Medical Image Analysis 110 (2026) 103953

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Reliable uncertainty quantification for 2D/3D anatomical landmark localization using multi-output conformal prediction

Jef Jonkers [a,*], Frank Coopman [b], Luc Duchateau [b], Glenn Van Wallendael [c], Sofie Van Hoecke [c]

[a] Department of Electronics and Information Systems, IDLab, Ghent University, Belgium
[b] Department of Morphology, Imaging, Orthopedics, Rehabilitation and Nutrition, Biometrics Research Group, Ghent University, Belgium
[c] Department of Electronics and Information Systems, IDLab, Ghent University - IMEC, Belgium

ARTICLE INFO

2020 MSC:
62G30
62G99
62H99
68T45
68T07
68T99

Keywords:
Landmark localization
Uncertainty quantification
Conformal prediction
Multi-output conformal prediction
Prediction region

ABSTRACT

Automatic anatomical landmark localization in medical imaging requires not just accurate predictions but reliable uncertainty quantification for effective clinical decision support. Current uncertainty quantification approaches often fall short, particularly when combined with normality assumptions, systematically underestimating total predictive uncertainty. This paper introduces conformal prediction as a framework for reliable uncertainty quantification in anatomical landmark localization, addressing a critical gap in automatic landmark localization. We present two novel approaches guaranteeing finite-sample validity for multi-output prediction: multi-output regression-as-classification conformal prediction (M-R2CCP) and its variant multi-output regression to classification conformal prediction set to region (M-R2C2R). Unlike conventional methods that produce axis-aligned hyperrectangular or ellipsoidal regions, our approaches generate flexible, non-convex prediction regions that better capture the underlying uncertainty structure of landmark predictions. Through extensive empirical evaluation across multiple 2D and 3D datasets, we demonstrate that our methods consistently outperform existing multi-output conformal prediction approaches in both validity and efficiency. This work represents a significant advancement in reliable uncertainty estimation for anatomical landmark localization, providing clinicians with trustworthy confidence measures for their diagnoses. While developed for medical imaging, these methods show promise for broader applications in multi-output regression problems.

1. Introduction

Automatic anatomical landmark localization is crucial in advancing diagnostic capabilities based on medical images. Despite the increasing volume of publications within the biomedical domain addressing this area (Payer et al., 2019; Thaler et al., 2021; Ryu et al., 2022; Pei et al., 2023; de Queiroz Tavares Borges Mesquita et al., 2023), the current emphasis on maximizing predictive accuracy neglects uncertainty quantification. Uncertainty quantification is vital for unlocking the full potential of clinical artificial intelligence by providing a more nuanced understanding of model predictions, enhancing diagnostic accuracy and decision support, and subsequently, improving human health (Banerji et al., 2023).

The central emphasis on predictive accuracy prompts the need to focus on the reliability of machine learning models, especially in clinical settings. Deployed models must not only deliver accurate predictions but also provide reliable uncertainty measures, increasing trust in their outputs and enabling decision support frameworks (Begoli et al., 2019; Kompa et al., 2021).

Existing approaches to uncertainty quantification in anatomical landmark localization often fall short of providing reliable estimates (Kumar et al., 2019, 2020; Drevický and Kodym, 2020; Lee et al., 2020; Ma et al., 2020; Payer et al., 2020; Thaler et al., 2021; Kwon et al., 2021; Schobs et al., 2023b,a), i.e., failing to accurately quantify the degree of confidence in predictions or measurements, which is a fundamental characteristic of any uncertainty quantification method. Approaches often also fail to address the total predictive uncertainty (Kumar et al., 2019, 2020; Ma et al., 2020; Lee et al., 2020; Kwon et al., 2021; McCouat and Voiculescu, 2022; Schobs et al., 2023b), i.e., encompassing all sources of variability and error. Consequently, there is a need to explore alternative methodologies that offer robust uncertainty estimates in medical image analysis.

We propose introducing conformal prediction to anatomical landmark localization tasks, a methodology capable of generating prediction

* Corresponding author.
E-mail address: jef.jonkers@ugent.be (J. Jonkers).

https://doi.org/10.1016/j.media.2026.103953
Received 26 February 2025; Received in revised form 8 December 2025; Accepted 16 January 2026
Available online 27 January 2026

regions with finite-sample coverage guarantees. Besides evaluating several existing multi-output conformal prediction approaches for the landmark localization task, this work also introduces two novel methods, i.e., multi-output regression to classification conformal prediction set to region (M-R2C2R) and multi-output regression-as-classification conformal prediction (M-R2CCP), an extension of the R2CCP (Guha et al.) approach to multi-output regression. These approaches successfully overcome key challenges in landmark localization and in multi-output regression problems in general, including handling multi-dimensional outputs effectively and managing varying response domains across different examples.

The remainder of the paper is structured as follows. Section 2 provides a comprehensive review of existing uncertainty quantification approaches for landmark localization, highlighting their limitations in accounting for different sources of uncertainty and the lack of reliability guarantees in current methods. Section 3 describes different uncertainty quantification approaches for landmark localization and introduces our novel multi-output conformal prediction approaches. Section 4 discusses the experimental setup and datasets, followed by Section 5, which discusses and presents the results. Finally, Section 7 concludes the work and outlines directions for future research.

2. Related work

Below, we briefly discuss different landmark localization approaches and the related work to uncertainty quantification for this problem.

2.1. Landmark localization

Landmark localization in medical imaging has evolved significantly with the emergence of deep learning approaches. Mainly, approaches have been based on coordinate and heatmap regression (Jonkers et al., 2025). While coordinate regression directly predicts landmark coordinates, heatmap regression has gained prominence since its introduction by Tompson et al. (2014), who demonstrated superior performance by predicting likelihood maps for each landmark.

Recent advances in heatmap regression have led to several methodological variations. Direct heatmap regression methods can be categorized into static and adaptive approaches (Jonkers et al., 2025). Static approaches use fixed hyperparameters for heatmap distribution during training, while adaptive methods dynamically adjust these parameters. Some works have explored treating heatmap parameters as learnable model parameters (Payer et al., 2019, 2020; Thaler et al., 2021), while others have implemented scheduling methods that modify parameters based on evaluation metrics (Teixeira et al., 2019).

The spatial configuration network (Payer et al., 2019) has emerged as a particularly effective architecture for medical imaging applications where anatomical structures maintain consistent spatial relationships, such as cephalograms or pelvis radiographs. This approach integrates spatial configuration information into the heatmap regression framework, improving localization accuracy for positionally consistent medical images.

Another approach that allows for a better representation of predictive uncertainty is the contour-hugging method (McCouat and Voiculescu, 2022), which transforms landmark localization into a classification task using one-hot encoded heatmaps.

Fully convolutional neural networks with differentiable decoding operations, such as the soft-argmax operation (Luvizon et al., 2018; Dong et al., 2018; Bulat et al., 2021), generate heatmaps as an intermediate layer and optimize using coordinate-based loss. Some approaches further account for annotation ambiguity through uncertainty-aware loss functions (Kumar et al., 2019, 2020; Zhou et al., 2023), enhancing the robustness of landmark predictions.

Two-stage approaches have proven particularly effective in the medical domain (Jiang et al., 2022; Song et al., 2020; Zhong et al., 2019). These methods first localize landmarks on low-resolution images before refining predictions on high-resolution patches centered around initial estimates. These approaches have been successfully applied across various medical imaging tasks, including cephalometric landmark localization (Song et al., 2020).

2.2. Uncertainty quantification

In clinical machine learning, a model's prediction is insufficient on its own; a reliable measure of its confidence must accompany it. For a clinician, the total predictive uncertainty is the most critical quantity, as it directly answers the key question: *how much might this prediction deviate from the true outcome?* This total uncertainty arises from two distinct yet interrelated sources: aleatoric (data-related) uncertainty and epistemic (model-related) uncertainty (Hüllermeier and Waegeman, 2021). Aleatoric uncertainty originates from inherent ambiguity in the data itself, such as ambiguous landmark definitions or high inter-rater variability (Kamoen et al., 2001). High aleatoric uncertainty indicates that the case is intrinsically complex and that even an expert should proceed with caution. In contrast, epistemic uncertainty stems from the model's lack of knowledge, for instance, due to limited training data or out-of-distribution inputs (Hüllermeier and Waegeman, 2021). Elevated epistemic uncertainty is a critical warning, signaling that the model's prediction may be unreliable and should be manually reviewed or discarded.

In this work, we focus on quantifying the total predictive uncertainty, encompassing both aleatoric and epistemic components. This choice is motivated by two considerations. First, in clinical applications involving diagnosis or prognosis, the clinician's primary concern is the overall deviation between the model's prediction and the actual outcome, regardless of whether the uncertainty arises from model limitations or inherent data variability. Second, although much of the literature seeks to isolate or disentangle these uncertainty sources, it is essential to recognize their interdependence (Valdenegro-Toro and Mori, 2022; Banerji et al., 2023); aleatoric variability can influence epistemic uncertainty, and the estimation of one inherently contains aspects of the other. While our focus remains on total predictive uncertainty, the underlying methodologies employed, such as density estimation and ensemble modeling, can, if desired, be adapted to quantify aleatoric and epistemic components separately. Such granularity may prove valuable in specific clinical contexts where understanding the relative contributions of uncertainty sources provides deeper insight (Banerji et al., 2023), though these approaches do not offer the finite-sample guarantees provided by the conformal prediction framework.

2.2.1. Uncertainty quantification in landmark localization

Table B.1 provides an overview of the current literature on uncertainty quantification in landmark localization, highlighting the key distinctions and identifying existing gaps. A more elaborate literature review can be found in B.1.

Current works exhibit varied approaches to uncertainty quantification in landmark localization, such as MC dropout (Lee et al., 2020; Drevický and Kodym, 2020; Jafari et al., 2022), Gaussian processes (Schobs et al., 2023a), ensembles (Drevický and Kodym, 2020; Schobs et al., 2023b), test-time augmentation (TTA) (Ma et al., 2020), maximum likelihood estimation (MLE) (Kumar et al., 2019; Avisdris et al., 2021; McCouat and Voiculescu, 2022), or inferring uncertainty from the predicted heatmaps (Drevický and Kodym, 2020; Payer et al., 2020; Thaler et al., 2021; Kwon et al., 2021; McCouat and Voiculescu, 2022; Schobs et al., 2023b). However, all of these works fail to reliably account for the total predictive uncertainty. Some only consider aleatoric uncertainty by performing maximum likelihood estimation (Kumar et al., 2019; Avisdris et al., 2021; McCouat and Voiculescu, 2022). Others only try to model the approximation uncertainty and fail to consider aleatoric and model uncertainty (Drevický and Kodym, 2020; Lee et al., 2020; Schobs et al., 2023b). Moreover, significant number of works also fail to evaluate the validity of their proposed approaches and uncertainty

measures (Kumar et al., 2019, 2020; Drevický and Kodym, 2020; Lee et al., 2020; Ma et al., 2020; Payer et al., 2020; Thaler et al., 2021; Kwon et al., 2021; Jafari et al., 2022; Schobs et al., 2023a). Most existing methods are also restricted to 2D landmark localization despite the growing use of 3D imaging as a standard in clinical practice. This shift from 2D to 3D imaging highlights the urgent need for uncertainty quantification techniques tailored to anatomical landmarks in 3D images. The existing literature should address this development more prominently, as only Ma et al. (2020), Payer et al. (2020) addresses 3D landmark localization.

This work aims to resolve the shortcomings by producing prediction regions with marginal validity guarantees for a specified confidence level, ensuring reliable uncertainty estimates for 2D and 3D landmark predictions. This is achieved by introducing the conformal prediction framework for anatomical landmark detection to account for all sources of uncertainty and to give a notion of validity. We will also empirically validate the reliability/validity of these regions and those produced by other approaches.

3. Uncertainty quantification for landmark localization

3.1. Notation and problem setup

In this work, we denote the training data as a sequence of pairs $(X_i, Y_i) = Z_i \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where X_i is the input image and Y_i is the corresponding anatomical landmark. The input space $\mathcal{X}$ is defined as $\mathcal{X} \subseteq \mathbb{R}^{H \times W \times C}$ for 2D images of height H, width W, and C channels, or $\mathcal{X} \subseteq \mathbb{R}^{D \times H \times W \times C}$ for 3D volumetric images with depth D. The response space $\mathcal{Y}$ satisfies $\mathcal{Y} \subseteq \mathbb{R}^d$, where $d \in 2, 3$ denotes the spatial dimensionality of the input. $\hat{C}_\alpha(X_{n+1}; Z_{1:n})$ represents the estimated prediction interval with a targeted confidence level $1 - \alpha$ for the image X_{n+1} based on the previous examples $Z_{1:n} := Z_1, \ldots, Z_n$. For conciseness, we will often drop the conditioning on the previous examples in the prediction interval notation, i.e., $\hat{C}_\alpha(X_{n+1})$.

3.2. Conformal prediction

Conformal prediction (Vovk et al., 1999; Saunders et al., 1999; Papadopoulos et al., 2002; Vovk et al., 2022) is a framework that quantifies predictive uncertainty by outputting a prediction set with a related confidence level $1 - \alpha$, which comes with a finite-sample coverage guarantee,

$$\mathbb{P}_{Z \sim P_Z}\left(Y_{n+1} \in \hat{C}_\alpha\left(X_{n+1}; Z_{1:n}\right)\right) \geq 1 - \alpha, \tag{1}$$

under distribution-free assumptions, only requiring exchangeability of the training observations $Z_1, \ldots, Z_n$ and test object Z_{n+1}. The prediction sets in conformal prediction are formed by comparing nonconformity scores of examples to see how unusual a predicted label is. These scores measure the disagreement between the prediction and the actual target. In regression problems, the prediction sets are often referred to as prediction intervals in the case of a univariate response and prediction regions in the multivariate case, such as landmark localization.

One of the main advantages of conformal prediction is its model-agnostic nature, i.e., it can be applied to almost every machine learning model, as well as neural networks and anatomical landmark localization algorithms. However, the initial proposed approach, transductive (full) conformal prediction (TCP) (Saunders et al., 1999), is computationally expensive for most algorithms since it requires a refitting of the model for every possible label in the response domain. For regression problems, one refits the model for several grid points spanning the domain to make it somewhat computationally feasible, but still, for computationally expensive fitting procedures such as neural networks, this approach is practically infeasible.

To overcome this limitation, we adopt inductive (split) conformal prediction (ICP) (Papadopoulos et al., 2002), a computationally efficient variant of TCP. ICP achieves this efficiency by dividing the available data into two disjoint subsets: a proper training set, used to fit the predictive model, and a calibration set, used solely to estimate the distribution of prediction errors and determine prediction intervals. This separation avoids the need for repeated model refitting, enabling the use of conformal prediction with computationally demanding models such as neural networks. The trade-off, however, is that the calibration data cannot be used to optimize the model parameters, which can make ICP slightly less statistically efficient than TCP. The general procedure of ICP is summarized in Algorithm 1.

Algorithm 1 General procedure of ICP.

1: **Assumption:** Exchangeability of calibration set $Z_{n-m+1:n}$ and test example Z_{n+1}.
2: **Input:** Training data sequence $Z_{1:n}$, test object X_{n+1}, landmark localization model $\mathcal{L}$, and size of the calibration set m.
3: The data sequence $Z_{1:n}$ is split into a proper training data set $Z_{1:n-m}$ and a calibration set $Z_{n-m+1:n}$.
4: The proper training dataset $Z_{1:n-m}$ is used to train a landmark localization model $\mathcal{L}$.
5: For each example i in the calibration set $Z_{n-m+1:n}$, calculate the nonconformity score $S_i = s(X_i, Y_i)$, which relies on the landmark localization model.
6: The nonconformity scores from the calibration set are sorted in descending order: $S^*_1, \ldots, S^*_m$.
7: For each new test object X_{n+1} and confidence level $1 - \alpha$, we return the prediction region

$$\hat{C}_\alpha(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(X_{n+1}, y) \leq S^*_{\lfloor \alpha(m+1) \rfloor} \right\} \tag{2}$$

Validity and efficiency are the two main criteria to evaluate a confidence predictor (Fontana et al., 2022); they are also sometimes referred to as reliability and sharpness. In conformal prediction, confidence predictions are marginally valid under the assumption of exchangeability and can be exactly valid when using a smoothed conformal predictor (Vovk et al., 2022; Fontana et al., 2022). Because validity is generally guaranteed under these assumptions, the main objective becomes to optimize efficiency, i.e., to minimize the uncertainty associated with predictions. However, efficiency should never be optimized at the expense of validity. Fontana et al. (2022) phrase it quite clearly that *"Validity is the priority: without it, the meaning of predictive regions is lost, and it becomes easy to achieve the best possible performance"*. This is the primary problem with the current state-of-the-art uncertainty quantification for landmark localization. Few of these works that discuss and propose methods for uncertainty quantification, evaluate the reliability of their methods (see Table B.1), while that should be the starting point. For regression problems, efficiency in conformal prediction is often assessed with the width of the resulting prediction intervals, which we want to minimize to increase efficiency. However, in our landmark localization problem, the target has two or three dimensions. Thus, we also require multi-dimensional prediction regions, where the area or volume of the prediction region evaluates efficiency.

3.2.1. Beyond marginal coverage

Conformal prediction traditionally provides calibrated prediction intervals with marginal coverage. However, we often desire conditional coverage, which guarantees coverage across different data strata. However, unfortunately, true conditional coverage,

$$\mathbb{P}_{Z \sim P_Z}\left(Y_{n+1} \in \hat{C}_\alpha\left(X_{n+1}; Z_{1:n}\right) | X_{n+1}\right) \geq 1 - \alpha, \tag{3}$$

is impossible without making modeling assumptions or returning trivial intervals (Vovk, 2012; Foygel Barber et al., 2021). Several approaches have emerged to address this limitation and create more adaptive prediction intervals, guaranteeing something between finite-sample marginal and true conditional coverage or giving asymptotic conditional guarantees.

Mondrian conformal prediction (Vovk et al., 2003) offers an intuitive solution by partitioning the example space using a measurable function that assigns each data point to a specific category. By applying the general conformal prediction framework to each category, one can obtain coverage guarantees conditional on these categories. However, this approach reduces the calibration set size, potentially increasing variance around the target coverage level.

3.2.2. Adaptive prediction regions under marginal coverage

A more sophisticated approach leverages nonconformity scores that incorporate conditional uncertainty estimates. Instead of using classical absolute residual errors, these methods use distributional predictions or statistics like quantiles or moments to create more adaptive prediction intervals. One can leverage the conditional estimate to obtain more conditionally relevant prediction intervals while still controlling the marginal coverage through the conformal prediction framework. This would make marginally valid prediction intervals more adaptive to the local variability of prediction error. A large region should indicate a high point prediction error, and a small region should indicate a low prediction error.

Several works have been proposed following this idea. Normalized nonconformity scores, proposed by Papadopoulos and Haralambous (2011), combine absolute error with a reciprocal uncertainty estimate. The uncertainty can be derived from separate models, MC dropout variance, or MLE. In regression problems, generally, these normalized scores consist of the product of the absolute error $|Y_i - \hat{f}(X_i)|$ and the reciprocal of a 1D uncertainty estimate $\hat{u}(X_i)$, $S_i = \frac{|Y_i - \hat{f}(X_i)|}{\hat{u}(X_i)}$. Romano et al. (2019) introduced conformalized quantile regression (CQR), which uses a novel nonconformity score using conditional upper and lower quantile estimates. Lei et al. (2011, 2013) proposed to leverage the conditional density estimation $\hat{f}_{X_i}$, e.g., kernel density estimation, to construct the nonconformity score $S_i = -\hat{f}_{X_i}(Y_i)$ to create more adaptive prediction intervals. Similarly, Chernozhukov et al. (2021) proposed distributional conformal prediction, leveraging the probability integral transform in the nonconformity score. Sesia and Romano (2021) also leveraged the conditional distribution by using a histogram binning approach and nested prediction sets (Gupta et al., 2022), allowing to get more efficient intervals for heavily skewed distributions.

A recent innovative approach is the regression to classification conformal prediction (R2CCP) method proposed by Guha et al.. This technique transforms the regression problem into a classification task by binning the response domain and estimating a discrete probability distribution. The discrete distribution is then interpolated to create a continuous conditional distribution score, which is used to construct prediction intervals. The idea behind this approach is that it results in more stable training and is better at learning conditional expectations (Stewart et al., 2023; Guha et al.). Specifically, they propose to transform the response domain into K bins covering the entire response domain $\mathcal{Y}$, resulting in a described label space $\hat{\mathcal{Y}} = \{\hat{Y}_1, \dots, \hat{Y}_K\}$, where $\hat{Y}_k$ represent the midpoints of the kth bin. They then fit a classifier with K classes for estimating a discrete probability distribution $\hat{p}(\cdot|X_i)$. This distribution is transformed to a continuous conditional distribution score $\hat{f}_{X_i}(y)$ by taking the linear interpolation of the discrete probability estimates,

$$\hat{f}_{X_i}(y) = \gamma_k \hat{p}\big(\hat{Y}_k|X_i\big) + \big(1-\gamma_k\big)\hat{p}\big(\hat{Y}_{k+1}|X_i\big), \tag{4}$$

where $\hat{Y}_k \le y < \hat{Y}_{k+1}$ and $\gamma_k = \frac{\hat{Y}_{k+1} - y}{\hat{Y}_{k+1} - \hat{Y}_k}$.

This estimated continuous conditional distribution is then leveraged in the same manner as in Lei et al. (2011, 2013), using the nonconformity scores $S_i = -\hat{f}_{X_i}(Y_i)$.

3.2.3. Multi-output conformal prediction

Most existing conformal prediction approaches primarily address univariate output regression problems despite landmark localization being inherently a multi-output regression task. Multi-output conformal prediction has emerged as a growing research area that provides valid prediction regions for multiple target variables.

Since the pioneering work by Lei et al. (2013), researchers have proposed various approaches to extend single-output conformal prediction techniques. These approaches focus on critical aspects such as validity guarantees, efficiency, adaptivity, and computational feasibility. Our work categorizes multi-output conformal prediction approaches into five classes: statistically corrected, maximum nonconformity, copula-based, ellipsoidal, and density-based conformal prediction. A comprehensive review of these approaches is available in Appendix B.2.

This work discusses and benchmarks all these approaches except copula-based methods. Copula-based approaches require two calibration sets when employing split conformal prediction (Sun and Yu, 2024), which can significantly reduce prediction efficiency. Furthermore, these methods produce hyperrectangular prediction regions aligned with the axes of the output dimensions, which may be suboptimal for coordinate predictions like landmark localization (see Figs. 1 and 2). Such regions tend to overcompensate for uncertainty, reducing the prediction regions' efficiency.

In this section, we review statistically corrected (Figs. 1 and 2, top-left), maximum nonconformity (top-middle), and ellipsoidal (top-right) conformal prediction, focusing on their application to landmark localization. We also introduce two novel density-based approaches for multi-output regression in Section 3.2.4. The first is a method termed multi-output regression-as-classification conformal prediction (M-R2CCP), which extends the R2CCP framework proposed by Guha et al.. Second, we introduce multi-output regression to classification conformal prediction set to region (M-R2C2R). This approach allows us to leverage conformal prediction approaches for classification problems, such as adaptive prediction sets (APS) (Romano et al., 2020). APS is a conformal prediction approach that specifically addresses the challenge of constructing prediction sets for classification problems that adapt to the difficulty of each instance, ensuring better efficiency (smaller sets) while maintaining coverage. As shown in the bottom panels of Figs. 1 and 2, our approaches are way more flexible in the prediction regions they can represent and are, therefore, often more efficient.

Statistical correction. In statistically corrected multi-output conformal prediction, predictions are initially generated separately for each output dimension. Global coverage across dimensions is achieved by applying dimension-wise corrections; see Algorithm 2.

While previous work (Messoudi et al., 2020) used the Šidák correction, setting individual significance levels to $\alpha_t = 1 - \sqrt[d]{1-\alpha}$, which provides exact validity for independent scores but becomes conservative when the nonconformity scores are positively dependent and can result in invalid prediction regions when they are negatively dependent, we instead opt for the Bonferroni correction, $\alpha_t = \frac{\alpha}{d}$, for benchmarking purposes (Dunn, 1961). This correction comes with the marginal coverage guarantee (Eq. (1)) without any assumptions about the dependence structure of the nonconformity scores across dimensions (see Proposition 1). Note that one can use any nonconformity score; here we use normalized scores based on variance estimates of uncertainty, which yield hyperrectangular prediction regions aligned with the coordinate axes (see Figs. 1 and 2). We obtain these variance estimates using one of the approaches discussed in Section 3.3.

Proposition 1. *Suppose that $Z_{1:n+1}$ are exchangeable, and $\alpha_t = \frac{\alpha}{d}$ (Bonferroni correction). Then, the prediction region defined by Algorithm 2 satisfies the marginal coverage guarantee in Eq. (1).*

Proof. See Appendix A. □

Maximum nonconformity score. The maximum nonconformity (Dietterich and Hostetler, 2022; Diquigiovanni et al., 2022) approach does not change anything to Algorithm 1. It only uses a specific family of nonconformity scores,

$$S_i := \max\big(S_{i,1}, \dots, S_{i,d}\big), \tag{6}$$

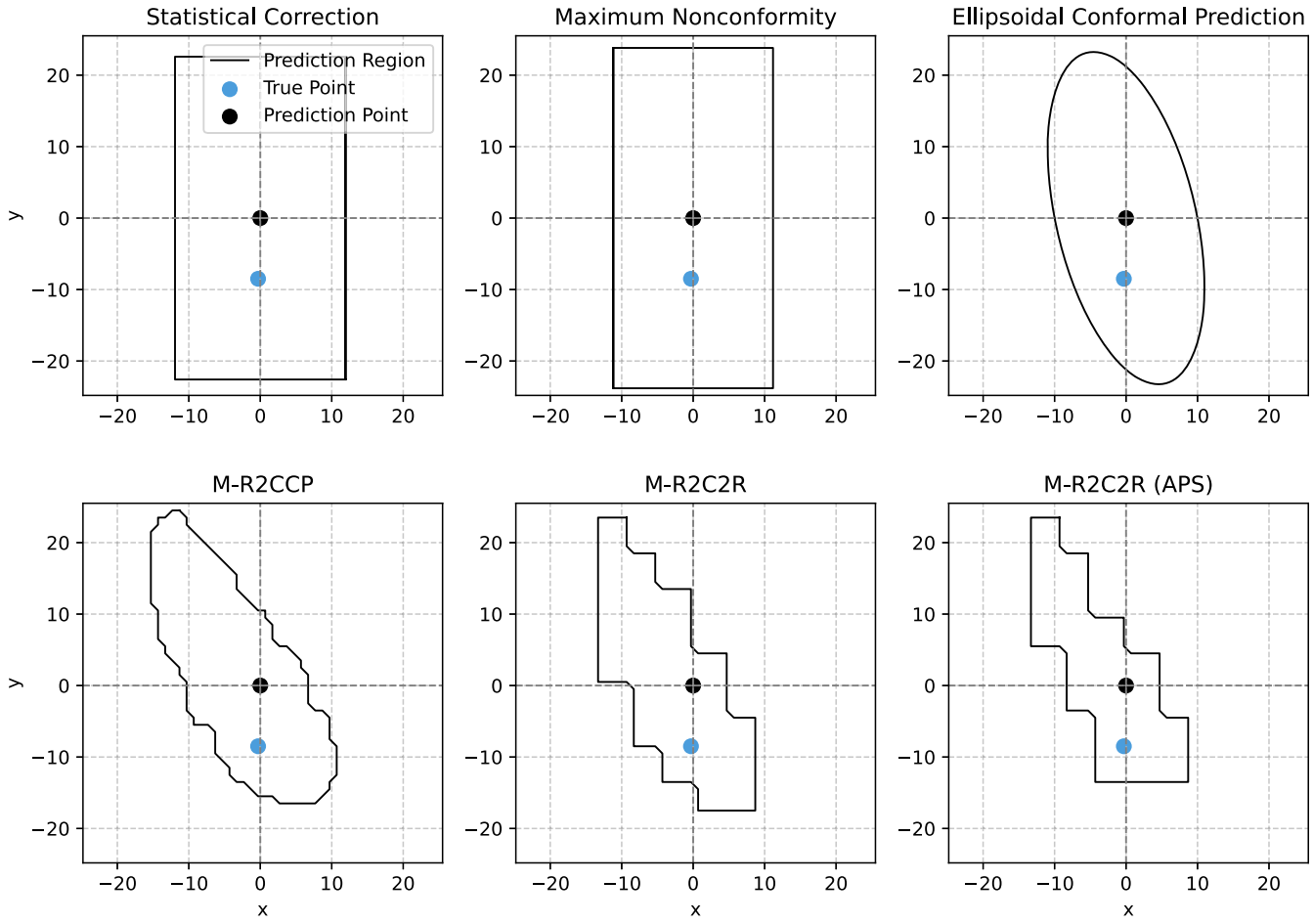

Fig. 1. Examples of prediction region produced by different multi-output conformal prediction approaches (2D).

Algorithm 2 Statistically corrected ICP.

1: **Assumption:** Exchangeability of calibration set $Z_{n-m+1:n}$ and test example Z_{n+1}.
2: **Input:** Training data sequence $Z_{1:n}$, test object X_{n+1}, landmark localization model $\mathcal{L}$, size of the calibration set m, and output dimension d.
3: The data sequence $Z_{1:n}$ is split into a proper training data set $Z_{1:n-m}$ and a calibration set $Z_{n-m+1:n}$.
4: The proper training dataset $Z_{1:n-m}$ is used to train a landmark localization model $\mathcal{L}$.
5: **for** $j \leftarrow 1$ to d **do**
6: For each example i in the calibration set $Z_{n-m+1:n}$, calculate the nonconformity score $S_{i,j} = s(X_i, Y_{i,j})$ for dimension j, which relies on the landmark localization model.
7: The nonconformity scores from the calibration set are sorted in descending order: $S^*_{1,j}, \ldots, S^*_{m,j}$
8: **end for**
9: For each new test sample X_{n+1} and confidence level $1-\alpha$, we return the prediction region

$$\hat{C}_\alpha(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(X_{n+1}, y_j) \leq S^*_{\lfloor \alpha_t (m+1) \rfloor, j} \text{ for } j = 1, \ldots, d \right\} \quad (5)$$

where α_t is the adjusted significance level based on α.

which allows for dealing with multi-output response values. The nonconformity score $S_{i,j}$ for example i and dimension j, can be any nonconformity score used for single-output regression problems, such as the absolute error, normalized absolute error, or CQR. This work will use the normalized absolute error, like for the Bonferroni correction. Similarly, as for the Bonferroni correction, this approach will result in hyperrectangular prediction regions aligned with the axes (see Figs. 1 and 2).

Ellipsoidal conformal prediction. Approaches (Johnstone and Cox, 2021; Messoudi et al., 2022; Henderson et al., 2024) using Mahalanobis-based nonconformity score enable multi-output response handling without changing Algorithm 1 by calculating scores as:

$$S_i = \sqrt{\left(Y_i - \hat{f}(X_i)\right)^T \hat{\Sigma}_i^{-1} \left(Y_i - \hat{f}(X_i)\right)}. \quad (7)$$

These methods produce ellipsoidal prediction regions that can be oriented in any direction based on the estimated covariance matrix. Uncertainty estimates can be derived from heatmaps, MC dropout, deep ensembles, or test-time augmentations (see Section 3.3). Note that this can be seen as an extension of the normalized absolute residual scores to multivariate response domain.

3.2.4. *Multi-output regression-as-classification conformal prediction*

The existing single-output R2CCP approach proposed by Guha et al. presents limitations when applied to multi-output problems, such as anatomical landmark localization. This work extends the R2CCP framework to address three key challenges:

Challenge 1: Handling multi-output response scenarios.
Challenge 2: Allow leveraging conformal prediction approaches for the classification setting (e.g., APS Romano et al., 2020) for regression problems.

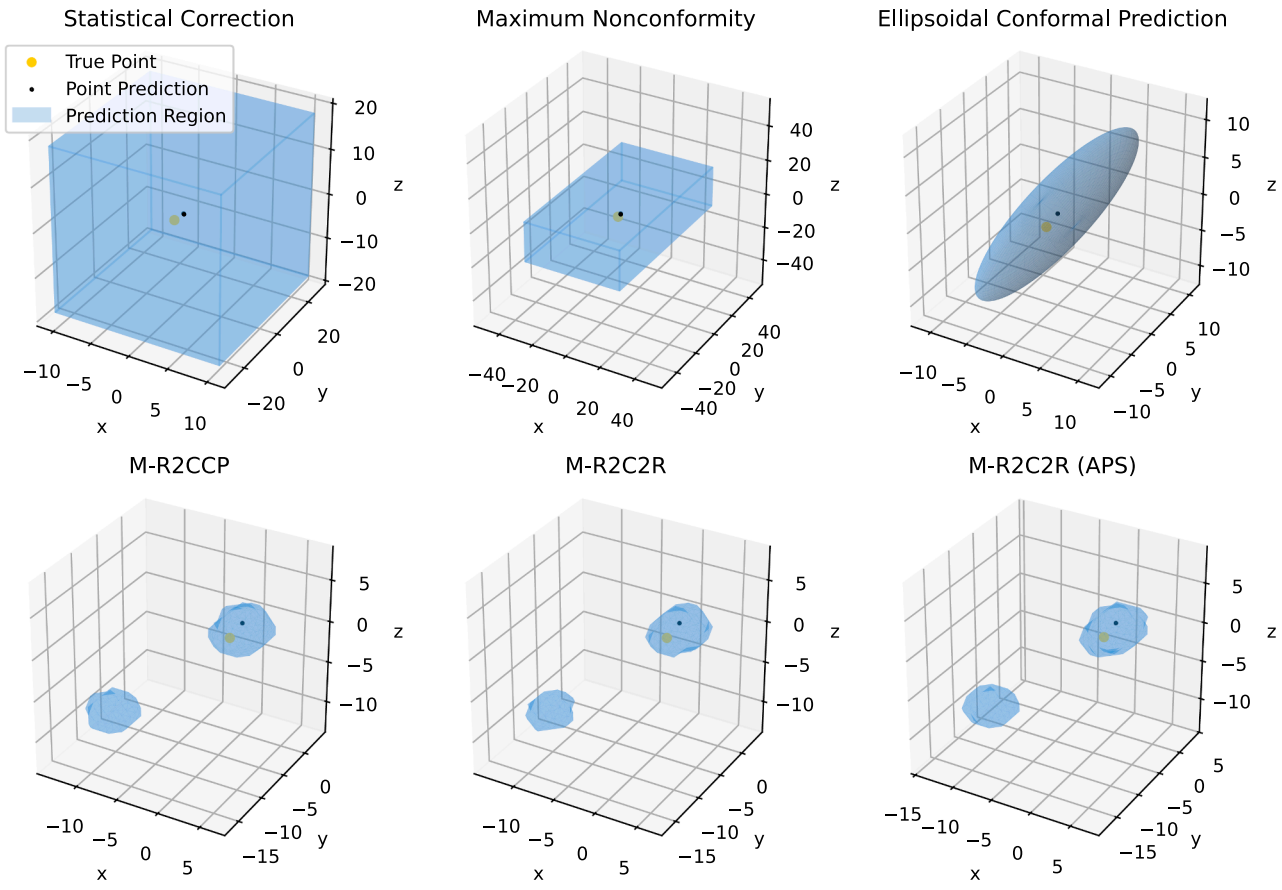

Fig. 2. Examples of prediction region produced by different multi-output conformal prediction approaches (3D).

Challenge 3: Managing varying response domains across different examples (e.g., different image resolutions).

To resolve these challenges, we introduce two key methodological advances. The first challenge leads to the development of multi-output regression to classification conformal prediction (M-R2CCP). The solutions to the second and third challenges are formally established in Propositions 2 and 3, respectively. Combining these solutions results in multi-output regression to classification conformal prediction set to region (M-R2C2R).

Challenge 1: Addressing multi-output responses. We propose binning the response space into equally sized hyperrectangular regions to handle multi-output responses. While pixel-level binning is conceptually straightforward for medical images, it quickly becomes computationally intractable. Instead, we recommend creating bins that encompass multiple pixels or voxels. Like the univariate approach, we fit a classifier with K classes for estimating a discrete probability distribution $\hat{p}(\cdot|X_i)$. We then transform this discrete distribution into a continuous conditional distribution score $\hat{f}_{X_i}(y)$ using linear interpolation:

$$\hat{f}_{X_i}(y) = \sum_{(k_1,\ldots,k_d)\in\mathcal{I}} \left[\prod_{j=1}^{d} \gamma_{k_j,j}\right] \hat{p}\left(\hat{Y}_{(k_1,\ldots,k_d)}|X_i\right), \tag{8}$$

where $\mathcal{I}$ is the set of all interpolation indices, $\gamma_{k_j,j} = \frac{Y_{k_j+1,j}-y_j}{Y_{k_j+1,j}-Y_{k_j,j}}$ for $\hat{Y}_{k_j,j} \leq y_j < \hat{Y}_{k_j+1,j}$. In Eq. (8), the summation covers all hypercube vertices involved in the n-linear interpolation and $\prod_{j=1}^{d} \gamma_{k_j,j}$ computes the interpolation weight for each vertex. The nonconformity scores are computed as $S_i = -\hat{f}_{X_i}(Y_i)$. Algorithm 3 provides a more detailed implementation of the M-R2CCP. While various interpolation methods can transform discrete bins into continuous responses, the specific approach may impact interval efficiency. Future research could systematically investigate the performance of different interpolation techniques.

One minor problem with (M-)R2CCP is that it requires an evaluation of the entire target domain to construct the prediction interval (Algorithm 3, line 9). This is often resolved by performing a grid search evaluation on the target domain.

Challenge 2: Classification conformal prediction for regression. Rather than treating the multi-output regression problem through traditional grid search methods, we propose leveraging classification conformal prediction approaches. This strategy allows us to leverage advanced classification conformal prediction techniques that generate more adaptive prediction sets.

Our multi-output regression to classification conformal prediction set to region (M-R2C2R) approach transforms the regression problem into a classification task. By generating prediction sets through classification conformal prediction and converting these discrete sets to continuous prediction regions, we can utilize sophisticated methods like adaptive prediction sets (APS) (Romano et al., 2020).

Standard conformal prediction approaches typically use $S_i = 1 - \hat{p}_{Y_i}$ as a conformity score for each class y independently when constructing prediction sets. While these methods guarantee coverage, the resulting set sizes can vary substantially: "easy" inputs (with high model confidence) may still produce unnecessarily large sets, whereas "hard" inputs (with uncertain or flat probability distributions) may yield sets that are too small or even empty, providing little practical benefit. APS addresses this issue by redefining the conformity score. Instead of scoring each class y independently, APS bases the score on the entire sorted probability distribution (from highest to lowest). Specifically, the conformity

Algorithm 3 Multi-output regression to classification conformal prediction (M-R2CCP).

1: **Assumption:** Exchangeability of calibration set $Z_{n-m+1:n}$ and test example Z_{n+1}.
2: **Input:** Training data sequence $Z_{1:n}$, test object X_{n+1}, landmark localization model $\mathcal{L}$, size of the calibration set m, and output dimension d.
3: **Hyperparameters:** number of bins $K_j > 1$ for each dimension j
4: For each dimension j of the output space $\mathcal{Y}$, discretize the output space dimension into K_j equidistant bins with midpoints $\{\hat{Y}_{1,j}, \dots, \hat{Y}_{K_j,j}\}$
5: The data sequence $Z_{1:n}$ is split into a proper training data set $Z_{1:n-m}$ and a calibration set $Z_{n-m+1:n}$.
6: Find a discrete probability distribution $\hat{p}(\cdot|X_i)$ by optimizing the landmark localization model $\mathcal{L}$ on the proper training dataset $Z_{1:n-m}$. For example, by using the one-hot heatmap approach.
7: For each example i in the calibration set $Z_{n-m+1:n}$, calculate the nonconformity score $S_i = -\hat{f}_{X_i}(Y_i)$ (see Eq. (8)).
8: The nonconformity scores from the calibration set are sorted in descending order: $S_1^*, \dots, S_m^*$.
9: For each new test sample X_{n+1} and confidence level $1-\alpha$, we return the prediction region

$$\hat{C}_\alpha(X_{n+1}) = \left\{ y \in \mathcal{Y} : -\hat{f}_{X_{n+1}}(Y_{n+1}) \leq S^*_{\lfloor \alpha(m+1) \rfloor} \right\} \tag{9}$$

score corresponds to the cumulative probability mass of classes more likely than the observed class Y_i. As a result, for "easy" inputs, prediction sets are forced to be small, while for "hard" inputs, the quantile threshold naturally expands the set to include more classes-achieving size adaptivity while maintaining rigorous coverage guarantees.

Classification conformal prediction approaches offer several advantages for regression problems. They can potentially produce more example-driven prediction sets that adapt more flexibly to the underlying data distribution; in some cases, we can also reduce the computation burden of the M-R2CCP grid search.

The M-R2C2R approach outlined in Algorithm 4 and its validity of the generated prediction regions is guaranteed by Proposition 2. The transformation of the prediction region in M-R2C2R is handled next.

Proposition 2. *Suppose a transformation $g : \mathcal{Y} \to \mathcal{P}(\mathcal{Y}')$, where $g(y)$ maps each $y \in \mathcal{Y}$ to a set in $\mathcal{Y}'$, and assuming we have a marginal coverage guarantee on $Y_{n+1} \subseteq \mathcal{Y}$, then the marginal coverage guarantee holds under the transformation of the response and prediction set,*

$$\mathbb{P}_{Z\sim P_Z}\left(g(Y_{n+1}) \in g\left(\hat{C}_\alpha(X_{n+1}; Z_{1:n})\right)\right) \geq 1-\alpha, \tag{10}$$

provided that the transformed prediction set is constructed as

$$g\left(\hat{C}_\alpha(X_{n+1}; Z_{1:n})\right) = \bigcup_{y \in \hat{C}_\alpha(X_{n+1}; Z_{1:n})} g(y). \tag{11}$$

Proof. See Appendix A. □

Challenge 3: Addressing varying response domains. We propose a domain transformation approach that standardizes input and response dimensions. We prove that we maintain marginal coverage guarantees by converting all examples to a fixed dimension (e.g., 512×512), computing nonconformity scores, and subsequently rescaling prediction regions. This method seamlessly integrates with existing anatomical landmark localization pipelines, which typically perform inference on rescaled images. The procedure still comes with the marginal coverage guarantee since it is a monotonic transformation of the response domain; see Proposition 3. Note that it can also be applied to previously described conformal prediction approaches.

Proposition 3. *Suppose the transformation of the response domain is denoted by $g : \mathcal{Y} \to \mathcal{Y}'$, and assuming we have a marginal coverage guarantee (see Eq. (1)) on $Y_{n+1} \subseteq \mathcal{Y}$, then the marginal coverage guarantee holds under the transformation of the response and prediction set,*

$$\mathbb{P}_{Z\sim P_Z}\left(g(Y_{n+1}) \in g\left(\hat{C}_\alpha(X_{n+1}; Z_{1:n})\right)\right) \geq 1-\alpha, \tag{12}$$

when g is a monotonic OR invertible transformation.

Proof. See Appendix A. □

Algorithm 4 Multi-output regression to classification conformal prediction set to region (M-R2C2R).

1: **Assumption:** Exchangeability of calibration set $Z_{n-m+1:n}$ and test example Z_{n+1}.
2: **Input:** Training data sequence $Z_{1:n}$, test object X_{n+1}, landmark localization model $\mathcal{L}$, size of the calibration set m, and output dimension d.
3: **Hyperparameters:** number of bins $K_j > 1$ for each dimension j
4: Transform the output space $\mathcal{Y}$ to $\mathcal{Y}^*$. E.g., transform all images and corresponding coordinates to a 512×512 resolution. (Challenge 3)
5: For each dimension j of the output space $\mathcal{Y}$, discretize the output space dimension into K_j equidistant bins with midpoints $\{\hat{Y}_{1,j}, \dots, \hat{Y}_{K_j,j}\}$
6: The data sequence $Z_{1:n}$ is split into a proper training data set $Z_{1:n-m}$ and a calibration set $Z_{n-m+1:n}$.
7: Find a discrete probability distribution $\hat{p}(\cdot|X_i)$ by optimizing the landmark localization model $\mathcal{L}$ on the proper training dataset $Z_{1:n-m}$. For example, by using the one-hot heatmap approach.
8: Create a discrete prediction set $C'_\alpha(X_{n+1})$ using $\hat{p}(\cdot|X_i)$, the calibration set $Z_{n-m+1:n}$, test object X_{n+1}, and conformal prediction for categorical labels.
9: Transform the discrete prediction set $C'_\alpha(X_{n+1})$ to the original response domain creating the continuous prediction region $C_\alpha(X_{n+1})$. If a bin is present in $C'_\alpha(X_{n+1})$ then all response values corresponding to that bin will be present in $C^*_\alpha(X_{n+1})$.
10: Transform prediction region $C^*_\alpha(X_{n+1})$ to the original output space $\mathcal{Y}$ resulting in prediction region $C_\alpha(X_{n+1})$. (Challenge 3)

3.3. UQ approaches without validity guarantees

Several non-conformal techniques, prevalent in the anatomical landmark localization literature (see Table B.1), serve as benchmarks for our study. We specifically incorporate one of these, the heatmap-derived weighted sampled covariance matrix, into our statistical correction, maximum nonconformity, and ellipsoidal conformal prediction approaches. While deep ensembles are also used in our conformal prediction methods, their application is restricted to the heatmap localization model, which yields more accurate and stable heatmap predictions (see Section 5.2).

3.3.1. Test-time augmentation

Test-time augmentation (TTA) randomly samples multiple augmentations of the input image during inference. It performs prediction on these samples, resulting in sample predictions that can be represented by sample statistics or empirical distributions. These transformations can be spatial, such as affine or elastic transformations, or intensity transformations, such as adding Gaussian noise or applying histogram shifts. TTA is also often used to improve accuracy and robustness.

Wang et al. (2019) provided a mathematical framework for TTA in image segmentation, comparing it with Monte Carlo dropout. While they argue that TTA can quantify image-related aleatoric uncertainty, the approach likely does not capture all sources of aleatoric uncertainty, such as label noise arising from inter- and intra-rater variability. The authors additionally proposed a combined approach integrating TTA and Monte Carlo dropout to quantify the more exhaustive notions of uncertainty, i.e., predictive uncertainty.

3.3.2. Monte Carlo dropout

Introduced by Gal and Ghahramani (2016), Monte Carlo (MC) dropout is a method for modeling epistemic uncertainty in deep neural networks by leveraging dropout layers. Typically, dropout is used only during training as a regularization technique. In MC dropout, multiple forward passes are performed through the network with dropout enabled during inference, allowing uncertainty estimation through the variance of these passes. Similar to TTA, it can also be used to improve accuracy and robustness.

Gal and Ghahramani (2016) demonstrated that dropout approximately integrates over model weights and mathematically equates a neural network with dropout to a deep Gaussian process. In our work, we extend this approach by estimating the covariance of response output for our multivariate problem, which can be derived directly from the MC samples.

3.3.3. Deep ensemble

Lakshminarayanan et al. (2017) proposed deep ensembles as a method for predictive uncertainty estimation. This approach involves creating an ensemble of neural networks with identical architecture but different weight initializations. Uncertainty is estimated by generating multiple prediction samples and analyzing their distribution. While primarily addressing epistemic uncertainty in regression contexts, Fort et al. (2020) argued that deep ensembles generally outperform MC dropout due to more decorrelated inference models.

3.3.4. Temperature scaling

Temperature scaling, introduced by Guo et al. (2017), is a calibration technique applicable to multi-class problems, specifically useful for one-hot heatmap approaches (McCouat and Voiculescu, 2022). The method introduces a single parameter τ (temperature) to calibrate heatmap probabilities by dividing each channel pixel before softmax activation. Optimal τ is determined using a calibration set to minimize the negative log-likelihood of the one-hot encoded heatmap.

3.3.5. Heatmap derived uncertainty

Research by Payer et al. (2020), Thaler et al. (2021) has explored uncertainty derivation by fitting multivariate Gaussian distributions to predicted heatmaps using least-squares curve fitting. In our work, we derive anatomical landmarks and their corresponding covariance matrix by calculating the weighted sample mean and covariance. This is also the approach we will use for our CP ellipsoidal approach which requires a covariance estimate.

Naive R2C2R. In this work, we also introduce another approach, the naive regression-to-classification-to-regression (Naive R2C2R) approach. We primarily introduce this approach to show why we need conformalization, as we can not (always) trust the probabilities represented by the heatmaps. Naive R2C2R uses the heatmap as a probability distribution. It works similarly as the APS approach (Romano et al., 2020); however, it does so without an adjustment to control coverage and thus consequently assumes an "oracle" distribution, i.e., a perfectly calibrated distribution, and thus does not come with any coverage guarantees. In practice, it ranks pixel locations in descending order based on predicted probabilities and then progressively adds these pixels to the prediction regions until the cumulative probability of that region meets the desired confidence level. This method can also be enhanced with temperature scaling and/or pixel averaging through deep ensembles, MC dropout, or TTA.

4. Datasets

4.1. ISBI 2015 automatic cephalometric X-ray landmark detection challenge (ISBI 2015)

The ISBI 2015 automatic cephalometric X-ray landmark detection challenge selected 19 landmarks commonly used in clinical practice for evaluating the ability of automatic detection of these landmarks (Wang et al., 2015). These landmarks suffer from relatively high intra- and inter-observer errors (Kamoen et al., 2001). The dataset consists of 400 annotated radiographs. While the original challenge protocol defines two predefined test sets, several subsequent studies, such as Lindner et al. (2016), Zhong et al. (2019), and Thaler et al. (2021), have adopted an alternative 4-fold cross-validation scheme to improve robustness and mitigate the systematic annotation shift (in landmark 8 and 16) observed between Test 1 and Test 2. Following these works, we also employ a 4-fold split of the data and use only the annotations from the junior annotator to maintain consistency across folds. However, our protocol slightly deviates from these prior setups, as we use two folds for model training, one fold for calibration, and one fold for testing. This additional calibration fold is required for our conformal prediction framework.

4.2. Canine hip dysplasia (CHD) landmark dataset

A second dataset used to evaluate the proposed approach in this work is a private dataset consisting of annotated canine pelvis X-rays where each radiograph has 12 annotated landmarks (see Fig. C.4), which we refer to as the canine hip dysplasia (CHD) dataset. These landmarks are used to derive radiograph measurements, which in turn are used to assess the development of hip dysplasia, one of dogs' most prevalent diseases. The dataset consists of 472 radiographs, which we use to evaluate our approaches using a 4-fold subject-level cross-validation. Specifically, we split the radiographs into four equal folds and alternated the training, calibration, and test sets. The training set always contains two folds, and the calibration and test always contain one.

4.3. Mandibular molar landmarking dataset

Finally, we also use the mandibular molar landmarking (MML) dataset (He et al., 2024), which consists of 658 annotated CT volumes of the skull. The CT volumes are transformed to a uniform scale of $512 \times 512 \times 256$. Multiple junior and senior clinicians annotated the CT volumes. It includes 14 mandibular landmarks, specifically, 4 crowns, 8 roots of the second and third mandibular molars, and 2 cups of cuspids. Note that the number of landmarks in a single CT volume is arbitrary, i.e., some landmarks are absent in the images. Since this work only focuses on landmark localization, we filter out all CT volumes where not all landmarks are present, reducing the dataset size to 399. For evaluating uncertainty quantification approaches, we split the data into a proper training dataset of 199 images, and a calibration and test set of each 100 images. For benchmarking the landmark localization model against the state-of-the-art, we use the evaluation procedure proposed by He et al. (2024). However, we propose to correct their deterministic evaluation procedure. The authors did not correct the pixel spacing for the scaling that is possibly performed during their center crop transformation. The pixel spacing should be divided by the applied scaling to get a correct point error in metric units (e.g., mm). This adjustment gives a slightly higher point error for the evaluation dataset, as seen in Tables C.3 and C.4 presenting deterministic performance, where the only difference is the scale adjustment.

5. Experiments

5.1. Evaluation metrics

We evaluate landmark localization accuracy and UQ regions using the following metrics. Definitions here apply to all subsequent experiments.

Landmark localization metrics. Let samples be indexed by $i = 1, \dots, n$ and landmarks by $l = 1, \dots, n_l$. Denote the predicted and ground-truth

coordinates by $\hat{Y}_i^l, Y_i^l \in \mathbb{R}^d$ with $d \in \{2, 3\}$ and units in mm. The per-landmark Euclidean point error (PE) is

$$\mathrm{PE}_i^l = \left\| \hat{Y}_i^l - Y_i^l \right\|_2 = \sqrt{\sum_{j=1}^{d} \left(\hat{Y}_{i,j}^l - Y_{i,j}^l \right)^2}. \tag{13}$$

We report the dataset mean and standard deviation (SD) over all nm errors:

$$\mathrm{PE} = \frac{1}{n\, n_l} \sum_{i=1}^{n} \sum_{l=1}^{n_l} \mathrm{PE}_i^l, \qquad \mathrm{SD} = \frac{1}{n_l} \sum_{l=1}^{n_l} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\mathrm{PE}_i^l - \overline{\mathrm{PE}}^l \right)^2}. \tag{14}$$

The success detection rate (SDR) at threshold τ mm is the percentage of landmark errors below τ:

$$\mathrm{SDR}(\tau) = \frac{100}{n\, n_l} \sum_{i=1}^{n} \sum_{l=1}^{n_l} \mathbb{1}\left\{ \mathrm{PE}_i^l \le \tau \right\} \%, \quad \tau \in \{2,\ 2.5,\ 3,\ 4\} \text{ mm}. \tag{15}$$

Uncertainty quantification metrics. For an image X_i and landmark l, let $\hat{C}_\alpha^l(X_i) \subset \mathbb{R}^d$ be the prediction region at confidence $1 - \alpha$, and let $r(\cdot)$ denote region size (area in 2D, volume in 3D; mm^d). We assess:

1. *Validity (coverage)* using empirical marginal coverage:

$$\widehat{\mathrm{cov}}_\alpha = \frac{1}{n\, n_l} \sum_{i=1}^{n} \sum_{l=1}^{n_l} \mathbb{1}\left\{ Y_i^l \in \hat{C}_\alpha^l(X_i) \right\}. \tag{16}$$

   A method is calibrated when $\widehat{\mathrm{cov}}_\alpha \approx 1 - \alpha$.
2. *Efficiency* by calculating the per-example region size $A_i^l = r\left(\hat{C}_\alpha^l(X_i)\right)$. At matched coverage, smaller regions are preferable. We summarize the distribution of $\{A_i^l\}_{i,l}$ (mean ± SD and median [Q1-Q3]).
3. *Adaptivity*. To quantify whether regions enlarge for harder examples, we compute the Spearman rank correlation between region size and realized error:

$$\rho_s = \mathrm{Spearman}\left(\{A_i^l\}_{i,l},\ \{\mathrm{PE}_i^l\}_{i,l} \right). \tag{17}$$

   Higher ρ_s indicates that region sizes increase monotonically with errors, a desirable property of conditional adaptivity. Scatter plots in Figs. C.1-C.3 visualize this relationship.

Remark 1. When heatmaps are decoded separately for each stochastic sample and landmark predictions are then averaged (landmark sampling, LS), both PE_i^l and the estimated spread can increase, which may mechanically raise ρ_s.

5.2. Landmark localization

This work emphasizes model-agnostic uncertainty quantification in landmark localization. However, we employ two distinct approaches to landmark localization to assess the effectiveness of different uncertainty quantification approaches. Performance is reported using the PE and SDR metrics defined in Section 5.1.

First, we utilize the spatial configuration network (SCN) model, which incorporates learned homoscedastic aleatoric uncertainty for each landmark, as introduced by Payer et al. (2020), Thaler et al. (2021). The SCN is used as a landmark localization model across all 2D datasets presented in Section 4. We use the same loss function, hyperparameters, and optimization strategies described in Payer et al. (2020).

Second, we implement the one-hot approach, introduced in McCouat and Voiculescu (2022), which transforms landmark localization into a classification task using one-hot encoded heatmaps. For 2D landmark localization, we propose using a UNet architecture that employs a ResNet-34 encoder pre-trained on ImageNet, with a depth of 5 layers. The decoder features channels of size 256, 128, 64, 32, and 32, respectively. A softmax activation is applied at the output layer to ensure proper probability distribution. This approach is ideal for the M-R2CCP and M-R2C2R approaches proposed in this work.

In this work, we suggest adapting the one-hot technique for the 3D landmark task. The foundational principle remains unchanged: we utilize a 3D UNet that integrates an EfficientNet-B0 encoder with MedicalNet's pre-trained weights, featuring a depth of 5 layers. The decoder comprises of channels sized 256, 128, 64, 32, and 16. A softmax activation is implemented at the output layer to ensure a probability distribution summing to one. This adaptation demonstrates strong performance, surpassing other methods assessed on the MML datasets (see Table 1).

For both methodologies and across all datasets, we apply a data augmentation strategy that includes random spatial transformations (affine transformations) and intensity transformations, such as adding Gaussian noise, scaling, gamma adjustments, and nonlinear transformations of the image histogram.

We also assess how sampling-based uncertainty quantification methods, like TTA, MC dropout, and deep ensembles, enhance predictive performance. Our evaluation consists of two strategies: the first involves averaging the heatmaps at the pixel level, leading to a more robust heatmap that is subsequently decoded into predictions of anatomical landmarks. The second method aligns more with uncertainty quantification approaches, as it decodes the heatmap for each individual sample and averages the landmark predictions across samples. This approach is denoted as LS. Results are presented in Tables C.1, C.2, and C.3 in Appendix C). We observed that pixel-level sampling significantly increases performance across all datasets compared to the baselines and the LS approaches.

5.3. Empirical results of uncertainty quantification approaches

We assess the various prediction regions of the uncertainty quantification methods according to the three criteria defined in Section 5.1: validity (coverage), efficiency, and adaptivity. The evaluation is conducted on the two 2D datasets and the 3D dataset to ensure diversity in our assessments. Quantitative results for coverage, efficiency, and adaptivity (ρ_s) are summarized in Tables 2–4. MC dropout is evaluated with the SCN backbone (Payer et al., 2019) (the only setup with dropout in the heatmap head). In contrast, all other UQ variants, including TTA, deep ensembles, the heatmap-derived method, and all CP methods (M-R2CCP, M-R2C2R), use the one-hot/contour-hugging backbone (McCouat and Voiculescu, 2022). We adopt this pairing because SCN imposes a (bi)variate Gaussian and suits only the MC setting, while the one-hot/contour-hugging representation better captures uncertainty without a parametric heatmap assumption; moreover, the 2D contour-hugging network has no dropout, making MC dropout inapplicable.

Coverage of prediction regions

We evaluate the validity of this work by measuring the marginal coverage of the prediction regions of different landmarks on different datasets. The marginal coverage for the ISBI 2015 (2D), CHD (2D) and MML (3D) datasets are presented in box plots in respectively Figs. 4, 5 and 6. We immediately observe that the frequently used sample inference approaches (MC dropout, deep ensemble, and TTA), combined with the normality assumption of the predictive uncertainty, drastically underestimate the prediction uncertainty. This can be attributed to the inability of the approaches to account for all sources of uncertainty, as discussed in the related work (see Section 2.2). The heatmap-derived approach, proposed by Thaler et al. (2021), seems to account a bit more for the predictive uncertainty, but it still systematically underestimates the total predictive uncertainty. To resolve this underestimation of total predictive uncertainty, we propose the conformal prediction framework to give us a finite-sample probabilistic coverage guarantee. The coverage results, presented in Figs. 4, 5, and 6, show that the conformal prediction framework gives us a valid prediction region. We note that across all three datasets, as expected, the Bonferroni correction and max nonconformity approaches give slightly more conservative prediction regions.

When temperature scaling is applied, the Naive R2C2R approaches perform reasonably well on the ISBI 2015 and CHD datasets, meaning

Table 1
Landmark localization results on the MML (3D) dataset compared against existing approaches. This benchmark dataset is the same as the data subset, only with complete landmarks, used as a benchmark in He et al. (2024). Performance is reported using the PE, i.e., the Euclidean distance (in mm) between predicted and reference landmarks, reported as mean ± standard deviation (SD), and the SDR, defined as the percentage of landmarks with PE below predefined thresholds (2, 2.5, 3, and 4 mm). The best results are highlighted in bold.

Model name	Validation						Test					
	PE (mm)	SD (mm)	SDR (%)				PE (mm)	SD (mm)	SDR (%)			
			2 mm	2.5 mm	3 mm	4 mm			2 mm	2.5 mm	3 mm	4 mm
Pruning-ResUNet3D (He et al., 2024)	1.82	–	73.21%	82.14%	88.93%	**94.76%**	1.96	1.39	70.03%	79.97%	86.10%	92.73%
One hot ensemble	**1.60**	–	**77.81%**	**87.76%**	**89.92%**	93.37%	**1.39**	1.12	**81.67%**	**91.31%**	**93.33%**	**96.31%**

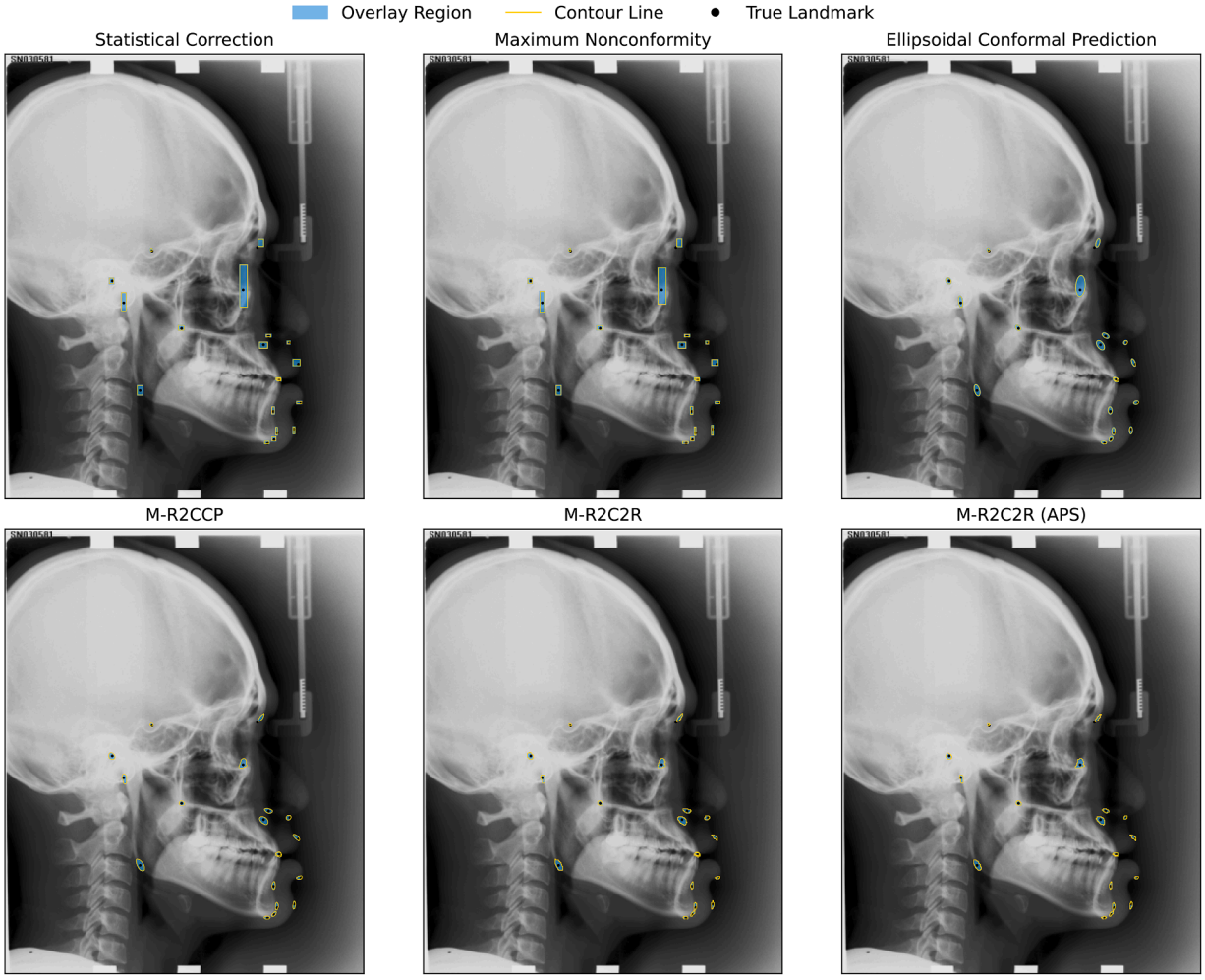

Fig. 3. Cephalogram example of conformal prediction regions.

that the temperature scaling ensures that probability distribution resembles the Oracle predictive distribution. However, temperature scaling cannot guarantee this, and this can be clearly seen for MML (3D) dataset, as the prediction regions are significantly undercover even with temperature scaling, pressing the need for the conformal prediction framework. It is interesting to note that, in general, the pixel-averaging approach using deep ensembles generally increases the validity of the prediction regions. Thus, in general, the pixel-averaging approach using deep ensembles increases the landmark localization approach and calibration of the pixel probabilities in the one-hot approach.

Efficiency and adaptivity of prediction regions

In the evaluation of prediction regions, validity serves as the primary criterion, as it underpins the interpretability and reliability of prediction regions. Additionally, it is easy to get good efficiency when you compromise on validity. Consequently, our efficiency analysis focuses exclusively on conformal prediction approaches, which provide theoretical finite-sample marginal validity guarantees. We assess efficiency through the measurement of prediction region area or volumes, utilizing both mean and median metrics to account for the variability in region size across different images, a variation that may appropriately reflect underlying predictive uncertainty. The results of our experiments are presented in Tables 2, 3, and 4. Additionally, to give an idea of the prediction regions, Fig. 3 shows the different conformal prediction regions on a cephalogram from the ISBI 2015 dataset.

We observe that the Bonferroni correction and max nonconformity approaches generate the largest prediction regions. This can be attributed to two factors: their inherently conservative nature in region

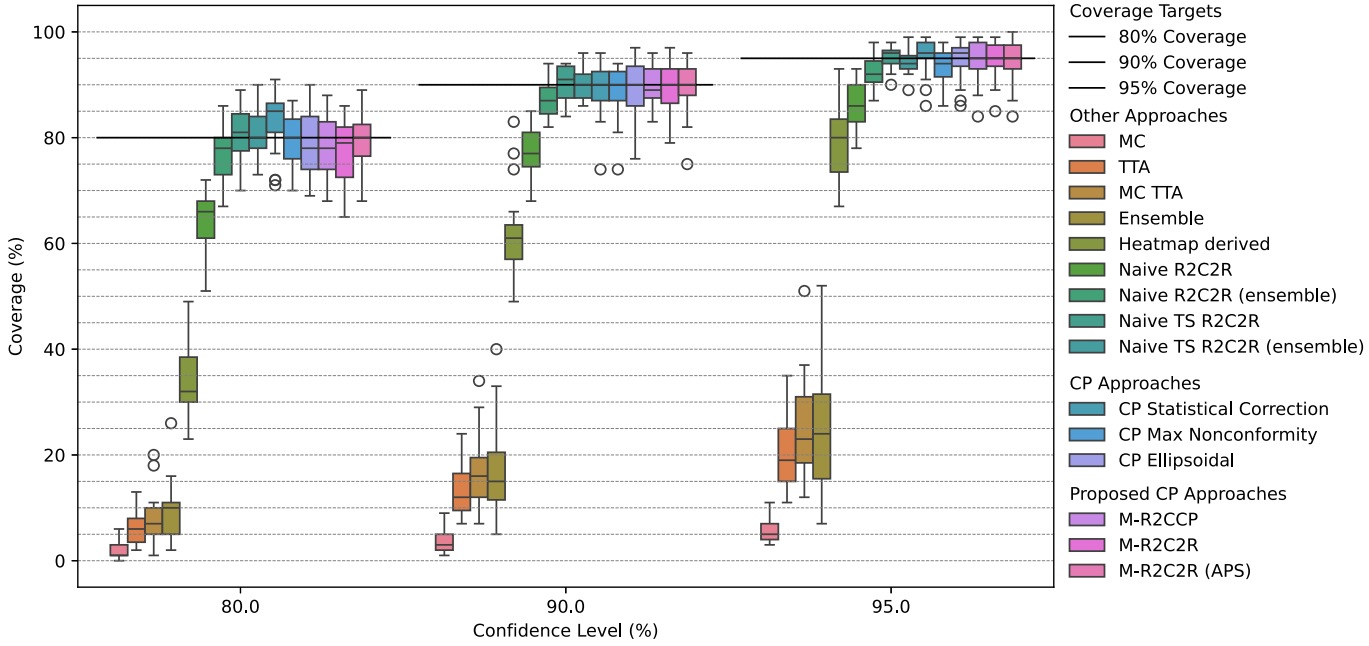

Fig. 4. Box plot showing the empirical *per-landmark* coverage of prediction regions (19 landmarks), evaluating various uncertainty quantification approaches at different confidence target levels on the ISBI 2015 (2D) test set. Each box summarizes the distribution of coverage values across landmarks, while the coverage percentages reported in Table 2 correspond to the overall empirical marginal coverage aggregated over all landmarks and subjects.

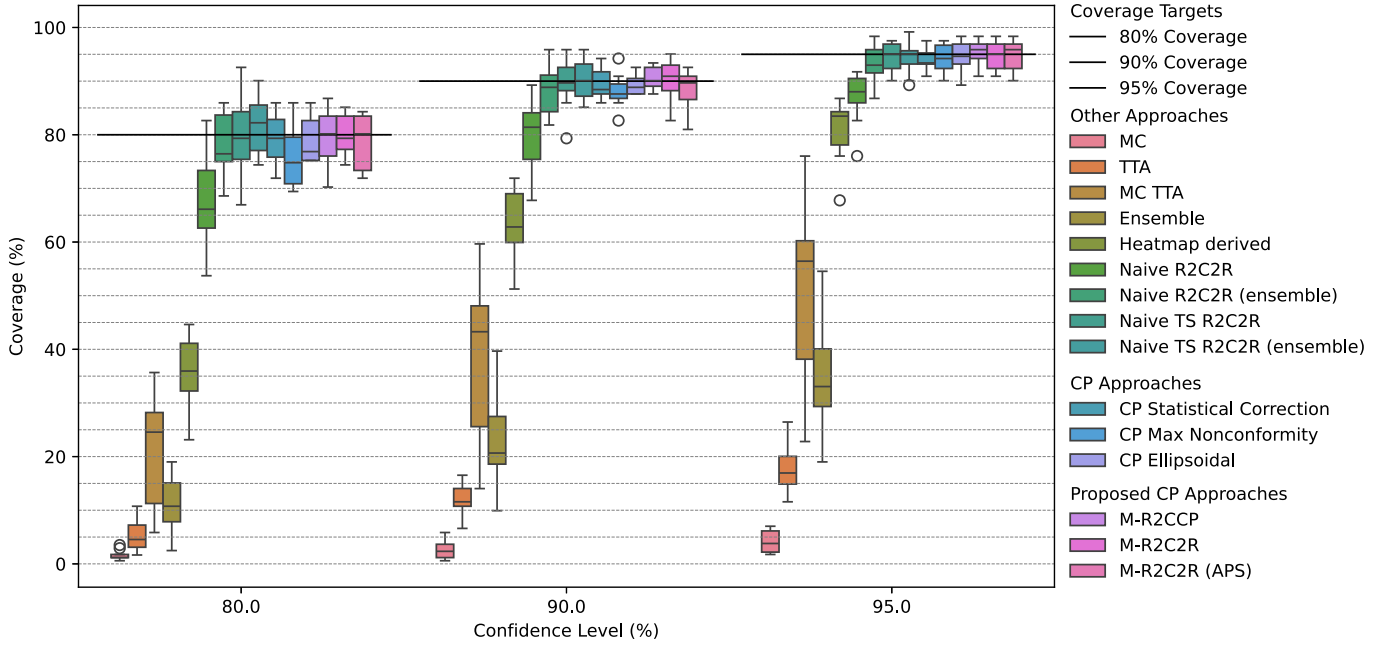

Fig. 5. Box plot showing the empirical coverage of prediction regions for different landmarks (12 landmarks), evaluating various uncertainty quantification approaches at different confidence target levels on the CHD (2D) dataset test set. Each box summarizes the distribution of coverage values across landmarks, while the coverage percentages reported in Table 2 correspond to the overall empirical marginal coverage aggregated over all landmarks and subjects.

estimation and, more significantly, their constraint to axis-aligned hyperrectangular prediction region. In contrast, the Mahalanobis nonconformity scores offer enhanced flexibility through ellipsoidal prediction regions, whose dimensions and orientations are determined by the predicted covariance matrix, resulting in improved efficiency.

Our M-R2CCP and M-R2C2R approaches demonstrate superior performance, capable of representing non-convex regions (as illustrated in Figs. 1 and 2), and consistently achieve the highest efficiency metrics. Across datasets, the two methods show similar typical behaviour: median region sizes and IQRs are comparable. The APS variant of M-R2C2R produces a small number of extremely large regions on the MML (3D) test set (mean = 4440.1 mm^3, std = 24056; median = 51.07 mm^3, IQR = 27.71-94.82). These outliers occur when the predicted heatmap is essentially flat (very high model uncertainty), causing the conformal region to span most of the image and thus inflating the reported mean and standard deviation. In practice, such extreme regions are more

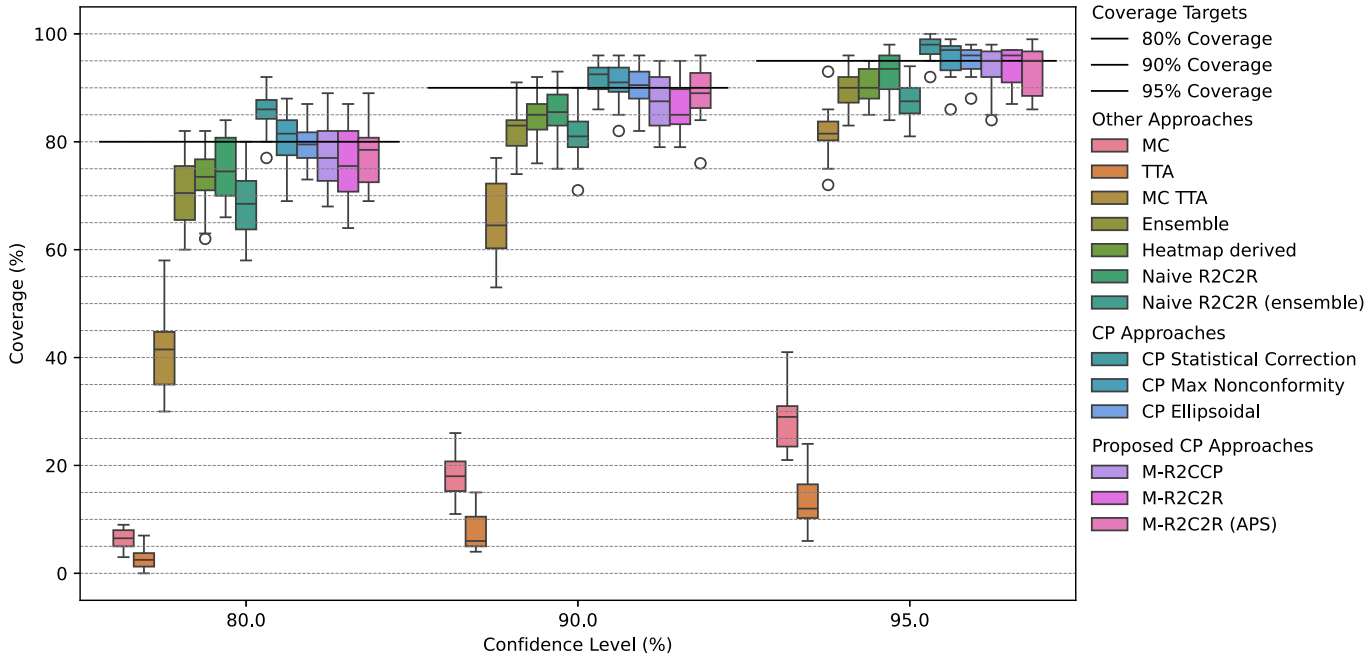

Fig. 6. Box plot showing the empirical coverage of prediction regions for different landmarks (14 landmarks), evaluating various uncertainty quantification approaches at different confidence target levels on the MML (3D) dataset test set. Each box summarizes the distribution of coverage values across landmarks, while the coverage percentages reported in Table 2 correspond to the overall empirical marginal coverage aggregated over all landmarks and subjects.

Table 2
Evaluation of 95% prediction regions of different UQ approaches on the ISBI 2015 (2D) test set. The best results are highlighted in bold.

UQ Approach	Model	Coverage (%)	Efficiency (mm^2)					Adaptivity (ρ_s)
			Mean	Std.	Median	Q1	Q3	
MC	SCN	5.47%	0.3015	1.2333	0.0862	0.0472	0.1836	0.2642
TTA	One hot	20.68%	2.7351	50.134	0.5793	0.3694	0.9650	0.3699
MC TTA	SCN	25.47%	1.4625	3.9245	0.5889	0.3778	1.0562	0.3732
Ensemble	One hot	25.21%	5.0906	65.185	0.7687	0.4734	1.6475	**0.4056**
Heatmap derived	One hot	78.63%	5.8981	5.5784	4.2713	3.2454	6.6309	0.3958
Naive R2C2R	One hot	86.26%	19.950	506.75	5.7200	3.9200	9.0150	0.4310
Naive R2C2R (ensemble)	One hot	92.47%	9.1205	6.9787	6.9300	4.8200	11.035	0.3867
Naive TS R2C2R	One hot	95.32%	676.41	3959.4	10.305	6.4650	17.940	0.3860
Naive TS R2C2R (ensemble)	One hot	94.26%	10.921	9.3519	8.0900	5.6300	13.263	0.3773
CP Statistical Correction	One hot	95.53%	55.291	245.29	14.620	7.9068	32.157	0.3703
CP Max Nonconformity		93.42%	44.982	226.92	11.465	6.4124	23.266	0.3833
CP Ellipsoidal	One hot	94.63%	15.812	18.067	9.7825	7.2443	18.094	0.3385
M-R2CCP	One hot	94.63%	**12.081**	8.0110	9.0000	7.2200	14.348	0.3360
M-R2C2R	One hot	94.84%	12.202	9.1147	**8.3350**	6.6300	15.163	0.3127
M-R2C2R (APS)	One hot	94.63%	13.213	13.278	8.5900	6.3175	16.463	0.3528

useful when surfaced as a low-confidence flag (suppressing the oversized region) rather than shown directly; to make the behaviour transparent, we report both mean (std) and median (IQR) in Table 4.

The relationship between prediction error and region size is visualized through scatter plots in Figs. C.1, C.2, and C.3. While the APS variant of M-R2C2R exhibits enhanced adaptivity, aligning with its design objectives, the practical significance of these adaptivity differences requires further investigation. It is noteworthy that conformal ellipsoid prediction regions can inherit adaptivity characteristics from sampling-based uncertainty quantification approaches (MC dropout, TTA, or deep ensembles) by leveraging their covariance matrices.

Furthermore, it is important to note that the high Spearman correlation observed for the MC dropout approach on the MML dataset (Table 4) does not completely reflect true conditional adaptivity. This inflated ρ_s arises from the *landmark sampling* decoding strategy, in which landmarks are decoded separately for each stochastic sample and then averaged. This process increases both the predicted covariance and the magnitude of the point error, thereby increasing the correlation between region size and error. Accordingly, these results should be interpreted as a methodological artifact rather than genuine adaptivity.

6. Discussion

Main findings

This work adapts conformal prediction to 2D/3D anatomical landmark localization and evaluates a family of model-agnostic UQ approaches; axis-aligned (Bonferroni, max nonconformity), ellipsoidal, and two region-building approaches (M-R2CCP, M-R2C2R). Across three datasets, CP-based methods achieve empirical marginal coverage close to target levels while offering competitive or superior efficiency relative to baselines lacking finite-sample guarantees. Ellipsoidal regions

Table 3
Evaluation of 95% prediction regions on the CHD (2D) dataset using 4-fold subject-level cross-validation. For each method and metric, we report the min–max across folds for disjoint train/calibration/test splits. The best results are highlighted in bold.

UQ Approach	Model	Coverage (%)	Efficiency (mm^2)					Adaptivity (ρ_s)
			Mean	Std.	Median	Q1	Q3	
MC	SCN	[4.2398 – 5.0195]	[0.0517 – 0.0803]	[0.1906 – 0.5304]	[0.0216 – 0.0275]	[0.0099 – 0.0141]	[0.0498 – 0.0612]	[0.3055 – 0.3622]
TTA	One hot	[17.769 – 20.417]	[0.2710 – 3.6059]	[0.1889 – 124.55]	[0.1935 – 0.2423]	[0.1349 – 0.1561]	[0.2734 – 0.3486]	[0.2088 – 0.3297]
MC TTA	SCN	[47.612 – 50.390]	[0.5793 – 0.7322]	[0.7075 – 1.1261]	[0.4487 – 0.5261]	[0.1666 – 0.2070]	[0.7345 – 0.9028]	[0.1243 – 0.1920]
Ensemble	One hot	[38.889 – 62.259]	[1.5745 – 17.044]	[16.373 – 124.74]	[0.4254 – 6.3289]	[0.1609 – 2.6780]	[0.7029 – 23.016]	[0.3219 – **0.8384**]
Heatmap derived	One hot	[83.678 – 85.139]	[1.5984 – 1.8587]	[0.7447 – 2.8803]	[1.4652 – 1.6682]	[1.0367 – 1.1093]	[2.0351 – 2.3113]	[0.3338 – 0.3855]
Naive R2C2R	One hot	[81.528 – 92.431]	[1.7363 – 3.4995]	[1.3431 – 38.172]	[1.4684 – 2.6083]	[1.0820 – 1.4298]	[2.0287 – 3.9028]	[0.3145 – 0.4380]
Naive R2C2R (ensemble)	One hot	[95.110 – 96.528]	[2.7887 – 13.113]	[1.4914 – 276.16]	[2.6083 – 2.9561]	[1.5795 – 1.7196]	[3.7145 – 4.2313]	[0.3216 – 0.3821]
Naive TS R2C2R	One hot	[93.750 – 95.868]	[4.0793 – 227.55]	[9.8052 – 1435.6]	[2.7532 – 3.3232]	[1.7727 – 2.1640]	[4.0574 – 4.9655]	[0.3200 – 0.4372]
Naive TS R2C2R (ensemble)	One hot	[94.284 – 95.694]	[2.7222 – 7.9374]	[1.5363 – 191.53]	[2.5504 – 2.7436]	[1.4684 – 1.5843]	[3.7096 – 3.9415]	[0.3184 – 0.3740]
CP Statistical Correction	One hot	[94.215 – 96.667]	[4.7676 – 15.412]	[5.2142 – 315.40]	[3.2554 – 3.9176]	[2.0653 – 2.5526]	[5.4749 – 6.2880]	[0.3122 – 0.3473]
CP Max Nonconformity	One hot	[93.595 – 96.319]	[4.5614 – 13.429]	[4.8519 – 260.16]	[3.0653 – 3.6965]	[1.9244 – 2.3359]	[5.3645 – 6.0715]	[0.3157 – 0.3618]
CP Ellipsoidal	One hot	[94.167 – 96.319]	[2.8205 – 3.3084]	[1.6544 – 4.2948]	[2.3653 – 2.6200]	[1.6675 – 1.9937]	[3.6202 – 4.2782]	[**0.3398** – 0.3709]
M-R2CCP	One hot	[93.958 – 96.042]	[2.7097 – 3.0595]	[1.3698 – 1.6368]	[2.3185 – 2.4151]	[1.6036 – 1.9321]	[3.6903 – 4.1927]	[0.3170 – 0.3562]
M-R2C2R	One hot	[93.264 – 96.597]	[**2.4629** – **2.9754**]	[1.4209 – 1.6502]	[**2.0094** – 2.3958]	[1.3911 – 1.7582]	[3.3425 – 4.0622]	[0.3147 – 0.3493]
M-R2C2R (APS)	One hot	[94.306 – 96.875]	[2.6752 – 13.278]	[1.6847 – 282.68]	[2.1350 – **2.3185**]	[1.4491 – 1.6954]	[3.5551 – 4.1057]	[0.3322 – 0.3641]

Table 4
Evaluation of 95% prediction regions of different UQ approaches on the MML (3D) dataset test set. The best results are highlighted in bold.

UQ Approach	Model	Coverage (%)	Efficiency (mm^3)					Adaptivity (ρ_s)
			Mean	Std.	Median	Q1	Q3	
MC	One hot	29.50%	232.50	1776.9	2.1958	0.8617	8.9424	**0.5794**
Ensemble	One hot	13.36%	63.108	574.84	0.2323	0.0732	1.3626	0.4437
Heatmap derived	One hot	81.86%	626.91	3453.1	21.555	9.2268	69.694	0.3015
Naive R2C2R	One hot	89.86%	1739.4	9990.2	39.653	22.812	75.474	0.3547
Naive R2C2R (ensemble)	One hot	90.29%	1661.5	10,607	29.803	20.506	53.173	0.3395
Naive TS R2C2R	One hot	92.57%	3133.1	13,212	52.697	28.582	116.43	0.3499
Naive TS R2C2R (ensemble)	One hot	87.50%	572.89	4783.6	23.842	16.767	41.376	0.3415
CP Statistical Correction	One hot	97.43%	3.6e + 07	5.6e + 08	542.33	109.93	7733.2	0.2652
CP Max Nonconformity	One hot	95.36%	2.8e + 07	4.1e + 08	483.26	101.36	6759.5	0.2520
CP Ellipsoidal	One hot	95.00%	2971.7	19,684	73.406	33.235	251.00	0.2869
M-R2CCP	One hot	93.21%	**66.044**	65.039	51.268	33.728	76.817	0.2756
M-R2C2R	One hot	93.71%	68.323	62.958	53.617	34.928	82.819	0.2922
M-R2C2R (APS)	One hot	93.07%	4440.1	24,056	**51.067**	27.707	94.819	0.2765

improve efficiency over axis-aligned sets by exploiting anisotropy in predicted covariance. The proposed M-R2CCP and M-R2C2R are the most efficient across datasets, with median sizes and IQRs comparable to or better than alternatives (Tables 2-4; examples in Fig. 3).

Adaptivity of prediction regions

We quantify adaptivity via the Spearman rank correlation ρ_s between per-example region size and realized error (Section 5.1). Larger ρ_s indicates regions expand for more complex cases, aiding conditional interpretability. We do not claim finite-sample conditional coverage per image/landmark; like most practical CP constructions, our guarantees are marginal.

Related to adaptivity, two empirical caveats merit attention: the high ρ_s observed for MC dropout on MML stems from the landmark-sampling decoding pipeline: decoding a landmark separately for each stochastic sample and averaging can inflate both covariance and point errors, inducing a mechanical correlation between region size and error (C). Second, the APS variant of M-R2C2R may produce very large regions when the predicted heatmap is nearly flat (very high model uncertainty), yielding heavy right tails in mean±SD; medians and IQRs remain comparable to other methods. In practice, such rare cases are best surfaced as low-confidence flags rather than displayed as oversized regions; hence, we report both mean±SD and median (IQR).

From learned regions to displays

Learned regions can be non-convex or fragmented. When desired, we propose displaying simple conservative summaries alongside raw regions. Convex supersets (e.g., convex hulls) provide smooth overlays whose coverage cannot decrease relative to the original region, preserving marginal validity while trading some efficiency. Applications can expose both views (raw and summarized) and flag highly fragmented cases. Future work will investigate algorithms that directly produce compact convex regions (e.g., minimum-volume or shape-regularized convex sets) to further simplify presentation without sacrificing coverage unduly.

Per-landmark versus image-level guarantees

Our guarantees are finite-sample per-landmark (marginal) coverage under exchangeability; they do not imply that all n_l landmarks in an image are simultaneously covered. Consequently, image-level success typically decreases with n_l, heuristically, under independence, $\mathbb{P}(\text{all covered}) \approx (1-\alpha)^{n_l}$, while a union bound gives $\mathbb{P}(\text{at least one miss}) \leq n_l\alpha$. When image-level control is required, one may trade efficiency for stronger joint protection via family-wise corrections across landmarks (Bonferroni with per-landmark level $\alpha' = \alpha/n_l$; Šidák under independence with $\alpha' = 1-(1-\alpha)^{1/n_l}$), or by calibrating a max-over-landmarks nonconformity that yields a single image-level region. If spatial consistency is enforced by post-processing (e.g.,

shape/graph priors, largest-component views, convex hulls), we recommend recalibrating with the operator in place to preserve nominal levels. While clinicians may not routinely reason at the all-landmarks-simultaneously level, joint control can be valuable for derived metrics (e.g., distances/angles) and robust decision rules.

Calibration behavior and training-conditional guarantees

Split CP requires a held-out calibration set of size n_{cal}, trading training signal for calibration precision. Realized marginal coverage fluctuates around $1-\alpha$ with variability $\mathcal{O}(1/\sqrt{n_{\text{cal}}})$; reporting a binomial confidence interval makes this uncertainty explicit. Beyond marginal coverage, the training-conditional coverage $\mathbb{P}(Y_{n+1} \in C(X_{n+1}) \mid D_n)$ under i.i.d. data and split CP with a fixed score stochastically dominates $\text{Beta}((1-\alpha)(n_{\text{cal}}+1),\ \alpha(n_{\text{cal}}+1))$ and equals it for continuous scores, implying concentration around $1-\alpha$ with variance at most $\alpha(1-\alpha)/(n_{\text{cal}}+2)$ (Angelopoulos et al., 2024, Sec. 4.1, Thm. 4.1). A PAC-style variant provides high-probability control by running split CP at $1-\alpha'$ with $\alpha' = \alpha - \mathcal{O}(n_{\text{cal}}^{-1/2})$ chosen via the Beta CDF (or a sub-Gaussian bound), ensuring

$$\mathbb{P}\Big(\mathbb{P}(Y_{n+1} \notin C(X_{n+1}) \mid D_n) \le \alpha\Big) \ge 1-\delta.$$

When data are scarce, data-reuse schemes such as cross-conformal (Vovk, 2015) or jackknife+ (Barber et al., 2021) reduce variance versus a single split while retaining exchangeability-based guarantees. All guarantees assume distributional similarity between calibration and deployment; under shift (e.g., scanners, hospitals, populations), recalibration on in-domain samples, covariate-shift-aware weighting (Tibshirani et al., 2019), Mondrian calibration (Vovk et al., 2003), and out-of-distribution (OOD) detection or abstention are advisable.

Practical guidance for deployment

Efficiency (region size) is sensitive to discretization and decoding, whereas coverage is robust post-calibration. We recommend aligning the conformal grid with the backbone's deterministic heatmap resolution; pushing to finer grids beyond the output stride typically yields diminishing gains while increasing memory and latency. For M-R2CCP, upsampling does not change the result because the refinement mapping is already encoded in the score. Prefer bilinear (2D) or trilinear (3D) interpolation over nearest-neighbor when mapping between grids: coverage remains intact after calibration, and smoother interpolation reduces jagged boundaries and often improves efficiency and visualization. Use M-R2CCP/M-R2C2R when uncertainty is anisotropic or multimodal (often yielding tighter, possibly non-convex regions). The APS variant of M-R2C2R offers a simple monotone probability-mass accumulation scheme but can return very large sets on near-flat heatmaps; this behavior is helpful for flagging or withholding predictions in low-confidence cases or OOD cases.

Computational overhead and code availability

Our CP procedures are model-agnostic and add negligible overhead beyond a single forward pass of the chosen backbone; optional sampling-based UQ (MC dropout, test-time augmentation, ensembles) is the main additional cost. We want to clarify that (i) conformal region construction only adds negligible overhead beyond a single forward pass (at most 0.5s for inference), and (ii) the dominant cost stems only from optional sampling-based UQ (MC dropout, TTA, ensembles), which can be parallelized or omitted if desired. Code and experiment scripts are available in a public repository,[1] and we integrate the CP-based UQ methods into `landmarker`[2] (PyTorch), which provides standardized data loaders, pretrained backbones, inference pipelines, and reference implementations.

[1] https://github.com/predict-idlab/landmark-uq
[2] https://github.com/predict-idlab/landmarker

Limitations and threats to validity

Key limitations include reliance on exchangeability between calibration and deployment, potential efficiency degradation under severe domain shift, occasional non-convex/fragmented regions that may hinder readability, and the lack of simultaneous image-level guarantees without additional corrections. Human-in-the-loop effects (e.g., how region shape/size influences reading time and error) were not measured and warrant study.

Future directions

Promising directions include covariate-shift-aware CP (e.g., weighting, Mondrian partitioning), training-conditional and locally adaptive variants that modulate scores by difficulty while preserving marginal validity, structured multi-landmark constructions with family-wise control, and systematic human-factors evaluations to quantify clinical impact.

7. Conclusion

Commonly used UQ approaches for landmark localization (e.g., MC dropout, deep ensembles, test-time augmentation) coupled with normality assumptions often underestimate total predictive uncertainty. We introduce a conformal prediction framework for anatomical landmark localization that delivers finite-sample marginal validity and practical efficiency in both 2D and 3D. In extensive experiments, M-R2CCP and M-R2C2R match nominal coverage while producing tighter regions than axis-aligned and ellipsoidal baselines, and can represent flexible, non-convex uncertainty patterns. These properties make the resulting confidence regions well-suited for clinical workflows. Future work could explore extending these methods to handle missing landmarks, investigating alternative interpolation techniques for discrete-to-continuous mapping, extending to other multi-output regression tasks, and integrating decision rules that leverage the prediction regions.

CRediT authorship contribution statement

Jef Jonkers: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization; **Frank Coopman:** Writing – review & editing, Conceptualization; **Luc Duchateau:** Writing – review & editing, Supervision, Conceptualization; **Glenn Van Wallendael:** Writing – review & editing, Supervision, Conceptualization; **Sofie Van Hoecke:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Data availability

All data are openly available except the CHD dataset, which is proprietary. The code related to this work can be found here: https://github.com/predict-idlab/landmark-uq. The methods presented in this work are also implemented in the `landmarker` package.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:
Jef Jonkers reports financial support was provided by Research Foundation Flanders. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding sources

Jef Jonkers is funded by the Research Foundation Flanders (FWO, Ref. 1S11525N). Part of the research was funded by the Research Foundation Flanders through an FWO Senior Research project on Hybrid AI

and Uncertainty Quantification for Trustworthy Time-to-Event Predictions (Ref. G0AH525N) and by the Flanders AI Research Program.

Appendix A. Proofs

Proposition 1. *Suppose that $Z_{1:n+1}$ are exchangeable, and $\alpha_t = \frac{\alpha}{d}$ (Bonferroni correction). Then, the prediction region defined by Algorithm 2 satisfies the marginal coverage guarantee in Eq. (1).*

Proof. The validity of each ICP follows from Proposition 4.1 in Vovk et al. (2022), and by applying Boole's inequality of the individual prediction intervals,

$$\mathbb{P}_{Z\sim P_Z}\left(Y_{n+1} \notin \hat{C}_\alpha(X_{n+1}; Z_{1:n})\right) \leq \alpha \tag{A.1}$$

$$\mathbb{P}_{Z\sim P_Z}\left(\bigcup_{j=1}^{d} Y_{n+1,j} \notin \hat{C}_{\alpha_t,j}(X_{n+1}; Z_{1:n})\right) \leq \alpha \tag{A.2}$$

$$\sum_{j=1}^{d} \mathbb{P}_{Z\sim P_Z}\left(Y_{n+1,j} \notin \hat{C}_{\alpha,j}(X_{n+1}; Z_{1:n})\right) \leq \alpha \tag{A.3}$$

which shows that inequality can only be guaranteed when $\alpha_t \leq \frac{\alpha}{d}$. □

Proposition 2. *Suppose a transformation $g : \mathcal{Y} \to \mathcal{P}(\mathcal{Y}')$, where $g(y)$ maps each $y \in \mathcal{Y}$ to a set in $\mathcal{Y}'$, and assuming we have a marginal coverage guarantee on $Y_{n+1} \subseteq \mathcal{Y}$, then the marginal coverage guarantee holds under the transformation of the response and prediction set,*

$$\mathbb{P}_{Z\sim P_Z}\left(g(Y_{n+1}) \in g\left(\hat{C}_\alpha(X_{n+1}; Z_{1:n})\right)\right) \geq 1-\alpha, \tag{10}$$

provided that the transformed prediction set is constructed as

$$g\left(\hat{C}_\alpha(X_{n+1}; Z_{1:n})\right) = \bigcup_{y\in\hat{C}_\alpha(X_{n+1};Z_{1:n})} g(y). \tag{11}$$

Proof. Starting from the marginal coverage guarantee in the original domain (see Eq. (1)), we aim to show that this extends to the transformed domain. By the construction of the transformed predictive set: the set $g(\hat{C}_\alpha(X_{n+1}; Z_{1:n})) = \bigcup_{y\in\hat{C}_\alpha(X_{n+1};Z_{1:n})} g(y)$ contains all possible transformed values $g(y)$ for $y \in \hat{C}_\alpha(X_{n+1}; Z_{1:n})$. By definition of the transformation g, the event $g(Y_{n+1}) \cap g\left(\hat{C}_\alpha(X_{n+1}; Z_{1:n})\right) \neq \emptyset$ is true if and only if there exists $y \in \hat{C}_\alpha(X_{n+1}; Z_{1:n})$ such that $g(Y_{n+1})$ overlaps with $g(y)$. Since $Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; Z_{1:n})$ implies that $g(Y_{n+1})$ is included in $\hat{C}'_\alpha(X_{n+1}; Z_{1:n})$, the inclusion relationship is preserved:

$$\mathbb{P}_{Z\sim P_Z}\left(Y_{n+1} \in \hat{C}_\alpha\left(X_{n+1}; Z_{1:n}\right)\right) \leq \mathbb{P}_{Z\sim P_Z}\left(g(Y_{n+1}) \cap g\left(\hat{C}_\alpha(X_{n+1}; Z_{1:n})\right) \neq \emptyset\right).$$

Finally, since the original coverage guarantee holds for Y_{n+1} in the untransformed space (Eq. (1)), the transformed set $g\left(\hat{C}_\alpha(X_{n+1}; Z_{1:n})\right)$ constructed as a union ensures the marginal guarantee in the transformed space:

$$\mathbb{P}_{Z\sim P_Z}\left(g(Y_{n+1}) \cap g\left(\hat{C}_\alpha(X_{n+1}; Z_{1:n})\right) \neq \emptyset\right) \geq 1-\alpha.$$

□

Proposition 3. *Suppose the transformation of the response domain is denoted by $g : \mathcal{Y} \to \mathcal{Y}'$, and assuming we have a marginal coverage guarantee (see Eq. (1)) on $Y_{n+1} \subseteq \mathcal{Y}$, then the marginal coverage guarantee holds under the transformation of the response and prediction set,*

$$\mathbb{P}_{Z\sim P_Z}\left(g(Y_{n+1}) \in g\left(\hat{C}_\alpha(X_{n+1}; Z_{1:n})\right)\right) \geq 1-\alpha, \tag{12}$$

when g is a monotonic OR invertible transformation.

Proof. The marginal coverage guarantee in Eq. (12) for monotonic and invertible transformation g are guaranteed because the inclusion relationships are preserved under transformation g. For the invertible transformation, this evident invertibility ensures that no information is lost, so $Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; Z_{1:n})$ is equivalent to $g(Y_{n+1}) \in g\left(\hat{C}_\alpha(X_{n+1}; Z_{1:n})\right)$. The same equivalence is present for monotonic transformation since the transformation preserves the relative ordering of elements. □

Appendix B. Extensive literature review

B.1. Uncertainty quantification in landmark localization

B.1.1. Dropout variational inference (Monte Carlo dropout)

Drevický and Kodym (2020) utilize Monte Carlo (MC) dropout Gal and Ghahramani (2016) to quantify the uncertainty of a static heatmap regression model, deriving solely approximation (epistemic) uncertainty from the variance in MC sample predictions.

Lee et al. (2020) adopt a two-stage model combining CNN-based region-of-interest (ROI) detection with a coordinate regression model using a CNN. They apply MC dropout in the second stage, use the variance of the MC samples as an uncertainty estimate, and leverage it for producing 95% confidence regions, assuming that the uncertainty distribution is a bivariate normal. The reliability of these regions is not validated, which is crucial for use in a clinical context.

Additionally, Jafari et al. (2022) applies MC dropout to measure pixel-level epistemic uncertainty in static heatmap regression. Epistemic uncertainty is leveraged to detect out-of-distribution (OOD) frames, as the authors believe and demonstrate that OOD frames generally have higher epistemic uncertainty. They demonstrate this on the critical video frame landmark detection tasks in ultrasound imaging videos of the heart. Besides quantifying epistemic uncertainty, in Jafari et al. (2022), they also quantify pixel-level aleatoric uncertainty, as they expect that in non-key frames, the regions around an obscured landmark region have high aleatoric uncertainty. They learn the aleatoric uncertainty by performing variational inference on the pixel level.

B.1.2. Test-time augmentations

Ma et al. (2020) introduce a test-time augmentation (TTA) strategy to improve prediction accuracy and robustness while generating 90% prediction regions. Their method involves applying random translation shifts as the augmentation technique. To form the prediction regions, they calculate the standard deviations of the TTA predictions along the three spatial axes, assuming that the error distribution follows an uncorrelated multivariate Gaussian. The evaluation of uncertainty estimates is limited to examining the correlation between the size of these prediction regions and the localization error.

B.1.3. Gaussian processes

Schobs et al. (2023a) propose a two-stage method for uncertainty quantification using Gaussian processes. The initial stage employs a static heatmap regression model to generate a coarse landmark prediction. This prediction defines a region of interest, which is input for a convolutional Gaussian process (CGP). The CGP then produces a distribution of probable landmark locations, modeled as a bivariate normal distribution, capturing total uncertainty. The authors primarily assess the model's fit to the bivariate normal distribution but do not evaluate the validity of the uncertainty estimates comprehensively. Additionally, this approach's point prediction performance is considerably lower than that of other methods.

B.1.4. Ensemble learning

In Drevický and Kodym (2020), Schobs et al. (2023b), they propose to use an ensemble model consisting of independently trained and randomly initialized static heatmap regression models; the uncertainty is quantified by measuring the variance in the landmark predictions of the ensembles. This approach only considers epistemic uncertainty, more specifically, approximation uncertainty. However, it is argued that this approach generally performs better than Monte Carlo dropout methods due to more decorrelated sampled inference models (Fort et al., 2020).

B.1.5. Maximum likelihood estimation

McCouat and Voiculescu (2022) approach landmark localization as a classification problem by utilizing one-hot encoded heatmaps as targets. They implement an encoder-decoder network architecture and apply a 2D softmax function separately to each output channel, with each

Table B.1
Overview of related works on uncertainty quantification for anatomical landmark localization.

Work	Approach	Uncertainty repres.	Epistemic			Validity eval.	Adaptivity eval.	Dim.
			Aleatoric	Model	Approx.			
Lee et al. (2020)	MC dropout	Variance, region	✗	✗	✓	✗	✗	2D
Drevický and Kodym (2020)	MC dropout, ensemble, heatmap derived	Variance	✗	✗	✓	✗	✗	2D
Jafari et al. (2022)	MC dropout, variational inference	Variance	✓	✗	✓	✗	✗	2D
Ma et al. (2020)	TTA	Region	✓	✗	✗	✗	✓	3D
Schobs et al. (2023b)	Ensemble, heatmap derived	Variance	?	?	?	✓	✓	2D
Schobs et al. (2023a)	Gaussian process	Covariance	✓	✗	✓	✗	✗	2D
Kwon et al. (2021)	Heatmap derived	Covariance	✓	✗	✗	✗	✗	2D
Payer et al. (2020); Thaler et al. (2021)	Heatmap derived	Covariance, entropy	?	?	?	✗	✗	2D/3D
Kumar et al. (2019, 2020)	MLE	Covariance	✓	✗	✗	✗	✓	2D
McCouat and Voiculescu (2022)	MLE, heatmap derived	Expected radial error	✓	✗	✗	✓	✓	2D

✓ Yes ✗ No ? Not Clear

channel corresponding to a specific landmark's heatmap. To improve model calibration, they use temperature scaling. This approach produces non-parametric, contour-hugging heatmaps closely following anatomical features, such as the skull in cephalometric landmark detection. The authors assess the reliability of pixel-wise probability estimates, which are generally very low (below 0.10), achieving an expected calibration error (ECE) of 0.7. However, since the probabilities are so small, the effectiveness of calibrated confidence regions remains uncertain, as the study does not implement or evaluate these regions. The authors introduce the expected radial error (ERE) as an uncertainty estimate, which shows a strong correlation (0.96) with the true radial error and can be used to identify potentially inaccurate predictions. They use this uncertainty estimate to flag erroneous predictions (ERE $\geq$ 4 mm).

Another approach is to leverage an intermediate heatmap approach and use a maximum likelihood estimation, assuming a specific probability distribution and an estimated covariance matrix. In Kumar et al. (2019), they assume a bivariate Gaussian distribution and minimize the negative log-likelihood. They evaluate the uncertainty estimate by plotting the estimated uncertainty (represented as the determinant of the covariance matrix) and the localization error. In Kumar et al. (2020), they extend the approach to deal with occluded landmarks.

B.1.6. Infering uncertainty from heatmaps

Drevický and Kodym (2020) introduce normalized maximum heatmap activation (MHA) from a static heatmap regression model as a confidence measure for landmark predictions. A limitation of this method is that it only identifies landmarks at the peak value of the heatmap, lacking subpixel accuracy. Additionally, the confidence metric lacks a true statistical interpretation of uncertainty, instead serving as a correlate of confidence. Determining the source of uncertainty quantified is challenging, though it likely encompasses elements of both aleatoric and epistemic uncertainty. Schobs et al. (2023b) enhance this approach by using an ensemble of models to compute an ensemble-MHA, averaging MHA values across ensemble predictions for greater robustness.

In Payer et al. (2020), Thaler et al. (2021), they propose to learn the Gaussian covariances of target heatmaps such that they represent, as they call, a *dataset-based annotation uncertainty*. This could be seen as homoscedastic aleatoric uncertainty. They learn these covariance parameters of the anisotropic Gaussian function $(\theta_i, \sigma_i^{maj}, \sigma_i^{\min})$ simultaneously with weights of the network. For the ith landmark, θ_i represents the rotation of the Gaussian function's major axis, and σ_i^{maj} and $\sigma_i^{\min}$ present its extent in major and minor axis, respectively, for the ith landmark. They add to the mean squared error loss function, which minimizes the predicted $\hat{h}_i$ and target h_i heatmaps, a regularization term which penalizes $\sigma_i^{maj}\sigma_i^{\min}$ with α to avoid that $\sigma_i^{maj}\sigma_i^{\min} \to \infty$ and $h_i \approx 0$. The authors also propose quantifying so-called *sample-based annotation uncertainty*, which you could see as a combination of aleatoric and epistemic uncertainty, by fitting a Gaussian function to the predicted landmarks using a robust least squares curve fitting method. Besides the proposed approach uncertainty quantification capabilities, they also show strong performance in terms of accuracy and robustness. They evaluate the reliability of their uncertainty estimates by plotting the relation between the learned Gaussian sizes $\hat{\sigma}_i^{maj}\hat{\sigma}_i^{\min}$ and the point prediction errors, but fail to evaluate the calibration of their fitted Gaussian distribution properly.

Payer et al. (2020), Thaler et al. (2021) also try to connect the inter-observer variability of landmark annotation and the learned homoscedastic aleatoric uncertainty, as the latter represents this inter-observer variability. To evaluate this hypothesis, Thaler et al. (2021) annotated five selected landmarks by a total of 11 annotators for 100 images and compared it with the learned homoscedastic aleatoric uncertainty and the average of the distribution parameters of the sample-based annotation uncertainty. Based on their visual comparisons of the fitted annotation Gaussian distribution, using the 11 annotations and

their proposed approach, they conclude that there is a correspondence between them. However, the estimated inter-variability through the learned homoscedastic aleatoric uncertainty and *sample-based annotation uncertainty* is consistently smaller. We believe this is primarily due to an error in their training approach since they trained their proposed method on the mean coordinate of the 11 annotations per image. Following this procedure, you are actually trying to determine the intra-variability of the mean estimate of the 11 annotators, assuming their proposed approach can even model the true aleatoric uncertainty. We believe training the model with augmented datasets would be a better approach to estimating the inter-variability. Each image is presented 11 times with a different annotator label. Additionally, due to the construction of the training procedure and loss function, the parameters of the uncertainty distribution are learned in a heuristic way, even assuming the true aleatoric uncertainty is homoscedastic and follows the bivariate normal distribution. We leave this for future work.

In Kwon et al. (2021), they propose to use intermediate heatmaps, which represent probability density function, and a softmax operation (Luvizon et al., 2018) to get the desired anatomical landmarks. They use these intermediate heatmaps to get confidence regions of the estimated landmark position. They propose to visualize it with standard deviational ellipses of 3σ, thus essentially assuming a bivariate normal distribution.

B.2. Multi-output conformal prediction

B.2.1. Density-based

The foundational work by Lei et al. (2011, 2013) introduced the first framework for multi-dimensional target conformal prediction, combining density estimation, e.g., kernel density estimation (KDE), and conformal prediction. Their method uses the estimated density evaluated at y as a nonconformity measure. Lei et al. (2015) further extend this work to functional data, which was later generalized by Diquigiovanni et al. (2021, 2022). In English and Lippert (2024), they proposed using conditional normalizing flows for producing the density estimation to deal with the computational burden, mitigate a grid search, and generate prediction regions; they propose to use Monte Carlo sampling of the multivariate label space.

B.2.2. Statistical correction

Much research has focused on constructing hyperrectangular prediction regions by considering each target dimension as a separate conformal predictor and treating it as the multiple comparison problem prevalent in statistics. Messoudi et al. (2020) proposed extending single-dimensional nonconformity scores to m-dimensional problems by applying conformal prediction for each dimension separately. To achieve the desired global coverage $(1-\alpha)$, they employed the Šidák correction, setting the individual significance levels to $\alpha_t = 1 - \sqrt[d]{1-\alpha}$. While computationally straightforward, this approach has limitations. It only maintains exact validity for independent scores (considering the individual conformal prediction approaches have an exact validity guarantee) and provides conservative validity for positively dependent nonconformity scores. A problem arises when the scores are negatively dependent, which can invalidate the prediction regions.

Concurrently, Stankeviciute et al. (2021) proposed a conformal prediction approach to provide validity guarantees for prediction regions of multi-step time-series forecasting using a Bonferroni correction, setting the individual significance levels to $\alpha_t = \frac{\alpha}{d}$. The validity proof relied on the validity guarantee inductive conformal prediction and Boole's inequality. Note that this often results in highly conservative prediction regions.

B.2.3. Copula-based

Messoudi et al. (2021) introduced a more sophisticated approach than the simple and conservative statistical correction approaches, using copulas to model dependencies between target variables. Their method uses Sklar's theorem to define individual (Sklar, 1959) to define individual significance levels α_t based on global significance level α and m-dimensional copula C. They proposed three specific copulas: the independent copula (equivalent to Šidák correction), the Gumbel copula (single-parameter estimation), and the empirical copula (non-parametric estimation). A downside of their approach is that validity guarantees are only given when the copula is known. In Mukama et al. (2024), they applied copula-based conformal prediction to the object detection problem. Copula-based conformal prediction was extended by Zhang et al. (2023), who allowed different significance values across dimensions by optimizing prediction region volume. Sun and Yu (2024) further formalized the approach, providing validity proofs under IID assumptions and using the empirical copula. They achieve validity by using two calibration sets, one for getting the cumulative distribution of the conformity score for each time-step or output variable and the other to estimate the copula. A downside of this approach is that you need quite a large calibration set.

B.2.4. Ellipsoidal prediction regions

A parallel research stream focuses on ellipsoidal prediction regions. Kuchibhotla (2021) introduced the Mahalaobis distance, a nonconformity measure, requiring both m-dimensional output predictions and an $m \times m$-dimensional sample inverse-covariance matrix. Johnstone and Cox (2021) formalized ellipsoid conformal prediction regions in the context of robust optimization and full conformal prediction. Messoudi et al. (2022) extended the method to generate object-specific ellipsoidal regions by estimating conditional covariance matrices. Henderson et al. (2024) proposed the conformal conditional linear expectation method, providing adaptive ellipsoid regions with validity guarantees by leveraging the Mahalanobis distance and ridge estimations of covariance matrix condition on the object x. Additionally, they study the asymptotic properties of the ellipsoid.

B.2.5. Multi-modal prediction regions

Several works focus on multi-output regression problems with multi-modal prediction regions. Conventionally, conformal prediction approaches often produce single-connected areas, which can be inefficient when the true conditional distribution is multi-modal. The density-based approach of Lei et al. (2011) can capture a multi-modal structure. A possible problem of the produced prediction regions of the density-based approach is that they are possibly non-convex and can be a problem for downstream tasks, such as decision-making (Johnstone and Cox, 2021). Additionally, the interpretability of these regions for end users can be confusing. Therefore, Tumu et al. (2023) proposed a method for multi-modal prediction regions, ensuring flexible but convex prediction regions, using a two-stage calibration process using two calibration sets. The first calibration set clusters residuals using KDE and defines shape template functions; the second set combines the shape template into a single nonconformity score. Besides the computational complexity of the method, a downside is that shapes are determined by the calibration set and, thus, are not fully adaptive.

Wang et al. (2023) introduced probabilistic conformal prediction (PCP), which allows discontinuous prediction regions through a sampling-based approach. Their methodology begins by fitting a conditional generative model on the training data. For calibration, the method generates K-independent prediction samples for each example and calculates the conformity score as the minimum distance between these samples and the true label. When making predictions for test examples, the approach generates K samples from the fitted distribution and constructs prediction regions by centering balls at each sample point, with radii equal to the calibrated conformity score quantile. The final prediction region is formed by taking the union of these individual balls.

Feldman et al. (2023) developed a distribution-free approach for multi-modal quantile regression using deep learning techniques. Their method uses a conditional variational autoencoder to learn a representation where the multi-output distribution becomes unimodal. In this

transformed space, they construct quantile regions using non-parametric directional quantile regression, which are then mapped back to the original space and conformally calibrated. While the approach effectively handles complex multi-modal structures and provides theoretical guarantees, it requires substantial training data and involves computationally intensive transformation steps that may lose some fine structural details.

B.2.6. Mixed-variables types

Dheur et al. (2024) addressed bivariate prediction with mixed continuous-discrete variables, proposing the conformal HDR (C-HDR) method, which generalizes the high predictive density approach (Izbicki et al., 2022). While they only consider the bivariate case, this generalization can be used for higher dimensions multi-output problems with continuous and discrete outcomes.

B.2.7. Other approaches

Notably, Diquigiovanni et al. (2022) introduced the maximum nonconformity score $S_i := \max\left(|Y_{i,1} - \hat{Y}_{i,1}|, \ldots, |Y_{i,d} - \hat{Y}_{i,d}|\right)$, which has been leveraged in many subsequent approaches. Dietterich and Hostetler (2022) built upon this and proposes SBox and SQbox, which adapt the maximum conformity score by replacing the absolute error as measure with a normalized absolute error (Papadopoulos and Haralambous, 2011) and CQR (Romano et al., 2019). This evolution continued with Cleaveland et al. (2024), who proposed a time series approach using linear complementary programming that generates prediction regions over multiple time steps. Their method introduces a weighted maximum nonconformity score $S_i := \max(a_1 S_{i,1}, \ldots, a_T S_{i,T})$ across time horizon T, where parameters a_t sum to one and are optimized using two calibration sets-one for parameter optimization to minimize region size, and another for computing nonconformity scores. While this approach yields less conservative regions than the Bonferroni correction, it requires re-optimization for different significance levels.

Appendix C. Additional experiments results

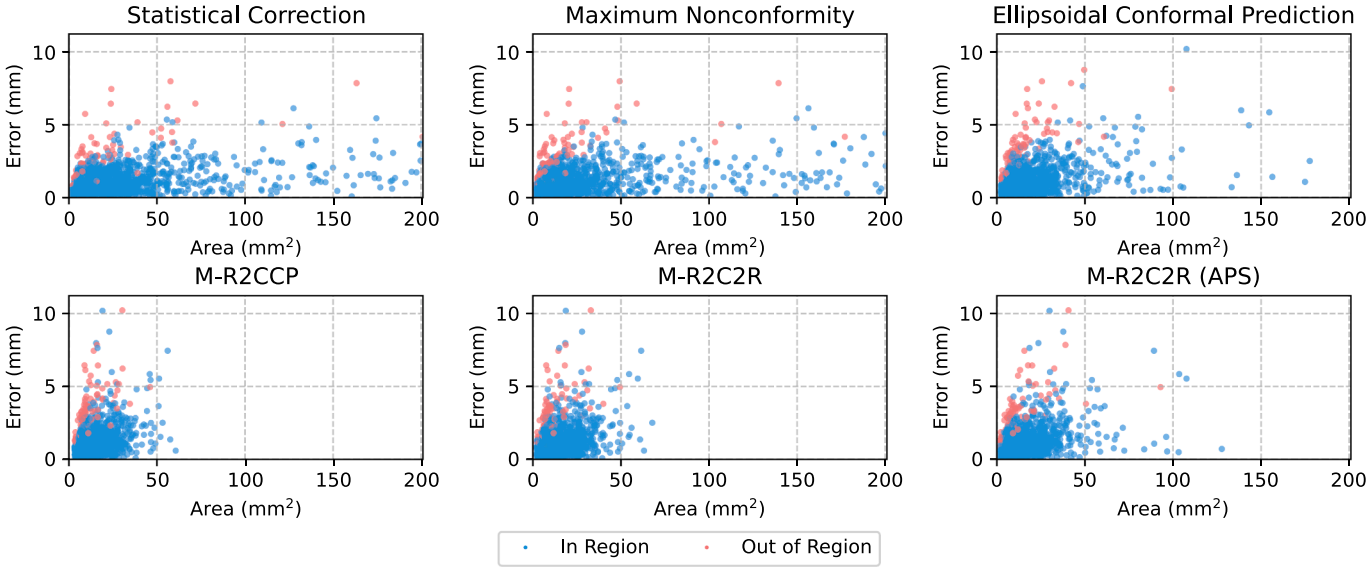

Fig. C.1. Scatter plot that illustrates the relationship between prediction error and efficiency (area) on the ISBI 2015 (2D) dataset.

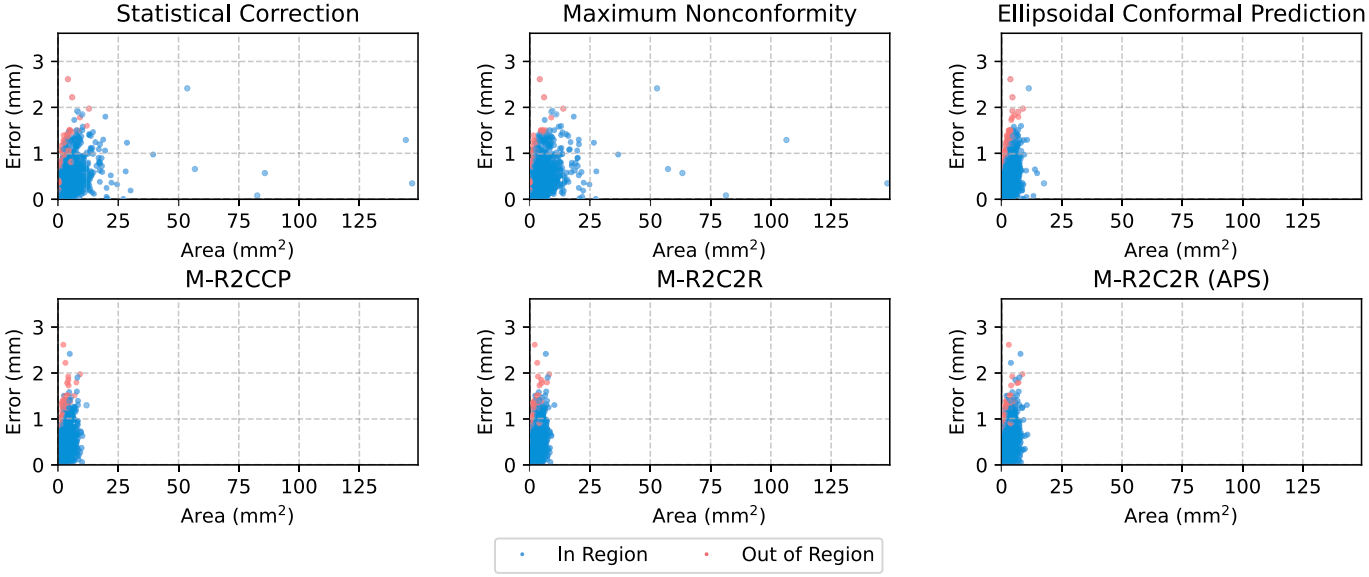

Fig. C.2. Scatter plot that illustrates the relationship between prediction error and efficiency (area) on the CHD (2D) dataset.

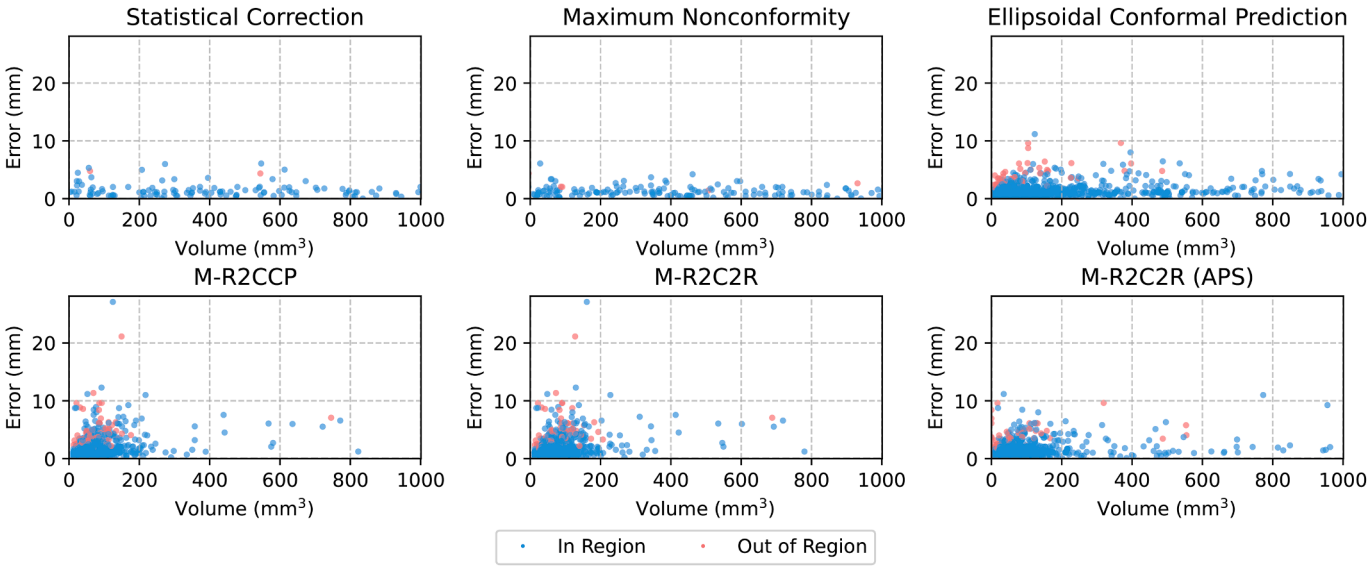

Fig. C.3. Scatter plot that illustrates the relationship between prediction error and efficiency (volume) on the MML (3D) dataset.

Table C.1
Anatomical landmark localization performance results on ISBI 2015 (2D) dataset. The best results are highlighted in bold.

Model name	Validation					Test				
	PE (mm)	SDR (%)				PE (mm)	SDR (%)			
		2 mm	2.5 mm	3 mm	4 mm		2 mm	2.5 mm	3 mm	4 mm
SCN	1.1446	86.26%	90.53%	93.84%	97.21%	1.2051	83.58%	89.68%	93.32%	96.84%
SCN ensemble	1.101	87.21%	91.16%	94.42%	97.58%	1.1672	84.95%	90.68%	93.89%	97.16%
SCN ensemble LS	**1.084**	87.21%	91.53%	95.11%	97.74%	1.1463	85.37%	91.37%	94.47%	97.42%
SCN MC	1.1364	86.11%	90.63%	94.05%	97.32%	1.2065	83.47%	89.84%	93.11%	96.95%
SCN MC LS	1.1255	86.58%	91.00%	94.47%	97.32%	1.1888	84.53%	90.32%	94.00%	97.11%
SCN MC TTA	1.0977	87.37%	91.16%	94.58%	97.47%	1.1617	85.05%	90.42%	93.68%	96.89%
SCN MC TTA LS	1.1029	86.95%	92.11%	94.68%	97.63%	1.163	85.58%	90.79%	93.58%	97.42%
SCN TTA	1.0939	87.21%	91.68%	94.42%	97.53%	1.1612	84.95%	90.42%	93.42%	97.11%
SCN TTA LS	1.1128	86.26%	91.74%	94.68%	97.68%	1.1654	85.00%	90.63%	93.74%	97.16%
One hot	1.3149	84.26%	90.47%	93.16%	96.58%	1.2502	84.53%	90.11%	93.58%	96.68%
One hot ensemble	1.0861	**88.11%**	**93.16%**	**95.95%**	**98.32%**	**1.116**	**87.42%**	**93.00%**	**95.53%**	**98.05%**
One hot ensemble LS	1.123	87.37%	92.47%	95.11%	97.68%	1.1366	86.74%	92.37%	94.89%	97.84%
One hot TTA	1.1958	86.32%	90.95%	94.05%	97.16%	1.1564	86.21%	91.37%	93.89%	97.21%
One hot TTA LS	1.1958	86.42%	91.74%	94.42%	97.47%	1.1688	86.11%	91.00%	94.05%	97.42%

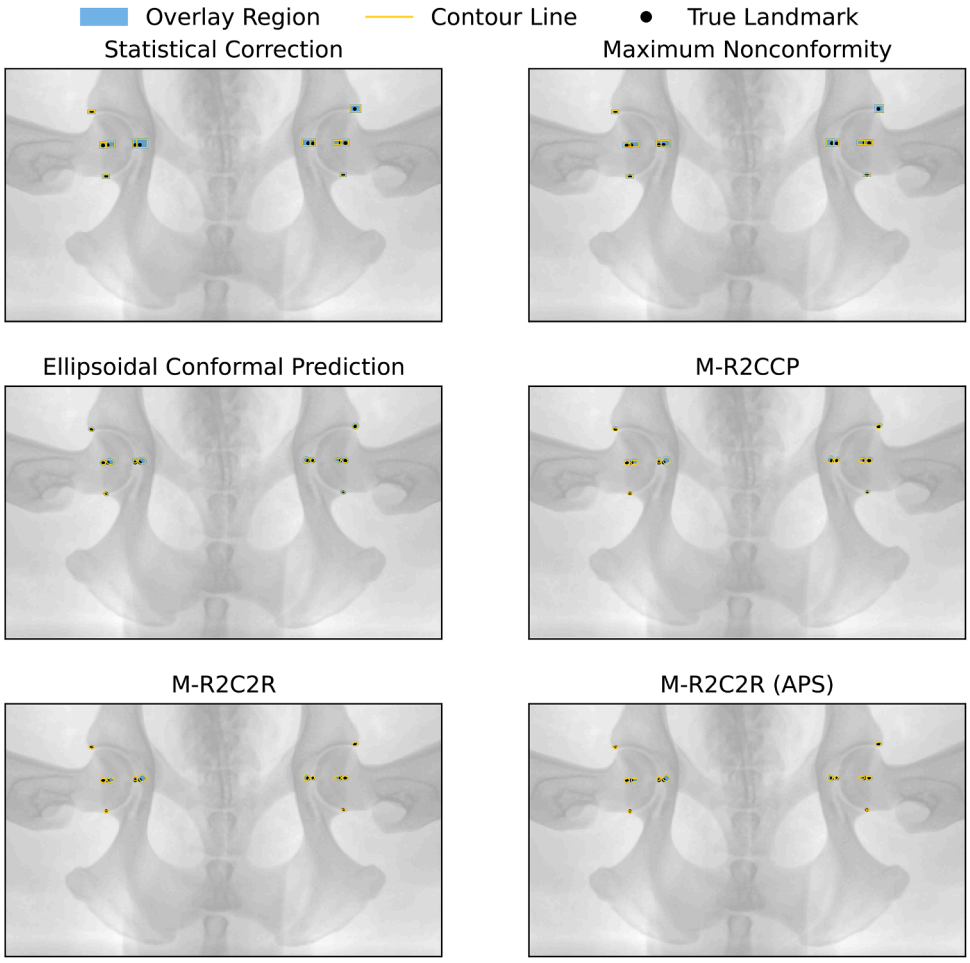

Fig. C.4. Canine pelvis example of conformal prediction regions.

Table C.2
Anatomical landmark localization performance results on CHD (2D) dataset. The best results are highlighted in bold.

Model name	Validation					Test				
	PE (mm)	SDR (%)				PE (mm)	SDR (%)			
		2 mm	2.5 mm	3 mm	4 mm		2 mm	2.5 mm	3 mm	4 mm
SCN	0.4591	99.20%	99.68%	99.84%	100.0%	0.4867	99.32%	99.76%	99.95%	100.0%
SCN ENSEMBLE	0.4358	99.52%	99.84%	99.84%	100.0%	0.4670	99.51%	100.0%	100.0%	100.0%
SCN ENSEMBLE LS	0.4231	**99.76%**	**100.0%**	**100.0%**	**100.0%**	0.4569	99.76%	100.0%	100.0%	100.0%
SCN MC	0.4599	99.20%	99.68%	99.84%	100.0%	0.4863	99.27%	99.76%	99.95%	100.0%
SCN MC LS	0.4519	99.36%	99.84%	99.92%	100.0%	0.4816	99.37%	99.85%	100.0%	100.0%
SCN MC TTA	0.3934	99.36%	99.84%	99.92%	100.0%	0.4161	99.61%	99.95%	100.0%	100.0%
SCN MC TTA LS	0.3775	99.68%	99.92%	100.0%	100.0%	0.4096	99.95%	100.0%	100.0%	100.0%
SCN TTA	0.3912	99.28%	99.76%	100.0%	100.0%	0.4131	99.76%	99.90%	99.95%	100.0%
SCN TTA LS	**0.3772**	99.52%	99.92%	100.0%	100.0%	**0.4059**	**99.95%**	**100.0%**	**100.0%**	**100.0%**
One hot	0.6115	99.24%	99.58%	99.65%	99.79%	0.5601	99.17%	99.72%	99.86%	99.86%
One hot ensemble	0.4835	99.72%	99.79%	99.93%	99.93%	0.4750	99.79%	99.93%	**100.0%**	**100.0%**
One hot ensemble LS	0.5646	98.89%	98.96%	99.17%	99.24%	0.4987	99.38%	99.52%	99.59%	99.66%
One hot TTA	0.6179	99.31%	99.72%	99.72%	99.79%	0.5672	99.10%	99.72%	99.86%	99.86%
One hot TTA LS	0.5998	98.96%	99.31%	99.51%	99.58%	0.5630	99.31%	99.79%	99.93%	99.93%

Table C.3
Anatomical landmark localization performance results on MML (3D) dataset. This benchmark dataset is the same as the data subset, only with complete landmarks, used as a benchmark in He et al. (2024). The best results are highlighted in bold.

Model Name	Validation					Test				
	PE (mm)	SDR (%)				PE (mm)	SDR (%)			
		2 mm	2.5 mm	3 mm	4 mm		2 mm	2.5 mm	3 mm	4 mm
One hot	1.983	71.81%	80.23%	84.57%	90.31%	1.6938	77.50%	86.43%	90.24%	94.29%
One hot ensemble	**1.601**	**77.81%**	**87.76%**	**89.92%**	93.37%	**1.3918**	**81.67%**	**91.31%**	**93.33%**	96.31%
One hot ensemble LS	1.632	76.02%	85.46%	89.41%	**93.62%**	1.4694	81.67%	89.40%	93.10%	**96.55%**
One hot MC	2.0249	73.09%	84.06%	86.86%	91.20%	1.6392	77.98%	87.98%	90.48%	94.05%
One hot MC LS	2.0557	71.30%	79.21%	84.57%	90.18%	1.6433	77.98%	86.07%	90.12%	95.24%

Table C.4
Anatomical landmark localization performance results on MML (3D) dataset with adjusted pixel spacing. This benchmark dataset is the same as the data subset, only with complete landmarks, used as a benchmark in He et al. (2024). The best results are highlighted in bold.

Model name	Validation					Test				
	PE (mm)	SDR (%)				PE (mm)	SDR (%)			
		2 mm	2.5 mm	3 mm	4 mm		2 mm	2.5 mm	3 mm	4 mm
One hot	2.1221	69.90%	78.32%	83.93%	89.41%	1.6937	77.50%	86.43%	90.24%	94.29%
One hot ensemble	**1.6947**	**75.64%**	**86.61%**	**89.29%**	**92.86%**	**1.3922**	**81.67%**	**91.31%**	**93.21%**	96.31%
One hot ensemble LS	1.7366	74.11%	83.42%	88.14%	**92.86%**	1.4693	81.67%	89.40%	93.10%	**96.55%**
One hot MC	2.2365	71.05%	82.14%	85.46%	89.41%	1.6769	75.48%	86.79%	89.29%	92.74%
One hot MC LS	2.1655	69.13%	78.70%	83.93%	89.29%	1.6556	78.10%	85.36%	90.24%	95.24%

References

Angelopoulos, A.N., Barber, R.F., Bates, S., 2024. Theoretical foundations of conformal prediction. arXiv preprint arXiv:2411.11824.

Avisdris, N., Ben Bashat, D., Ben-Sira, L., Joskowicz, L., et al., 2021. Fetal brain MRI measurements using a deep learning landmark network with reliability estimation. In: Sudre, C.H., Licandro, R., Baumgartner, C., Melbourne, A., Dalca, A., Hutter, J., Tanno, R., Abaci Turk, E., Van Leemput, K., Torrents Barrena, J., Wells, W.M., Macgowan, C. (Eds.), Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. Springer International Publishing, Cham, pp. 210–220. https://doi.org/10.1007/978-3-030-87735-4_20

Banerji, C. R.S., Chakraborti, T., Harbron, C., MacArthur, B.D., et al., 2023. Clinical AI tools must convey predictive uncertainty for each individual patient. Nat. Med. , 1–3 https://doi.org/10.1038/s41591-023-02562-7

Barber, R.F., Candes, E.J., Ramdas, A., Tibshirani, R.J., 2021. Predictive inference with the jackknife+. Ann. Stat. 49 (1), 486–507.

Begoli, E., Bhattacharya, T., Kusnezov, D., 2019. The need for uncertainty quantification in machine-assisted medical decision making. Nature Mach. Intell. 1 (1), 20–23.

Bulat, A., Sanchez, E., Tzimiropoulos, G., et al., 2021. Subpixel Heatmap Regression for Facial Landmark Localization. [CS] https://doi.org/10.48550/arXiv.2111.02360

Chernozhukov, V., Wüthrich, K., Zhu, Y., et al., 2021. Distributional conformal prediction. Proc. Natl. Acad. Sci. 118 (48), e2107794118. Publisher: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.2107794118

Cleaveland, M., Lee, I., Pappas, G.J., Lindemann, L., et al., 2024. Conformal prediction regions for time series using linear complementarity programming. Proceedings of the AAAI Conference on Artificial Intelligence 38 (19), 20984–20992. Number: 19. https://doi.org/10.1609/aaai.v38i19.30089

de Queiroz Tavares Borges Mesquita, G., Vieira, W.A., Vidigal, M. T.C., Travençolo, B. A.N., Beaini, T.L., Spin-Neto, R., Paranhos, L.R., de Brito Júnior, R.B., et al., 2023. Artificial intelligence for detecting cephalometric landmarks: a systematic review and meta-analysis. J. Digit. Imaging https://doi.org/10.1007/s10278-022-00766-w

Dheur, V., Bosser, T., Izbicki, R., Ben Taieb, S., et al., 2024. Distribution-free conformal joint prediction regions for neural marked temporal point processes. Mach. Learn. 113 (9), 7055–7102. https://doi.org/10.1007/s10994-024-06594-z

Dietterich, T.G., Hostetler, J., 2022. Conformal Prediction Intervals for Markov Decision Process Trajectories. arXiv:2206.04860. [CS, STAT]

Diquigiovanni, J., Fontana, M., Vantini, S., et al., 2021. The Importance of Being a Band: Finite-Sample Exact Distribution-Free Prediction Sets for Functional Data. [stat] https://doi.org/10.48550/arXiv.2102.06746

Diquigiovanni, J., Fontana, M., Vantini, S., et al., 2022. Conformal prediction bands for multivariate functional data. J. Multivar. Anal. 189, 104879. https://doi.org/10.1016/j.jmva.2021.104879

Dong, X., Yu, S.-I., Weng, X., Wei, S.-E., Yang, Y., Sheikh, Y., et al., 2018. Supervision-by-registration: an unsupervised approach to improve the precision of facial landmark detectors. pp. 360–368. https://openaccess.thecvf.com/content_cvpr_2018/html/Dong_Supervision-by-Registration_An_Unsupervised_CVPR_2018_paper.html.

Drevický, D., Kodym, O., 2020. Evaluating deep learning uncertainty measures in cephalometric landmark localization:. In: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies. SCITEPRESS - Science and Technology Publications, Valletta, Malta, pp. 213–220. https://doi.org/10.5220/0009375302130220

Dunn, O.J., 1961. Multiple comparisons among means. J. Am. Stat. Assoc. 56 (293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

English, E., Lippert, C., 2024. Conformalised conditional normalising flows for joint prediction regions in time series. https://openreview.net/forum?id=5Vkwer7im5.

Feldman, S., Bates, S., Romano, Y., et al., 2023. Calibrated multiple-output quantile regression with representation learning. J. Mach. Learn. Res. 24 (24), 1–48. http://jmlr.org/papers/v24/21-1280.html.

Fontana, M., Zeni, G., Vantini, S., et al., 2022. Conformal Prediction: a Unified Review of Theory and New Challenges. [CS, ECON, STAT]. https://doi.org/10.48550/arXiv.2005.07972

Fort, S., Hu, H., Lakshminarayanan, B., et al., 2020. Deep Ensembles: A Loss Landscape Perspective. https://doi.org/10.48550/arXiv.1912.02757

Foygel Barber, R., Candès, E.J., Ramdas, A., Tibshirani, R.J., et al., 2021. The limits of distribution-free conditional predictive inference. Inf. Inf. A J. IMA 10 (2), 455–482. https://doi.org/10.1093/imaiai/iaaa017

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. arXiv:1506.02142 [CS, STAT].

Guha, E.K., Natarajan, S., Möllenhoff, T., Khan, M.E., Ndiaye, E., . Conformal prediction via regression-as-classification. In: The Twelfth International Conference on Learning Representations.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., et al., 2017. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. PMLR, pp. 1321–1330. ISSN: 2640-3498. https://proceedings.mlr.press/v70/guo17a.html.

Gupta, C., Kuchibhotla, A.K., Ramdas, A., et al., 2022. Nested conformal prediction and quantile out-of-bag ensemble methods. Pattern Recognit. 127, 108496. https://doi.org/10.1016/j.patcog.2021.108496

He, T., Xu, G., Cui, L., Tang, W., Long, J., Guo, J., et al., 2024. Anchor ball regression model for large-scale 3D skull landmark detection. Neurocomputing 567, 127051. https://doi.org/10.1016/j.neucom.2023.127051

Henderson, I., Mazoyer, A., Gamboa, F., et al., 2024. Adaptive inference with random ellipsoids through Conformal Conditional Linear Expectation. https://hal.science/hal-04489842.

Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. 110 (3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Izbicki, R., Shimizu, G., Stern, R.B., et al., 2022. CD-split and HPD-split: efficient conformal regions in high dimensions. J. Mach. Learn. Res. 23 (87), 1–32. http://jmlr.org/papers/v23/20-797.html.

Jafari, M.H., Luong, C., Tsang, M., Gu, A.N., Van Woudenberg, N., Rohling, R., Tsang, T., Abolmaesumi, P., et al., 2022. U-land: uncertainty-driven video landmark detection. IEEE Trans. Med. Imaging 41 (4), 793–804. Conference Name: IEEE Transactions on Medical Imaging. https://doi.org/10.1109/TMI.2021.3123547

Jiang, Y., Li, Y., Wang, X., Tao, Y., Lin, J., Lin, H., et al., 2022. Cephalformer: incorporating global structure constraint into visual features for general cephalometric landmark detection. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2022. Springer Nature Switzerland, Cham, pp. 227–237. https://doi.org/10.1007/978-3-031-16437-8_22

Johnstone, C., Cox, B., 2021. Conformal uncertainty sets for robust optimization. In: Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and

Applications. PMLR, pp. 72–90. ISSN: 2640-3498. https://proceedings.mlr.press/v152/johnstone21a.html.
Jonkers, J., Duchateau, L., Wallendael, G.V., Hoecke, S.V., 2025. landmarker: a toolkit for anatomical landmark localization in 2D/3D images. arXiv:2501.10098.
Kamoen, A., Dermaut, L., Verbeeck, R., et al., 2001. The clinical significance of error measurement in the interpretation of treatment results. Eur. J. Orthod. 23 (5), 569–578. https://doi.org/10.1093/ejo/23.5.569
Kompa, B., Snoek, J., Beam, A.L., 2021. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Dig. Med. 4 (1), 4.
Kuchibhotla, A.K., 2021. Exchangeability, Conformal Prediction, and Rank Tests. arXiv:2005.06095. [stat].
Kumar, A., Marks, T.K., Mou, W., Feng, C., Liu, X., et al., 2019. UGLLI Face alignment: estimating uncertainty with Gaussian log-likelihood loss. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, Seoul, Korea (South), pp. 778–782. https://doi.org/10.1109/ICCVW.2019.00103
Kumar, A., Marks, T.K., Mou, W., Wang, Y., Jones, M., Cherian, A., Koike-Akino, T., Liu, X., Feng, C., et al., 2020. LUVLi Face alignment: estimating landmarks' location, uncertainty, and visibility likelihood. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8236–8246. https://openaccess.thecvf.com/content_CVPR_2020/html/Kumar_LUVLi_Face_Alignment_Estimating_Landmarks_Location_Uncertainty_and_Visibility_Likelihood_CVPR_2020_paper.html.
Kwon, H.J., Koo, H.I., Park, J., Cho, N.I., et al., 2021. Multistage probabilistic approach for the localization of cephalometric landmarks. IEEE Access 9, 21306–21314. Conference Name: IEEE Access. https://doi.org/10.1109/ACCESS.2021.3052460
Lakshminarayanan, B., Pritzel, A., Blundell, C., et al., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.
Lee, J.-H., Yu, H.-J., Kim, M.-j., Kim, J.-W., Choi, J., et al., 2020. Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks. BMC Oral Health 20 (1), 270. https://doi.org/10.1186/s12903-020-01256-7
Lei, J., Rinaldo, A., Wasserman, L., et al., 2015. A conformal prediction approach to explore functional data. Ann. Math. Artif. Intell. 74 (1), 29–43. https://doi.org/10.1007/s10472-013-9366-6
Lei, J., Robins, J., Wasserman, L., et al., 2011. Efficient Nonparametric Conformal Prediction Regions. https://doi.org/10.48550/arXiv.1111.1418
Lei, J., Robins, J., Wasserman, L., et al., 2013. Distribution-free prediction sets. J. Am. Stat. Assoc. 108 (501), 278–287. https://doi.org/10.1080/01621459.2012.751873
Lindner, C., Wang, C.-W., Huang, C.-T., Li, C.-H., Chang, S.-W., Cootes, T.F., et al., 2016. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. Sci. Rep. 6 (1), 33581. https://doi.org/10.1038/srep33581
Luvizon, D.C., Picard, D., Tabia, H., et al., 2018. 2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning. [cs] https://doi.org/10.48550/arXiv.1802.09232
Ma, T., Gupta, A., Sabuncu, M.R., et al., 2020. Volumetric landmark detection with a multi-scale shift equivariant neural network. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 981–985. ISSN: 1945-8452. https://doi.org/10.1109/ISBI45749.2020.9098620
McCouat, J., Voiculescu, I., 2022. Contour-hugging heatmaps for landmark detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20565–20573. ISSN: 2575-7075. https://doi.org/10.1109/CVPR52688.2022.01994
Messoudi, S., Destercke, S., Rousseau, S., et al., 2020. Conformal multi-target regression using neural networks. In: Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications. PMLR, pp. 65–83. ISSN: 2640-3498. https://proceedings.mlr.press/v128/messoudi20a.html.
Messoudi, S., Destercke, S., Rousseau, S., et al., 2021. Copula-based conformal prediction for multi-target regression. Pattern Recognit. 120, 108101. https://doi.org/10.1016/j.patcog.2021.108101
Messoudi, S., Destercke, S., Rousseau, S., et al., 2022. Ellipsoidal conformal inference for multi-target regression. In: Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications. PMLR, pp. 294–306. ISSN: 2640-3498. https://proceedings.mlr.press/v179/messoudi22a.html.
Mukama, B.C., Messoudi, S., Rousseau, S., Destercke, S., et al., 2024. Copula-based conformal prediction for object detection: a more efficient approach. In: Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications. PMLR, pp. 140–157. ISSN: 2640-3498. https://proceedings.mlr.press/v230/mukama24a.html.
Papadopoulos, H., Haralambous, H., 2011. Reliable prediction intervals with regression neural networks. Neural Netw. 24 (8), 842–851. https://doi.org/10.1016/j.neunet.2011.05.008
Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A., et al., 2002. Inductive confidence machines for regression. In: Elomaa, T., Mannila, H., Toivonen, H. (Eds.), Machine Learning: ECML 2002. Springer, Berlin, Heidelberg, pp. 345–356. https://doi.org/10.1007/3-540-36755-1_29
Payer, C., Stern, D., Bischof, H., Urschler, M., et al., 2019. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med. Image Anal. 54, 207–219. https://doi.org/10.1016/j.media.2019.03.007
Payer, C., Urschler, M., Bischof, H., Štern, D., et al., 2020. Uncertainty estimation in landmark localization based on Gaussian heatmaps. In: Sudre, C.H., Fehri, H., Arbel, T., Baumgartner, C.F., Dalca, A., Tanno, R., Van Leemput, K., Wells, W.M., Sotiras, A., Papiez, B., Ferrante, E., Parisot, S. (Eds.), Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis. Springer International Publishing, Cham, pp. 42–51. https://doi.org/10.1007/978-3-030-60365-6_5
Pei, Y., Mu, L., Xu, C., Li, Q., Gan, S., Sun, B., Li, X., Li, X., et al., 2023. Learning-based landmark detection in pelvis x-rays with attention mechanism: data from the osteoarthritis initiative. Biomed. Phys. Eng. Express 9 (2), 025001. https://doi.org/10.1088/2057-1976/ac8ffa
Romano, Y., Patterson, E., Candes, E., et al., 2019. Conformalized quantile regression. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.
Romano, Y., Sesia, M., Candes, E., et al., 2020. Classification with valid and adaptive coverage. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 3581–3591. https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html.
Ryu, S.M., Shin, K., Shin, S.W., Lee, S.H., Seo, S.M., Cheon, S.-u., Ryu, S.-A., Kim, J.-S., Ji, S., Kim, N., et al., 2022. Automated landmark identification for diagnosis of the deformity using a cascade convolutional neural network (FlatNet) on weight-bearing lateral radiographs of the foot. Comput. Biol. Med. 148, 105914. https://doi.org/10.1016/j.compbiomed.2022.105914
Saunders, C., Gammerman, A., Vovk, V., et al., 1999. Transduction with confidence and credibility. pp. 722–726. https://eprints.soton.ac.uk/258961/.
Schobs, L., McDonald, T.M., Lu, H., et al., 2023a. Bayesian uncertainty estimation in landmark localization using convolutional Gaussian processes. In: Sudre, C.H., Baumgartner, C.F., Dalca, A., Mehta, R., Qin, C., Wells, W.M. (Eds.), Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. Springer Nature Switzerland, Cham, pp. 22–31. https://doi.org/10.1007/978-3-031-44336-7_3
Schobs, L.A., Swift, A.J., Lu, H., et al., 2023b. Uncertainty estimation for heatmap-based landmark localization. IEEE Trans. Med. Imaging 42 (4), 1021–1034. https://doi.org/10.1109/TMI.2022.3222730
Sesia, M., Romano, Y., 2021. Conformal prediction using conditional histograms. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 6304–6315. https://proceedings.neurips.cc/paper_files/paper/2021/hash/31b3b31a1c2f8a370206f111127c0dbd-Abstract.html.
Sklar, M., 1959. Fonctions de répartition à N dimensions et leurs marges. Annales de l'ISUP VIII (3), 229–231. Publisher: Publications de l'Institut de Statistique de l'Université de Paris. https://hal.science/hal-04094463.
Song, Y., Qiao, X., Iwamoto, Y., Chen, Y.-w., et al., 2020. Automatic cephalometric landmark detection on x-ray images using a deep-learning method. Applied Sciences 10 (7), 2547. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/app10072547
Stankeviciute, K., Ahmed, M.A., van der Schaar, M., et al., 2021. Conformal time-series forecasting. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 6216–6228. https://proceedings.neurips.cc/paper/2021/hash/312f1ba2a72318edaaa995a67835fad5-Abstract.html.
Stewart, L., Bach, F., Berthet, Q., Vert, J.-P., et al., 2023. Regression as classification: influence of task formulation on neural network features. In: Proceedings of the 26th International Conference on Artificial Intelligence and Statistics. PMLR, pp. 11563–11582. ISSN: 2640-3498. https://proceedings.mlr.press/v206/stewart23a.html.
Sun, S., Yu, R., 2024. Copula Conformal Prediction for Multi-step Time Series Forecasting. https://doi.org/10.48550/arXiv.2212.03281
Teixeira, B., Tamersoy, B., Singh, V., Kapoor, A., et al., 2019. Adaloss: Adaptive Loss Function for Landmark Localization. https://doi.org/10.48550/arXiv.1908.01070
Thaler, F., Payer, C., Urschler, M., Stern, D., et al., 2021. Modeling Annotation Uncertainty with Gaussian Heatmaps in Landmark Localization. https://doi.org/10.48550/arXiv.2109.09533
Tibshirani, R.J., Foygel Barber, R., Candes, E., Ramdas, A., 2019. Conformal prediction under covariate shift. Adv. Neural Inf. Process. Syst., Curran Associates, Inc., 32.
Tompson, J.J., Jain, A., LeCun, Y., Bregler, C., et al., 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/hash/e744f91c29ec99f0e662c9177946c627-Abstract.html.
Tumu, R., Cleaveland, M., Mangharam, R., Pappas, G.J., Lindemann, L., et al., 2023. Multi-Modal Conformal Prediction Regions by Optimizing Convex Shape Templates. arXiv:2312.07434 [cs, eess]. https://doi.org/10.48550/arXiv.2312.07434
Valdenegro-Toro, M., Mori, D.S., 2022. A deeper look into aleatoric and epistemic uncertainty disentanglement. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1508–1516. ISSN: 2160-7516. https://doi.org/10.1109/CVPRW56347.2022.00157
Vovk, V., 2012. Conditional validity of inductive conformal predictors. In: Asian Conference on Machine Learning. PMLR, pp. 475–490.
Vovk, V., 2015. Cross-conformal predictors. Ann. Math. Artif. Intell. 74 (1), 9–28.
Vovk, V., Gammerman, A., Saunders, C., et al., 1999. Machine-learning applications of algorithmic randomness. In: Proceedings of the Sixteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 444–453.
Vovk, V., Gammerman, A., Shafer, G., et al., 2022. Algorithmic Learning in a Random World. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-06649-8
Vovk, V., Lindsay, D., Nouretdinov, I., Gammerman, A., et al., 2003. Mondrian Confidence Machine. https://pure.royalholloway.ac.uk/en/publications/mondrian-confidence-machine.
Wang, C.-W., Huang, C.-T., Hsieh, M.-C., Li, C.-H., Chang, S.-W., Li, W.-C., Vandaele, R., Marée, R., Jodogne, S., Geurts, P., Chen, C., Zheng, G., Chu, C., Mirzaalian, H., Hamarneh, G., Vrtovec, T., Ibragimov, B., et al., 2015. Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: a grand challenge. IEEE Trans. Med. Imaging 34 (9), 1890–1900. https://doi.org/10.1109/TMI.2015.2412951

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., et al., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34–45. https://doi.org/10.1016/j.neucom.2019.01.103

Wang, Z., Gao, R., Yin, M., Zhou, M., Blei, D., et al., 2023. Probabilistic conformal prediction using conditional random samples. In: Proceedings of the 26th International Conference on Artificial Intelligence and Statistics. PMLR, pp. 8814–8836. ISSN: 2640-3498. https://proceedings.mlr.press/v206/wang23n.html.

Zhang, R., Zhou, P., Chai, T., et al., 2023. Improved copula-based conformal prediction for uncertainty quantification of multi-output regression. J. Process Control 129, 103036. https://doi.org/10.1016/j.jprocont.2023.103036

Zhong, Z., Li, J., Zhang, Z., Jiao, Z., Gao, X., et al., 2019. An attention-guided deep regression model for landmark detection in cephalograms. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2019. Springer International Publishing, Cham, pp. 540–548. https://doi.org/10.1007/978-3-030-32226-7_60

Zhou, Z., Li, H., Liu, H., Wang, N., Yu, G., Ji, R., et al., 2023. STAR Loss: reducing semantic ambiguity in facial landmark detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15475–15484. https://openaccess.thecvf.com/content/CVPR2023/html/Zhou_STAR_Loss_Reducing_Semantic_Ambiguity_in_Facial_Landmark_Detection_CVPR_2023_paper.html.