bound Accuracies- SA dataset (Higher is better).

Fig. 3: Results from Quantile Binning for U-Net and PHD-Net across all landmarks & folds, using our three coordinate extraction & uncertainty estimation methods. Bins are in descending order of uncertainty (B_5 highest uncertainty, B_1 lowest uncertainty). (a) and (b) show the mean localization error of each bin, with error decreasing as we move towards the bins with lower uncertainty. (c) and (d) present the Jaccard Index, showing how similar the predicted bins are to the ground truth error quantiles. (e) and (f) visualize the estimated error bound accuracy, showing the percentage of predictions within the estimated error bounds for each bin. Best viewed on screen.

error quantiles, corroborating the findings of [5]. E-CPV is particularly good at identifying the worst predictions (B_5). For the easier SA data, no method has a significantly higher Jaccard Index. Therefore, when we generalize across both models and datasets, all uncertainty measures fared broadly similar on average in terms of error reduction between the entire set and the B_1 subset of predictions. S-MHA had an average error reduction of 35.07%, E-MHA 32.94% and E-CPV 32%.

Despite similar performances in uncertainty estimation, we found E-MHA yields the greatest localization performance overall. Table I shows E-MHA offers the best localization performance for B_1 across both datasets and models. This is due to the combination of offering the most robust coordinate extraction on average (Table. I), and similar uncertainty estimation performance (Fig. 3c, Fig. 3d). We more concretely demonstrate Quantile Binning's ability to identify low uncertainty predictions in Fig. 4. We clearly observe a significant increase in the percentage of images below the acceptable error

threshold of 5mm when considering only predictions in B_1 - with E-MHA giving the greatest proportion of acceptable predictions.

D. Analysis of Error Bound Estimation

We analyse how accurate the isotonically regressed estimated error bounds are for our quantile bins. Figs. 3e & 3f show the percentage of samples in each bin that fall between the estimated error bounds.

We found we can predict the error bounds for the two extreme bins better than the intermediate bins. Figs. 3e & 3f show a similar parabolic pattern to the Jaccard Index Figs. 3c & 3d, with the two extreme bins B_5 and B_1 predicting error bounds significantly more accurately than the inner bins. Again, this indicates the most useful uncertainty information is present at the extremes of the uncertainty distribution, with the predicted uncertainty-error function unable to capture a consistent relationship for the inner quantiles. Further, the increased accuracy of the outer bins can be explained by the fact that it is easier to predict a single lower/upper bound than

(a) PHD-Net - 4CH Images. (b) U-Net - 4CH Images. (c) PHD-Net - SA Images. (d) U-Net - SA Images.
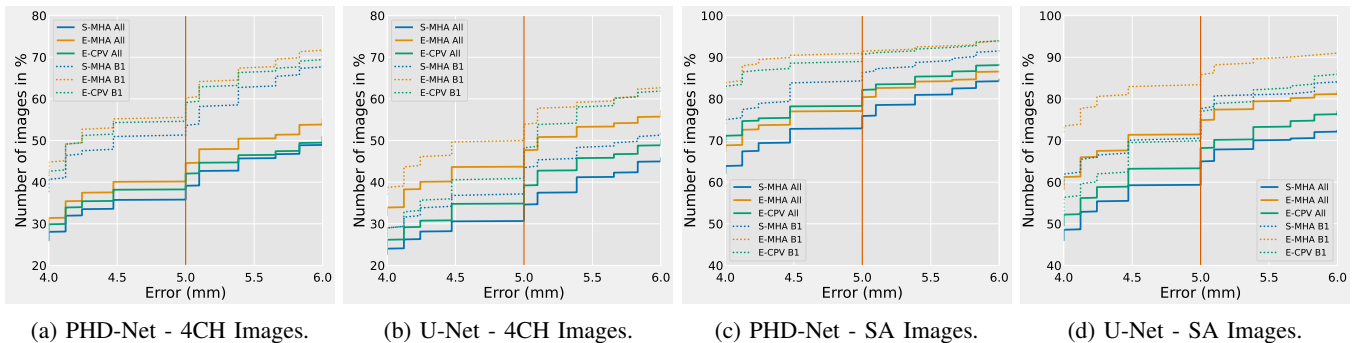
Fig. 4: Cumulative distribution of localization errors showing the % of predictions under a given error threshold, comparing all predictions (*All*) to the lowest uncertainty subset (B_1) for the uncertainty methods across all folds & landmarks. The vertical line is the acceptable error threshold, chosen by a radiologist. Higher percentage is better.

a pair of tighter bounds for the middling bins.

We also found that a well defined upper bound for heatmap activations is important for error bound estimates. For both the 4CH and SA datasets, S-MHA for PHD-Net is significantly more accurate at predicting error bounds for the highest uncertainty quantile B_5 compared to the lowest uncertainty quantile B_1 (56% & 72% compared to 30% & 27% for 4CH & SA, respectively), correlating with S-MHA capturing a greater proportion of those bins (Jaccard Indexes of 32% & 24% compared to 16% & 15%). On the other hand, U-Net using S-MHA predicts error bounds for low uncertainty bins better than high uncertainty bins. This suggests that although PHD-Net's heatmap activation is a robust predictor of gross mispredictions, the less tight upper bound of its heatmap activations make it hard to make an accurate prediction for the lowest uncertainty quantile (B_1). This is alleviated by using an ensemble of networks in E-MHA, where the B_1 bound accuracy is improved to 62%.

E-MHA and E-CPV are more consistent than S-MHA. Overall, there is no significant difference between the error bound estimation accuracy of E-MHA and S-MHA, but Figs. 3e & 3f show E-MHA has less variation in performance between U-Net and PHD-Net compared to S-MHA, suggesting an ensemble of models is more robust. For the 4CH dataset, PHD-Net using E-CPV is on average significantly more accurate at predicting error bounds than S-MHA and E-MHA. However, there are no significant differences for PHD-Net on the easier SA dataset, nor U-Net on either dataset. There are also no significant differences between U-Net and PHD-Net in error bound estimation accuracy, with each method broadly equally effective for both models.

E. Generalizability

We train U-Net on the Cephalometric dataset, applying Quantile Binning to E-MHA and E-CPV to test their generalizability across imaging modalities. For $Q = 5$, Figs. 5c & 5d show a predictive power of true error quantiles comparable with the CMR datasets. The mean Jaccard Index (JI) for B_5 is 22% for E-MHA and 34% for E-CPV on the Cephalometric dataset, compared to 22% & 32% for U-Net on the 4CH dataset. B_1 shows a better result than the CMR datasets, achieving a JI of 18% for E-MHA and 19% for E-CPV, compared to 15% & 14% for U-Net on the 4CH dataset. E-CPV more effectively identifies the extreme mis-predictions

compared to E-MHA, as evidenced by a higher JI for B_5 (left-most bin) in Figs. 5c & 5d, supporting the results from the challenging 4CH dataset. For $Q = 5$, Figs. 5a and 5b show a gradual reduction in error from B_5 (left-most) to B_1. Overall, the larger Cephalometric dataset (19 landmarks) shows a more consistent downward trend in error across bins compared to our smaller CMR datasets (3 landmarks).

Next, to test the robustness of using MHA as an uncertainty measure across target heatmaps of varying sizes, we repeated these experiments changing the standard deviation of the target heatmap from Eq. (1) to 2, 4, 8 and 12. We found the trends of our Quantile Binning results hold, with only the localization error deteriorating as we increased the size of the Gaussian. We conclude that as long as the standard deviation leads to a learnable heatmap, similar uncertainty estimation properties are exhibited by MHA.

### F. Varying Quantile Binning Resolution

We vary the number of Quantile Bins ($Q = \{2, 3, 5, 10, 20\}$) for the larger Cephalometric dataset to gain deeper insights on the flexibility of Quantile Binning. Figs. 5a & 5b show the localization error quantiles across Q for the Cephalometric dataset, with a gradual reduction in mean localization error (▲) from B_Q to B_1 for all values of Q. We find that the edge bins are most useful for all values of Q, with the Jaccard Indexes in Figs. 5c & 5d and error bound accuracies in Figs. 5e & 5f showing parabolic trends, confirming our results from the CMR datasets.

Further, Quantile Binning provides utility for a range of Q values. First, consider the extreme case of $Q = 2$, where the threshold is the median uncertainty of the validation set. Here, B_2 (the high uncertainty bin, left) captures the majority of the gross mispredictions and B_1 (the low uncertainty bin, right) captures the majority of the best predictions. Now, consider the effect of increasing Q, shown in Figs. 5a & 5b. As we increase the number of Quantile Bins, the mean error (▲) of B_Q (far left bin of each set) increases. This is because as Q increases, B_Q is pushed farther towards the edge of the uncertainty measure distribution, capturing progressively more extreme outliers. Therefore, as Q increases, we observe an increasingly logarithmic trend of the mean error across the bins as poor predictions are filtered out more gradually.
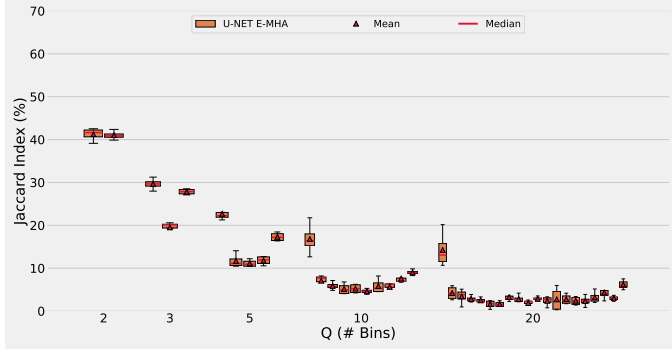
In practice, the higher the value of Q, the greater the resolution of separation of the data. For example, consider
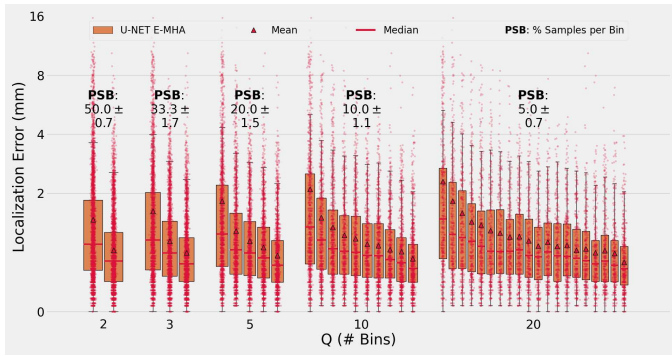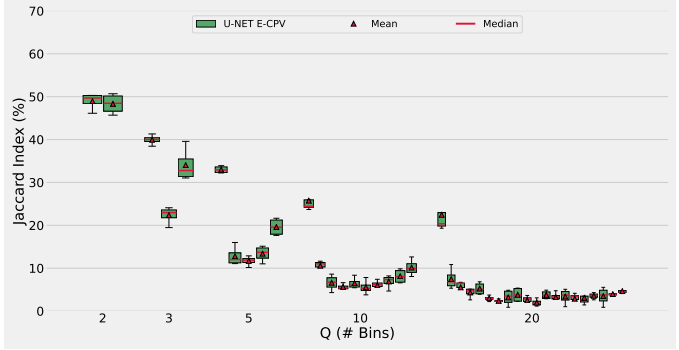
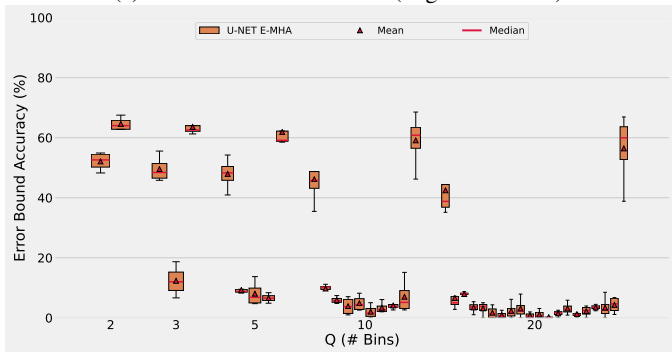(a) Localization error - E-MHA (Lower is better).
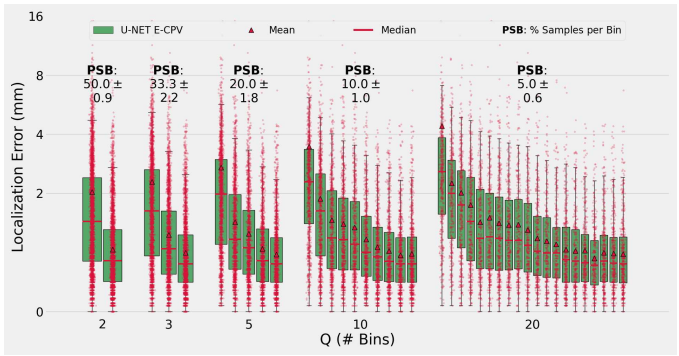
(b) Localization error - E-CPV (Lower is better).
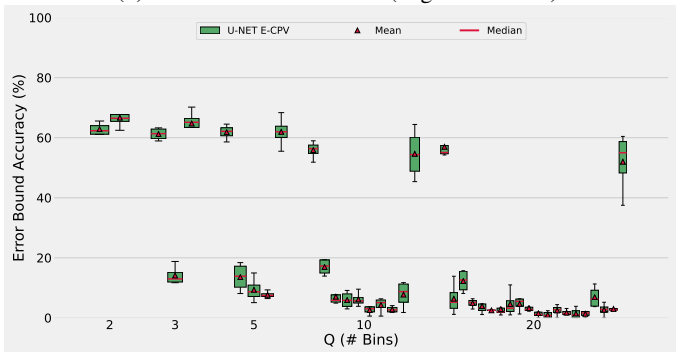

(c) Jaccard Index - E-MHA (Higher is better).

(d) Jaccard Index - E-CPV (Higher is better).

(e) Estimated Error Bound Accuracies - E-MHA (Higher is better).

(f) Estimated Error Bound Accuracies - E-CPV (Higher is better).

Fig. 5: Quantile Binning varying Q (Number of Quantile Bins) on the Cephalometric dataset. We show results for the uncertainty measures E-MHA and E-CPV, over all landmarks from a 4-fold CV, trained on the U-Net model. Red dots represent the errors of individual samples, best viewed on screen.

the task of flagging up uncertain landmark predictions for manual review. Using $Q = 2$ and flagging predictions from the highest uncertainty bin will lead to 50% of predictions requiring review and re-annotation. On the other hand, filtering out the highest uncertainty bin using $Q = 10$ leaves only 10% of predictions to be reviewed. In each case, the user will have an upper error bound estimate for the remaining predictions with reasonable accuracy (∼50% for E-MHA and ∼60% for E-CPV, the left-most bins, B_Q, in Figs. 5e & 5f). However, the contents of B_Q are more accurate when Q is small, with a Jaccard Index of 50% for $Q = 2$ compared to 40% for $Q = 10$ for E-CPV (Fig. 5d). Therefore, this trade-off between true error quantile accuracy and binning resolution means Q is a subjective choice that depends on the specificity of the downstream task and the resources available for reannotation.

Similar trends are present for our 4CH dataset and SA dataset, but we note that results are poor for $Q >= 10$

compared to the Cephalometric dataset. This is because when fitting the data for Quantile Binning, our CMR datasets had access to a much smaller validation set compared to the Cephalometric dataset (∼30 samples compared to ∼60 samples) and could not accurately estimate the quantile uncertainty distribution for large values of Q. Therefore, the larger the available validation set, the larger Q can be set.

G. Relationship with Aleatoric Uncertainty

Lastly, we study aleatoric uncertainty, which refers to uncertainty caused by internal randomness in the data. Using Quantile Binning, we explore how our epistemic uncertainty measures deal with landmarks of varying levels of aleatoric uncertainty. In landmark localization, one way to measure aleatoric uncertainty is from the inherent ambiguity of the landmark, quantified by the inter-observer variability of multiple annotators. We can infer that the higher the variation in annotator opinion, the greater the ambiguity of the landmark.
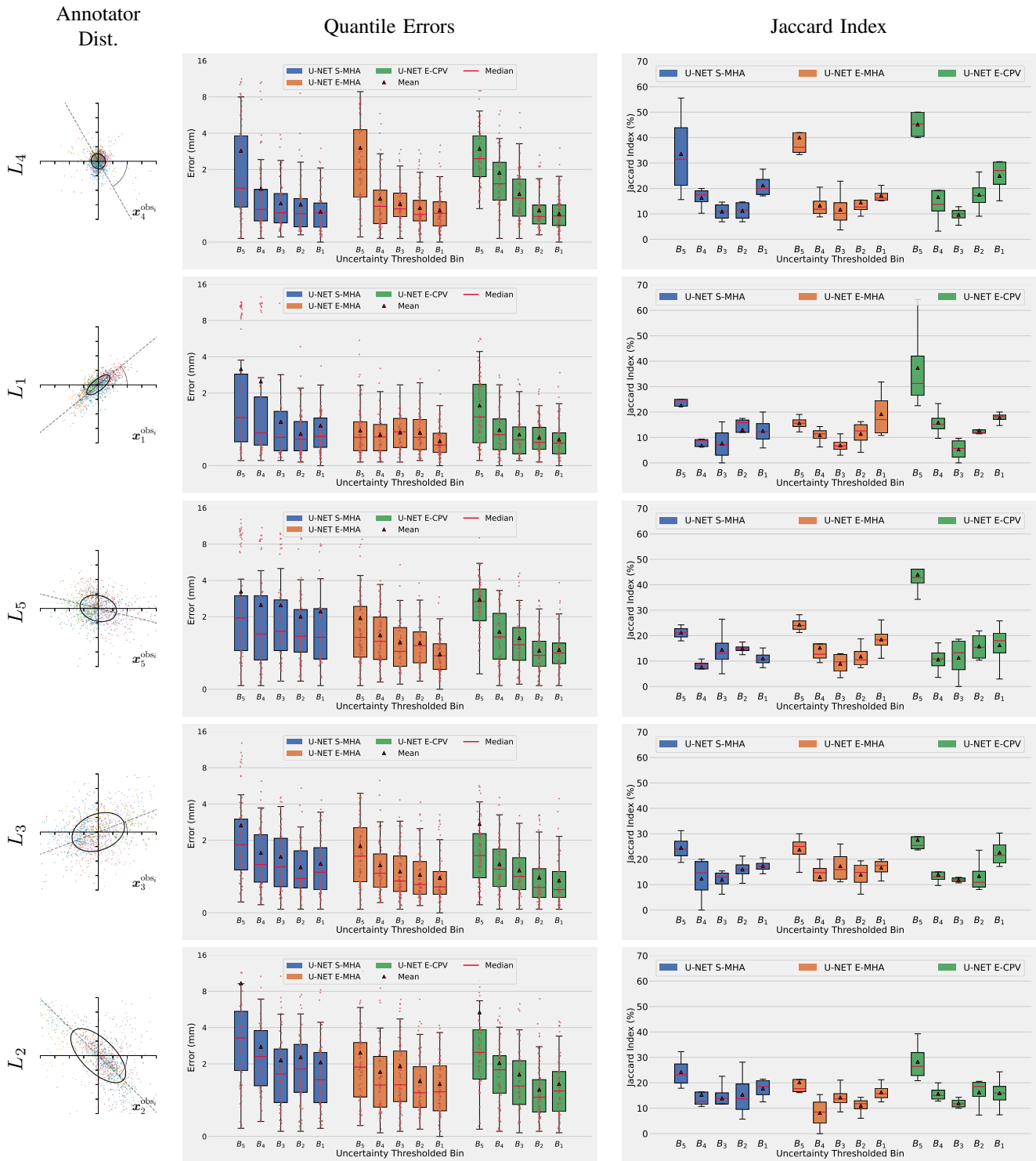
Fig. 6: Column *Annotator Dist.* shows the individual offsets from each of the 11 annotators to the mean annotation of each landmark [12]. The larger the fitted Gaussian, the more variance between annotators and the higher the aleatoric uncertainty. *Quantile Errors* column shows the boxplots of localization errors for each quantile bin, showing the landmarks across all folds. The *Jaccard Index* column shows the similarity between the predicted Quantiles and the true error quantiles.

We can observe the directional ambiguity of the landmark by fitting an anistotropic (directionally skewed) Gaussian function to the distribution of the annotations, seen in the *Annotator Dist.* column of Fig. 6. Thaler *et al.* [12] provide this ground truth measure of the aleatoric uncertainty, using a total of 11 annotators to label a subset of five landmarks (Fig. 2c)

from 100 images of the Cephalometric dataset. We assume the landmark-specific ambiguities hold for the full Cephalometric dataset.

Fig. 6 shows that all studied coordinate extraction methods are best for landmarks with low aleatoric uncertainty. The mean errors (▲) over the boxplots in the *Quantile Errors*

rows in Fig. 6 confirm that landmarks with higher aleatoric uncertainty (L_3, L_2) had worse localization performance than landmarks with low aleatoric uncertainty (L_4, L_1). However, the distribution of individual samples (represented by red dots, best seen on screen) show that E-MHA and E-CPV reliably capture the majority of gross mispredictions (B_5) regardless of landmark ambiguity. S-MHA performs poorly on some landmarks (L_5, L_2) due to the reliance on a single model capacity. In terms of filtering out poor predictions, we see the best results for all uncertainty methods for the landmark with the tightest annotation distribution (L_4), with B_5 Jaccard Index's showing a mean of 40% and 45% similarity with the true quantile bin for E-MHA and E-CPV, respectively.

However, MHA methods falter for landmarks with directional ambiguity, whereas E-CPV estimates uncertainty well for all types of ambiguity. The *Annotator Dist.* column of Fig. 6 shows that the annotation distribution of landmarks L_1 and L_2 have a distinct directional skew. The Jaccard Indexes of E-MHA for these landmarks ($L_1 = 16\%$, $L_2 = 20\%$ for B_5) are lower than the other landmarks with more isotropic annotation distributions ($L_4 = 40\%$, $L_5 = 24\%$, $L_3 = 23\%$ for B_5). Furthermore, the mean and median localization errors do not consistently trend down across bins for E-MHA for the anisotropic landmarks (L_1, L_2).

On the other hand, Quantile Binning shows E-CPV is consistently effective regardless of directional ambiguity, with mean Jaccard Indexes for B_5 no less than 28% across all landmarks. This is likely because the objective function (Eq. (1)) encourages the model to predict isotropic Gaussian Heatmaps, which better match isotropic annotator distributions. When we calculate the mean heatmap to extract the peak pixel using E-MHA, the resulting map will still be constrained to the isotropic properties defined by the objective function. This explains why E-MHA even performs well on the ambiguous yet isotropic landmarks L_3 and L_5, but poorly on the directionally ambiguous, anisotropic landmarks L_1 and L_2. In contrast, E-CPV calculates the variance between peak pixel activations of a group of individual models, where sampling enough independent predictions of an ensemble can effectively approximate the anisotropic distribution. In practice, if the Quantile Error bins for E-MHA show uniformity as they do in L_1 and L_2, this is an indication to the user that the landmark may contain some directional ambiguity.

VII. DISCUSSION AND CONCLUSION

A. Summary of Findings

This paper presented a general framework to assess any continuous uncertainty measure in landmark localization, demonstrating its use on three uncertainty metrics and two paradigms of landmark localization model. We introduced a new coordinate extraction and uncertainty estimation method, E-MHA, offering the best baseline localization performance and competitive uncertainty estimation.

Our experiments indicate that both heatmap-based uncertainty metrics (S-MHA, E-MHA), as well as the strong baseline of coordinate variance uncertainty metric (E-CPV) are applicable to both U-Net and PHD-Net. Despite the two models' distinctly different approaches to generating heatmaps,

using the maximum heatmap activation as an indicator for uncertainty is effective for both models. We showed that all investigated uncertainty metrics were effective at filtering out the gross mispredictions (B_Q) and identifying the most certain predictions (B_1), but struggled to capture useful information for the intermediate uncertainty bins (B_2-B_{Q-1}).

Our experiments also showed that E-MHA and S-MHA had a surprisingly similar ability to capture the true error quantiles of the best and worst 20% of predictions (Figs. 3c & 3d), but E-MHA was more consistent with its performance predicting the error bounds of those bins across models (Figs. 3e & 3f). This suggests that the correlation with localization error at the head and tail ends of the heatmap distributions are stable across our ensemble of models, but susceptible to variance when fitting our isotonically regressed line to predict error bounds. On the more challenging 4CH dataset, E-CPV broadly remained the strongest method for filtering out the worst predictions, but this trend did not continue in the easier SA dataset (Fig 4).

In terms of error bound estimation, we found bins B_Q and B_1 could offer good error bound estimates, but the intermediate bins could not (Figs. 3e & 3f). We found all uncertainty methods performed broadly the same: effective at predicting error bounds for B_1 and B_Q, but poor at predicting error bounds for B_2-B_{Q-1}. The one exception was PHD-Net using S-MHA, which could not accurately predict error bounds for B_1 due to the high variance in pixel activations of highly certain predictions.

We demonstrated our Quantile Binning and the three uncertainty metrics are generalizable across imaging modalities by reporting effective results on the Cephalometric dataset in Fig. 5. Here, we also showed the flexibility of Quantile Binning by varying the number of bins (Q), illustrating the trade-off between true error quantile accuracy and binning resolution as Q increases.

Finally, in Sec. VI-G we explored the effect of aleatoric uncertainty on our chosen epistemic uncertainty measures, using Quantile Binning to uncover weaknesses of E-MHA when dealing with landmarks with high directional ambiguity under conventional isotropic heatmap regression.

B. Recommendations

We offer the following recommendations:

- When resources are available, E-MHA should be used as the coordinate extraction and uncertainty estimation method since it offers the best baseline localization performance with a sufficient ability to filter out the gross mispredictions.
- If the definition of the landmark is known to be directionally ambiguous, use E-CPV over E-MHA for uncertainty estimation. If this is unknown, uniformity in the E-MHA Quantile Bins can be an indication of directional ambiguity in the landmark.
- When resources are constrained, S-MHA is surprisingly effective at capturing the true error quantiles for bins B_1 and B_Q, but note that when using a patch-based voting heatmap that is not strictly bounded, the error bound estimation for B_1 is not robust.

- The number of Quantile Bins used (Q) is a trade-off, with a larger Q offering a finer binning resolution at the cost of less accurate bins. Q is constrained by the size of the hold-out validation set and can perform poorly when $Q > 10$ and the validation set is smaller than 60 samples.

C. Conclusion

Beyond the above recommendations, we hope the framework described in this paper can be used to assess refined or novel uncertainty metrics for landmark localization, and act as a baseline for future work. Furthermore, we have shown that both the voting derived heatmap of PHD-Net, and the regressed Gaussian heatmap of U-Net can be exploited for uncertainty estimation. In this paper, we only explored the activation of the peak pixel, but it is likely that more informative measures can be extracted from the broader structure of the heatmap, promising greater potential for uncertainty estimation in landmark localization waiting to be uncovered.

REFERENCES

[1] R. Beichel, H. Bischof, F. Leberl, and M. Sonka, "Robust active appearance models and their application to medical image analysis," *IEEE Trans. Med. Imag.*, vol. 24, no. 9, pp. 1151–1169, 2005.

[2] H. J. Johnson and G. E. Christensen, "Consistent landmark and intensity-based image registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 5, pp. 450–461, 2002.

[3] K. Murphy, B. van Ginneken, S. Klein, M. Staring, B. J. de Hoop, M. A. Viergever, and J. P. Pluim, "Semi-automatic construction of reference standards for evaluation of image registration," *Med. Image Anal.*, vol. 15, no. 1, pp. 71–84, 2011.

[4] A. Holzinger, "Interactive machine learning for health informatics: When do we need the human-in-the-loop?" *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[5] D. Drevický and O. Kodym, "Evaluating deep learning uncertainty measures in cephalometric landmark localization." in *Proc. BIOIMAGING*, 2020, pp. 213–220.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Springer, 2015, pp. 234–241.

[7] L. Schobs, S. Zhou, M. Cogliano, A. J. Swift, and H. Lu, "Confidence-quantifying landmark localisation for cardiac MRI," in *Proc. ISBI*. IEEE, 2021, pp. 985–988.

[8] C.-W. Wang, C.-T. Huang, J.-H. Lee, C.-H. Li, S.-W. Chang, M.-J. Siao, T.-M. Lai, B. Ibragimov, T. Vrtovec, O. Ronneberger *et al.*, "A benchmark for comparison of dental radiography analysis algorithms," *Med. Image Anal.*, vol. 31, pp. 63–76, 2016.

[9] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Integrating spatial configuration into heatmap regression based CNNs for landmark localization," *Med. Image Anal.*, vol. 54, pp. 207–219, 2019.

[10] Z. Zhong, J. Li, Z. Zhang, Z. Jiao, and X. Gao, "An attention-guided deep regression model for landmark detection in cephalograms," in *Proc. MICCAI*. Springer, 2019, pp. 540–548.

[11] N. Torosdagli, D. K. Liberton, P. Verma, M. Sincan, J. S. Lee, and U. Bagci, "Deep geodesic learning for segmentation and anatomical landmarking," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 919–931, 2018.

[12] F. Thaler, C. Payer, M. Urschler, D. Štern *et al.*, "Modeling annotation uncertainty with gaussian heatmaps in landmark localization," *J. Mach. Learn. Biomed. Imag.*, vol. 1, pp. 1–10, 2021, Art. no. 014.

[13] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected U-Nets for efficient landmark localization," in *Proc. ECCV*, 2018, pp. 339–354.

[14] J. Yang, Q. Liu, and K. Zhang, "Stacked hourglass network for robust facial landmark localisation," in *Proc. CVPR*, 2017, pp. 79–87.

[15] J. Zhang, M. Liu, and D. Shen, "Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4753–4764, 2017.

[16] O. Emad, I. A. Yassine, and A. S. Fahmy, "Automatic localization of the left ventricle in cardiac MRI images using deep learning," in *Proc. EMBC*. IEEE, 2015, pp. 683–686.

[17] Y. Li, A. Alansary, J. J. Cerrolaza, B. Khanal, M. Sinclair, J. Matthew, C. Gupta, C. Knight, B. Kainz, and D. Rueckert, "Fast multiple landmark localisation using a patch-based iterative network," in *Proc. MICCAI*. Springer, 2018, pp. 563–571.

[18] J. Noothout, B. de Vos, J. Wolterink, T. Leiner, and I. Isgum, "CNN-based landmark detection in cardiac CTA scans," in *Proc. MIDL*, 2018, pp. 1–11.

[19] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, accepted in 2021, to appear.

[20] J. M. Noothout, B. D. De Vos, J. M. Wolterink, E. M. Postma, P. A. Smeets, R. A. Takx, T. Leiner, M. A. Viergever, and I. Išgum, "Deep learning-based regression and classification for automatic landmark localization in medical images," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4011–4022, 2020.

[21] A. Jungo, F. Balsiger, and M. Reyes, "Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation," *Frontiers in Neuroscience*, 2020, Art. no. 00282.

[22] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 3868–3878, 2020.

[23] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, "Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation," *Comput. Stat. Data Anal.*, vol. 142, 2020, Art. no. 106816.

[24] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Med. Image Anal.*, vol. 59, 2020, Art. no. 101557.

[25] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," in *Proc. NeurIPS*, vol. 33, 2020, pp. 4697–4708.

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

[27] R. Mehta, A. Filos, U. Baid, C. Sako, R. McKinley, M. Rebsamen, K. Dätwyler, R. Meier, P. Radojewski, G. K. Murugesan *et al.*, "QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results," *J. Mach. Learn. Biomed. Imag.*, vol. 1, 2022, Art. no. 026.

[28] D. Karimi, Q. Zeng, P. Mathur, A. Avinash, S. Mahdavi, I. Spadinger, P. Abolmaesumi, and S. E. Salcudean, "Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images," *Med. Image Anal.*, vol. 57, pp. 186–196, 2019.

[29] J.-H. Lee, H.-J. Yu, M.-j. Kim, J.-W. Kim, and J. Choi, "Automated cephalometric landmark detection with confidence regions using bayesian convolutional neural networks," *BMC Oral Health*, vol. 20, no. 1, pp. 1–10, 2020.

[30] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*. PMLR, 2017, pp. 1321–1330.

[31] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proc. AAAI*, vol. 2015, 2015, pp. 2901–2907.

[32] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. SIGKDD*, 2002, pp. 694–699.

[33] J. Hurdman, R. Condliffe, C. Elliot, C. Davies, C. Hill, J. Wild, D. Capener, P. Sephton, N. Hamilton, I. Armstrong *et al.*, "ASPIRE registry: assessing the spectrum of pulmonary hypertension identified at a REferral centre," *Eur. Resp. J.*, vol. 39, no. 4, pp. 945–955, 2012.

[34] C. Lindner, C.-W. Wang, C.-T. Huang, C.-H. Li, S.-W. Chang, and T. F. Cootes, "Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms," *Scientific Reports*, vol. 6, no. 1, pp. 1–10, 2016.