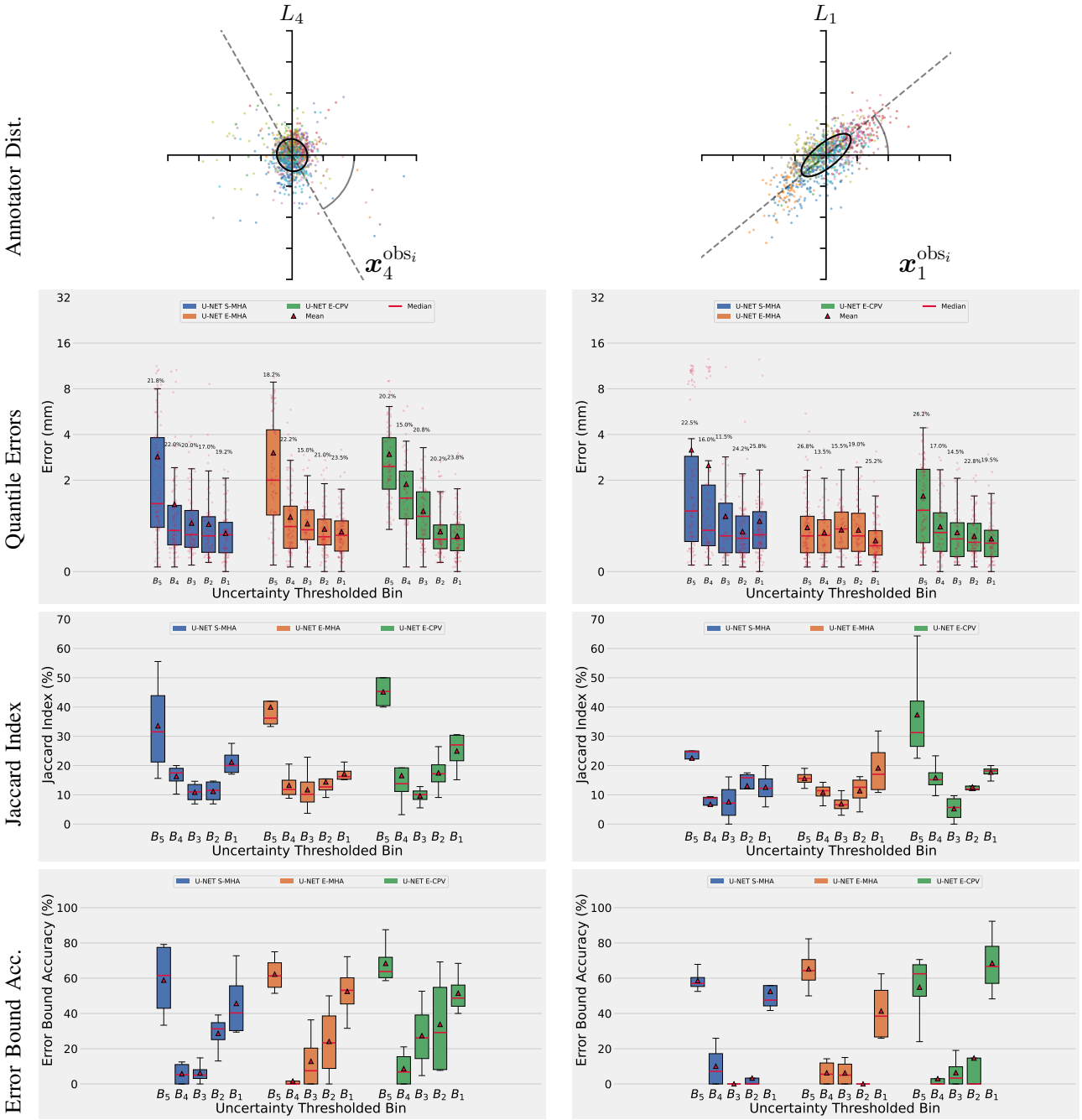
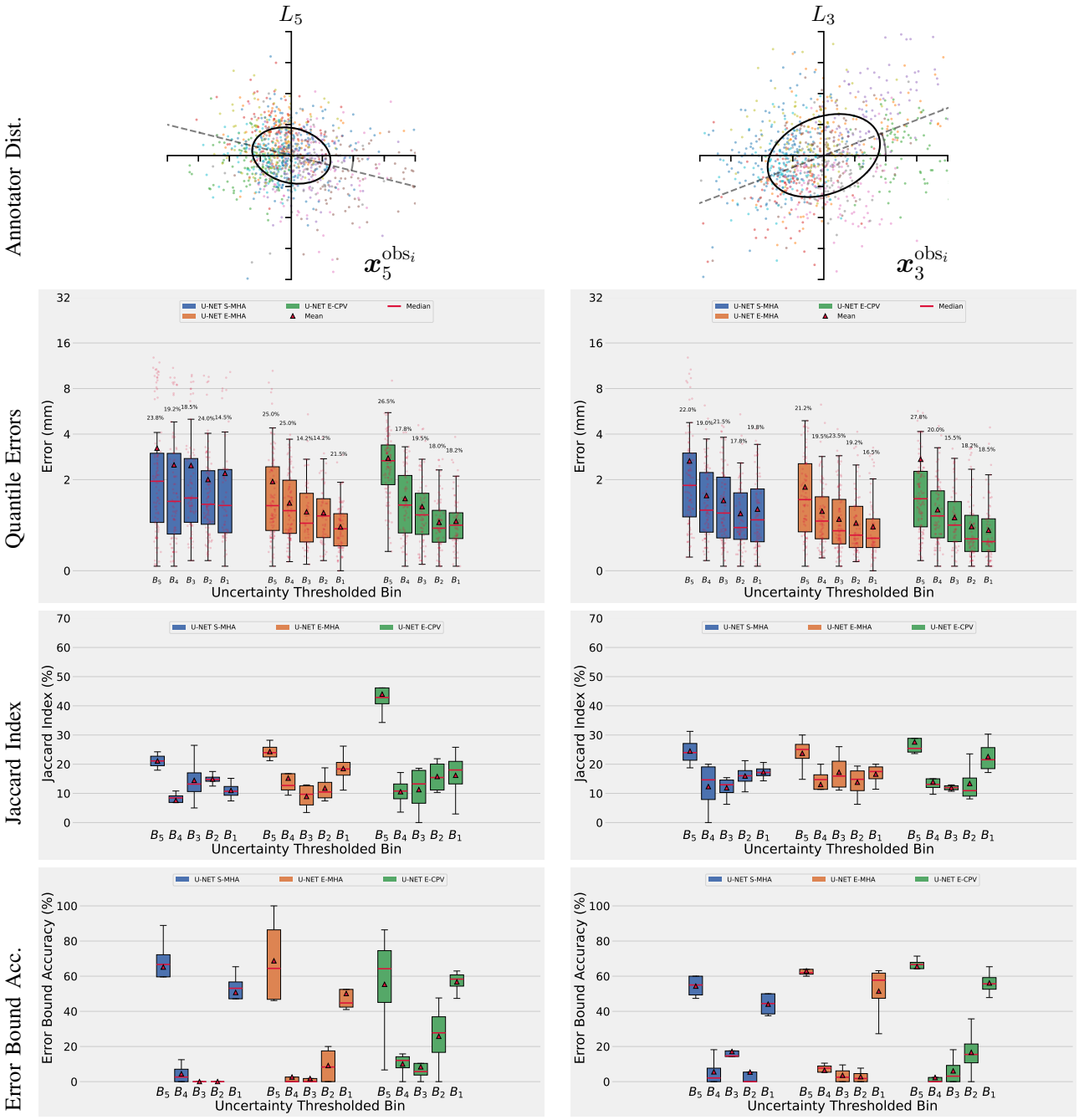
Please find additional experiments supporting our paper.

A. Aleatoric Uncertainty



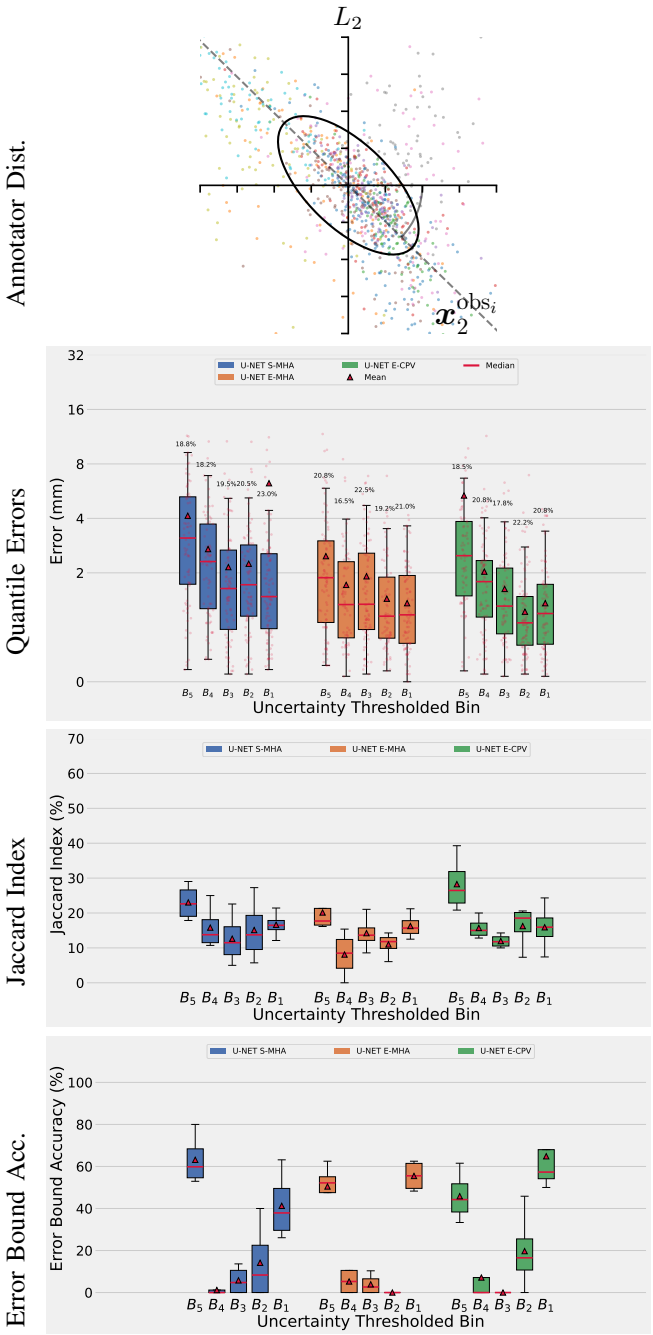
(a) Results for landmarks L_4 and L_1. L_4 has the lowest aleatoric uncertainty with an isotropic annotator distribution. L_1 has a low aleatoric uncertainty with an anisotropic (directionally skewed) annotator distribution.

Fig. 1 (Spans 3 pages, full caption on page 20)



(b) Results for landmarks L_5 and L_3. L_5 has the median aleatoric uncertainty with a mostly isotropic annotator distribution. L_3 has a high aleatoric uncertainty with a mostly isotropic annotator distribution.

Fig. 1 [Continued]. (Full caption on page 20)



(c) Results for landmark L_2. $L2$ has the highest aleatoric uncertainty with an anisotropic (directionally skewed) annotator distribution.

Fig. 1: [Continued] Quantile Binning results for a subset of landmarks from the Cephalometric dataset [1]. Row *Annotator Dist.* shows the individual offsets from each of the 11 annotators to the mean annotation of the each landmark, (data and images from [2]). The larger the fitted Gaussian, the more variance between annotators and the higher the aleatoric uncertainty. The remaining rows show the results from Quantile Binning for S-MHA, E-MHA and E-CPV uncertainty measures on each landmark over all folds. The *Quantile Errors* row shows the boxplots of localization errors for each quantile bin. The *Jaccard Index* row shows the similarity between the predicted Quantiles and the true error quantiles, and the *Error Bound Acc.* row shows the accuracy of the predicted error bounds for each quantile bin.

B. Standard Deviation of Target Heatmap Comparison

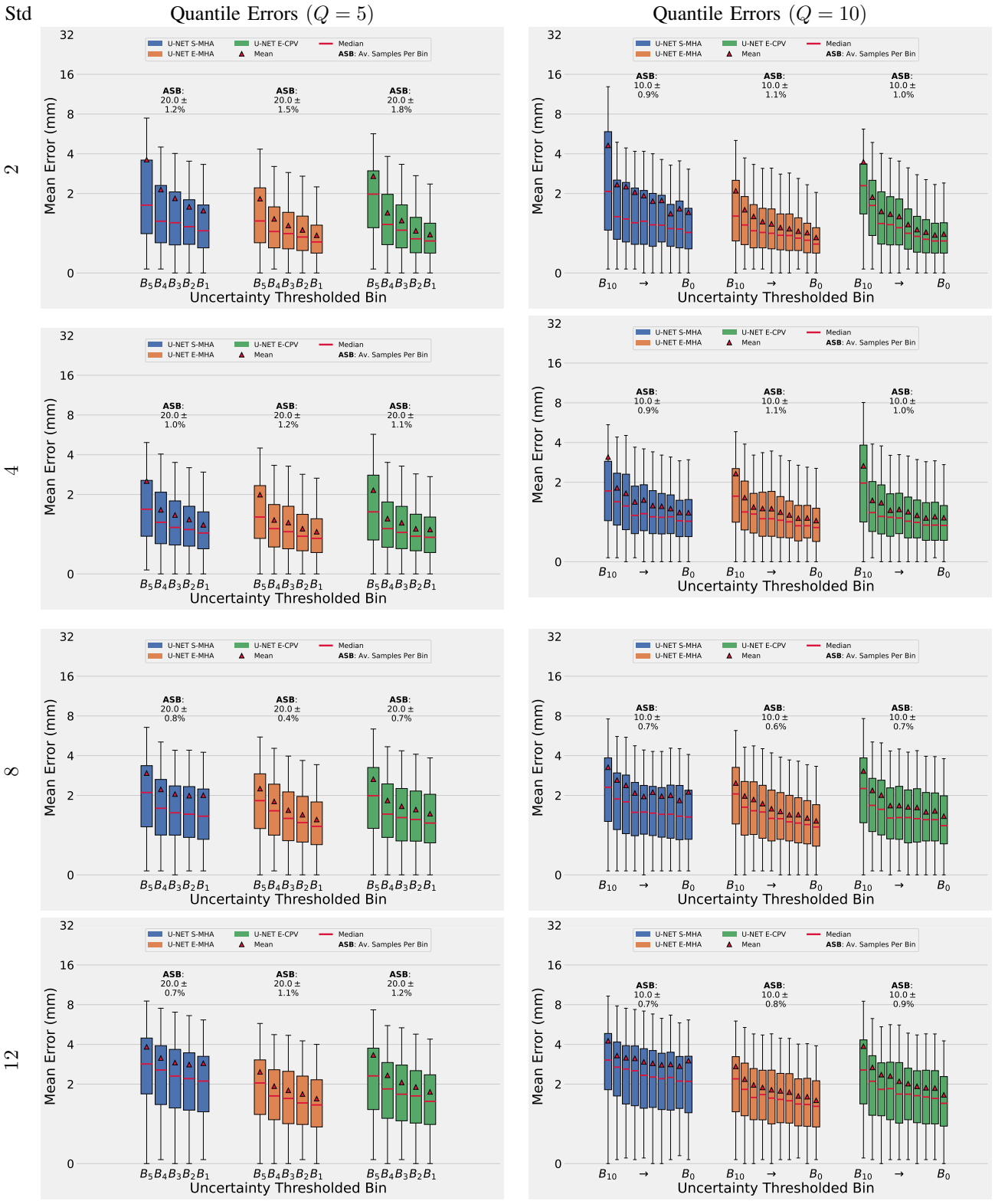


Fig. 2: Comparing results for models using different standard deviation values for the ground truth heatmap labels. We show the Quantile Localization Errors using 5 & 10 Quantile bins. We present results on all landmarks from a 4-fold CV on the Cephalometric dataset [1].

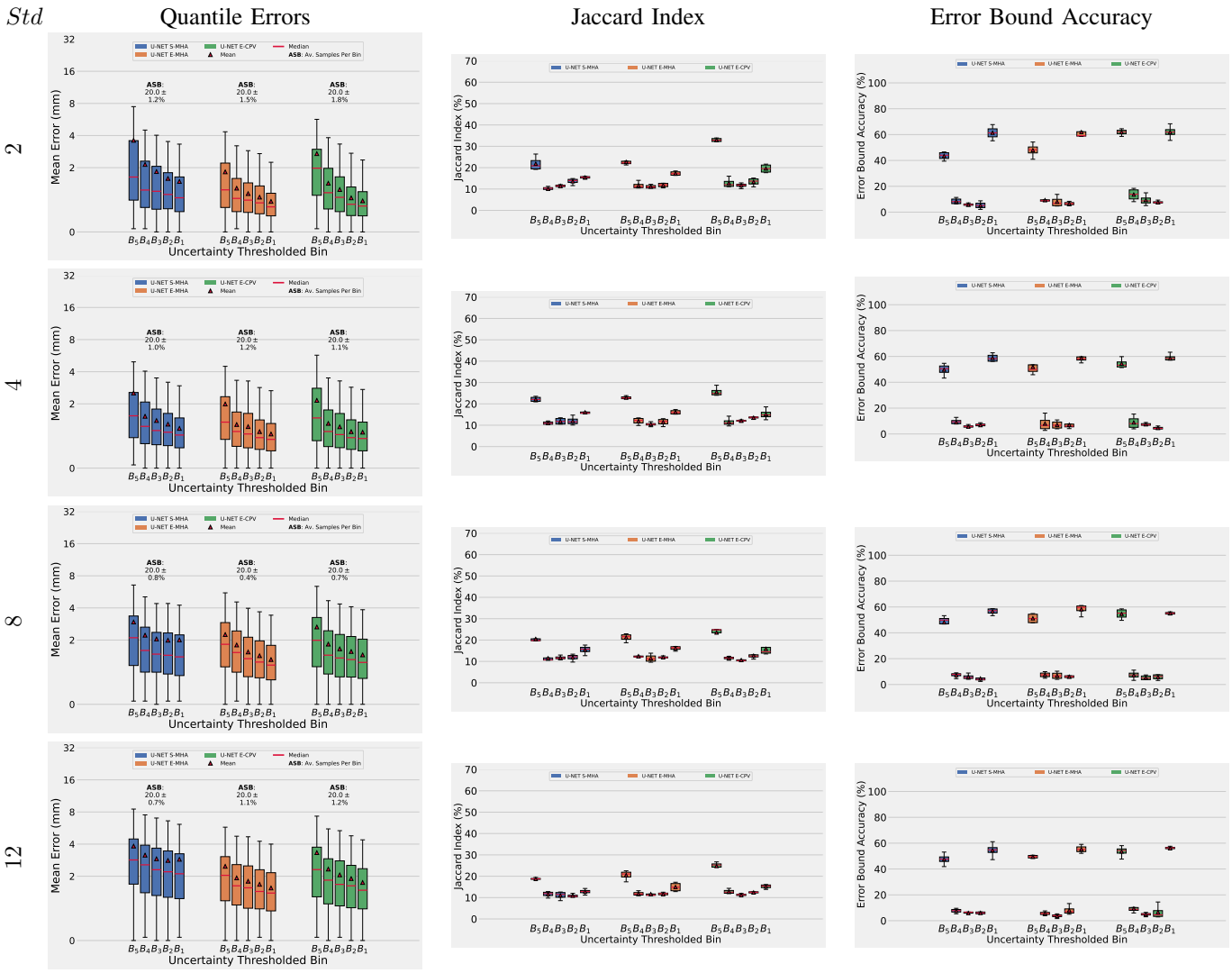


Fig. 3: Quantile Binning from a U-Net localization model trained using target heatmaps, varying the standard deviation of the Gaussian blob. The Quantile Errors, Jaccard index and Error Bound Accuracy are presented over a 4-fold Cross Validation on the Cephalometric dataset [1].

C. Merging Middle Bins

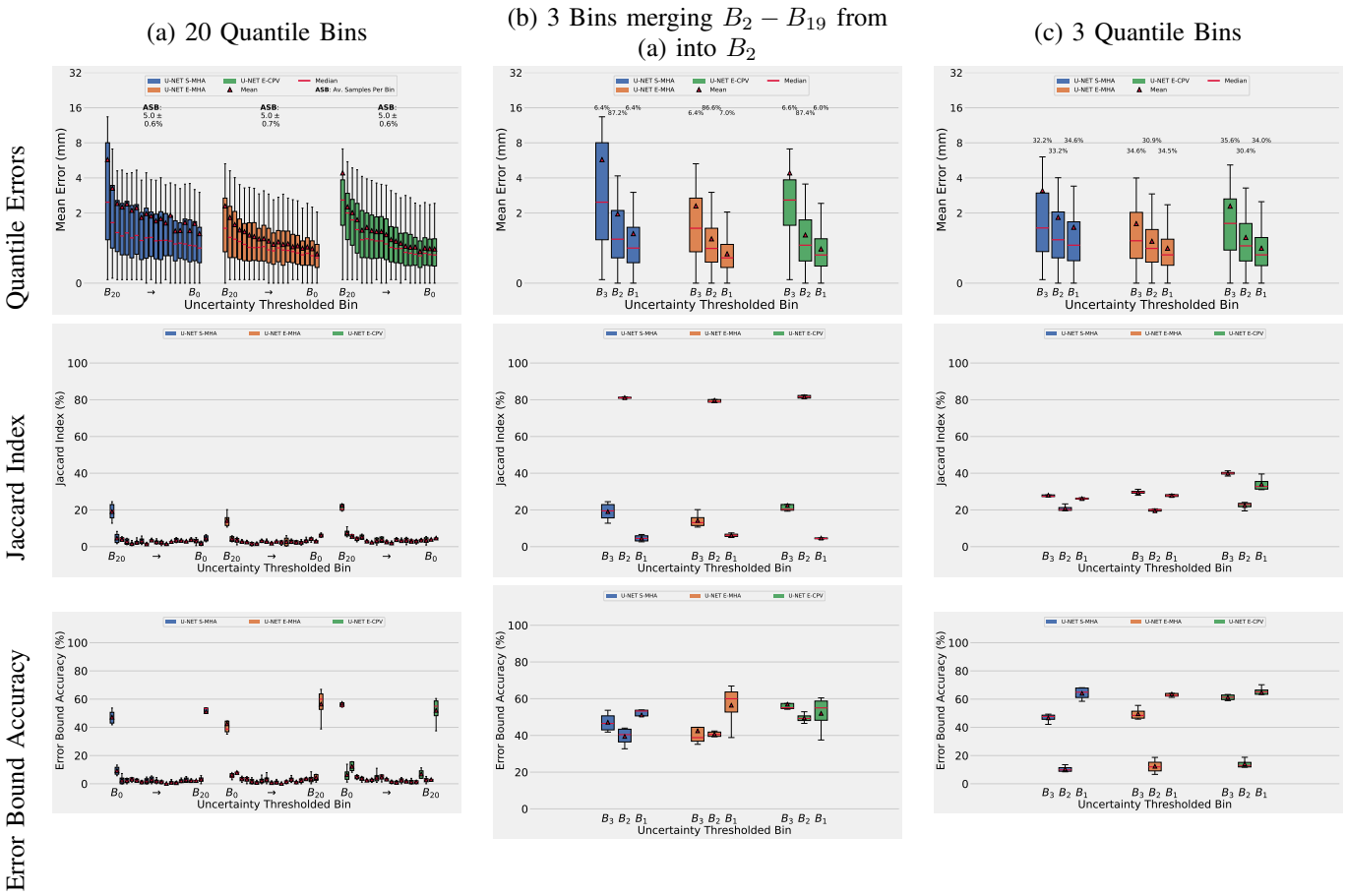


Fig. 4: Comparing using **(a)** 20 quantile bins, **(b)** 3 Bins where the edge bins are the same as (a) and the middle bin is a super bin from merging $B_{19} - B_2$, and **(c)** 3 Quantile Bins. We show the distribution of localization errors in each bin, the Jaccard index of each bin compared to the ground truth error quantiles the estimated error bound accuracies.

D. Uncertainty-Error Correlation

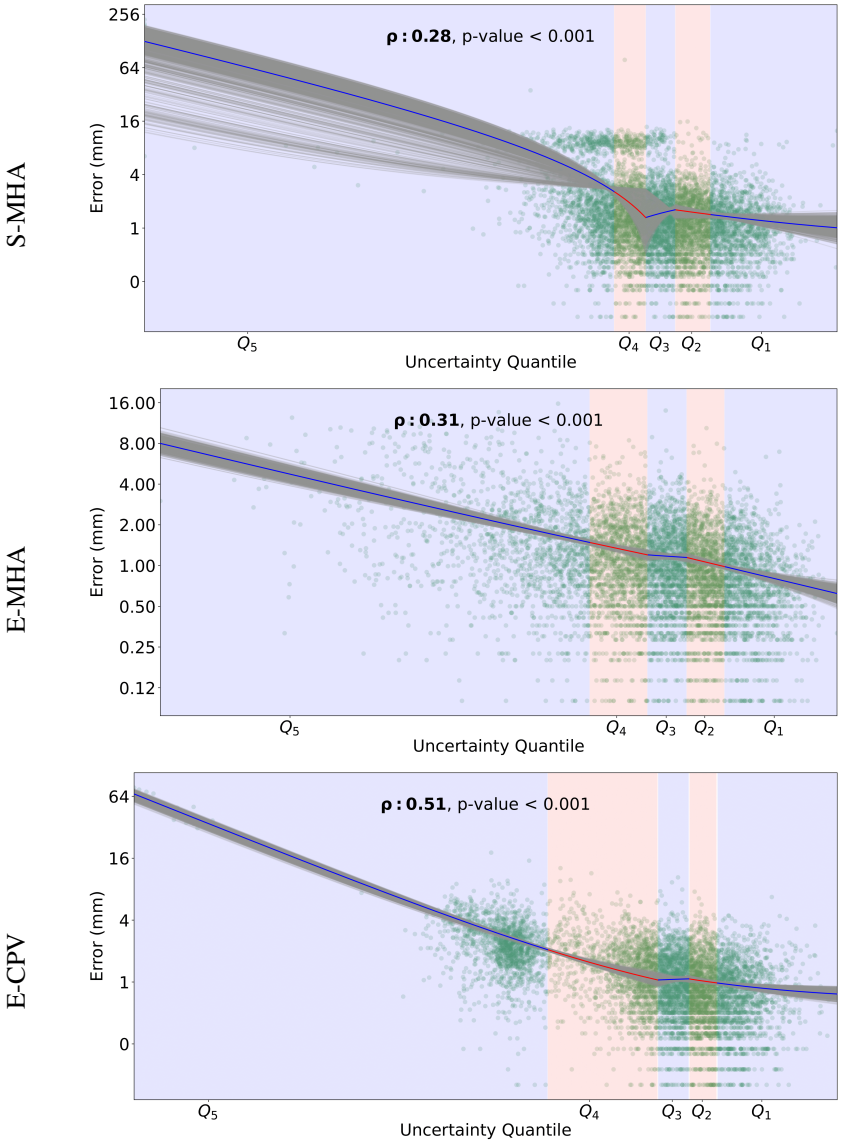


Fig. 5: Piece-wise linear regression of uncertainty with localization error, with breakpoints at the uncertainty quantiles. Grey represents bootstrap confidence intervals. Data is reported on all data from 4-fold cross validation on the Cephalometric dataset [1] using the U-Net model. ρ is the Spearman's Rank Correlation Coefficient between the uncertainty measure and error. Both the x-axis and y-axis are log-transformed.

E. Comparing Q Values

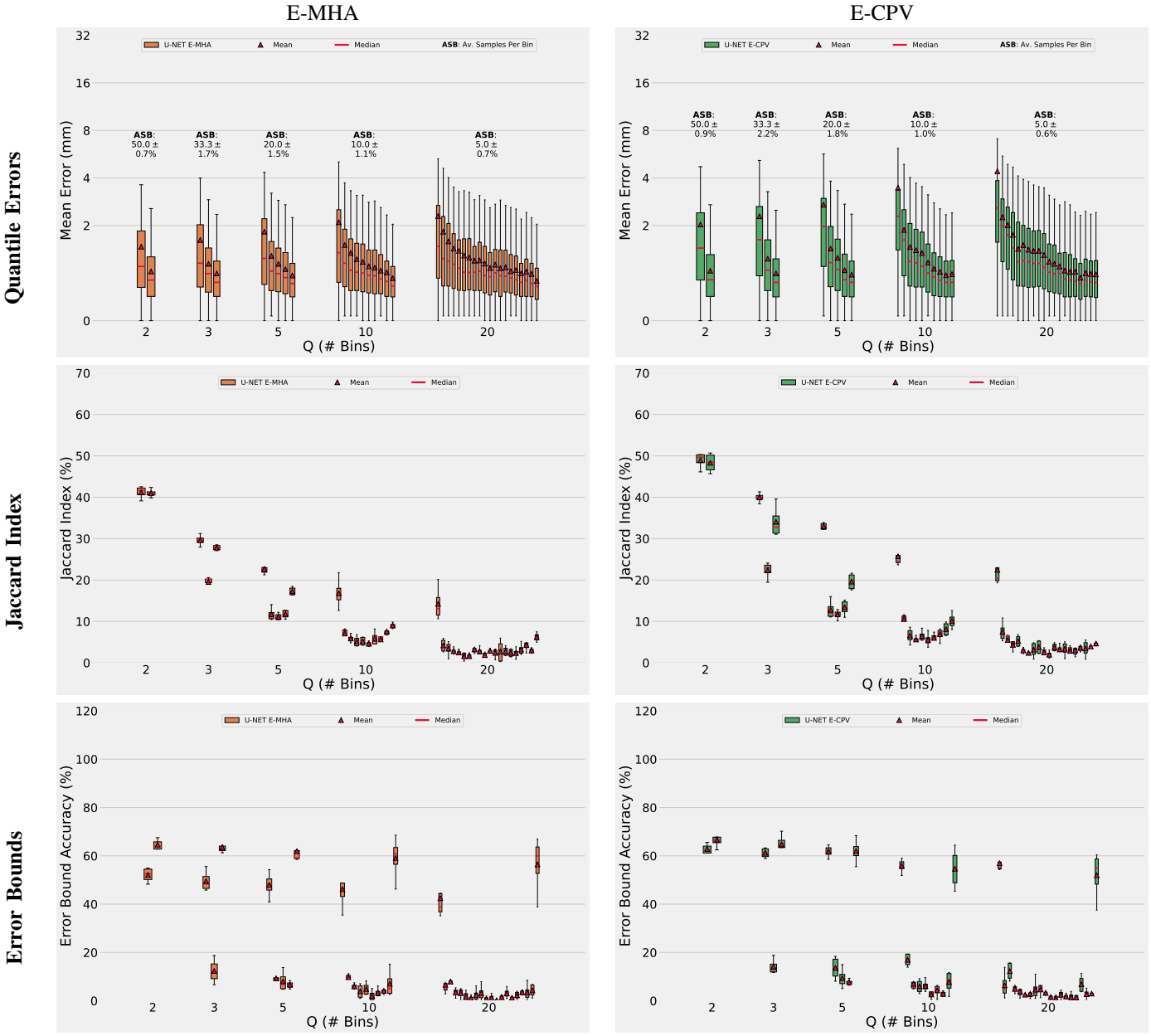


Fig. 6: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures E-MHA and E-CPV, over all landmarks from a 4-fold CV on the **Cephalometric dataset [1], trained on the U-Net model.**

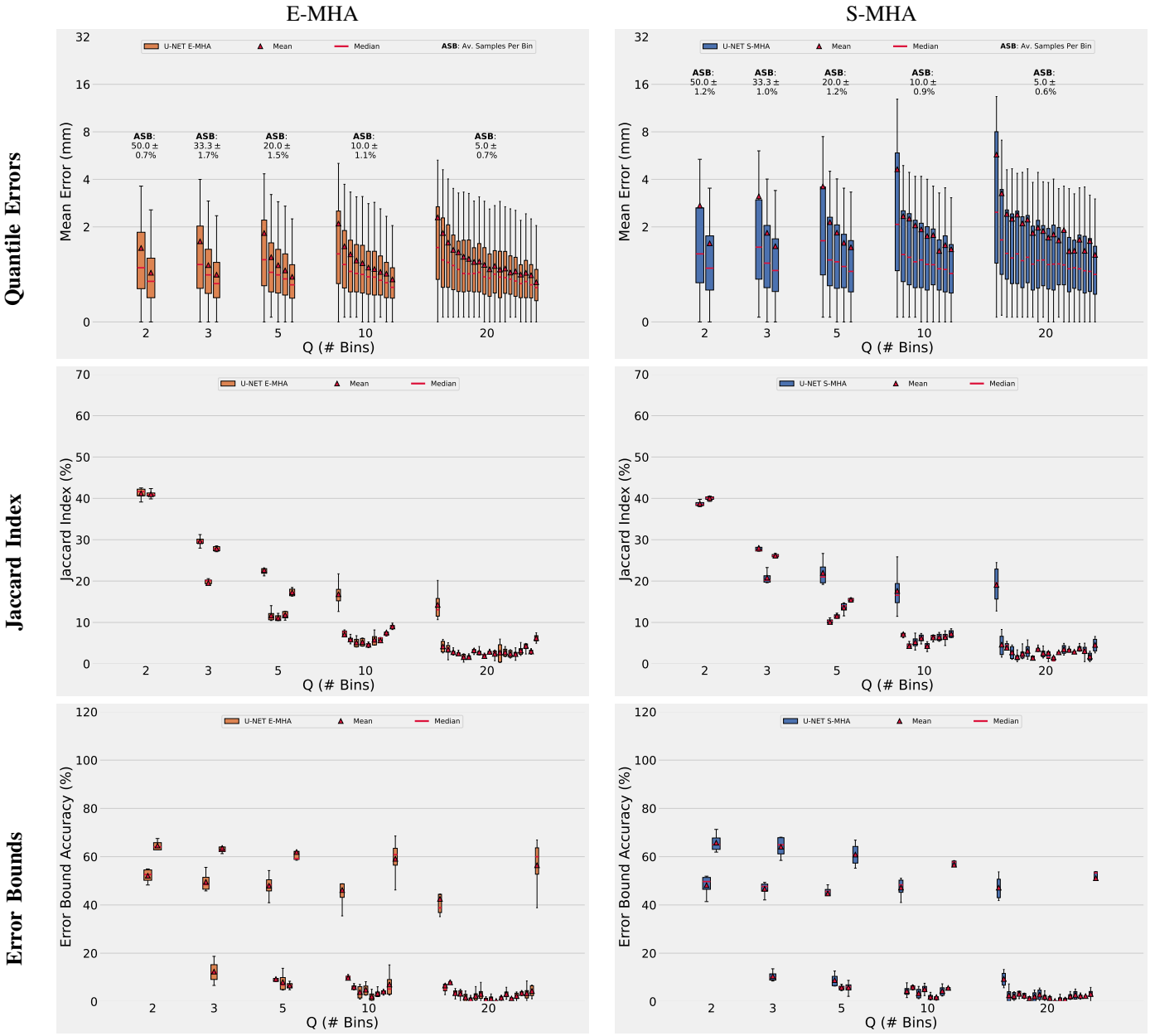


Fig. 7: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures E-MHA and S-MHA, over all landmarks from a 4-fold CV on the **Cephalometric dataset [1], trained on the U-Net model.**

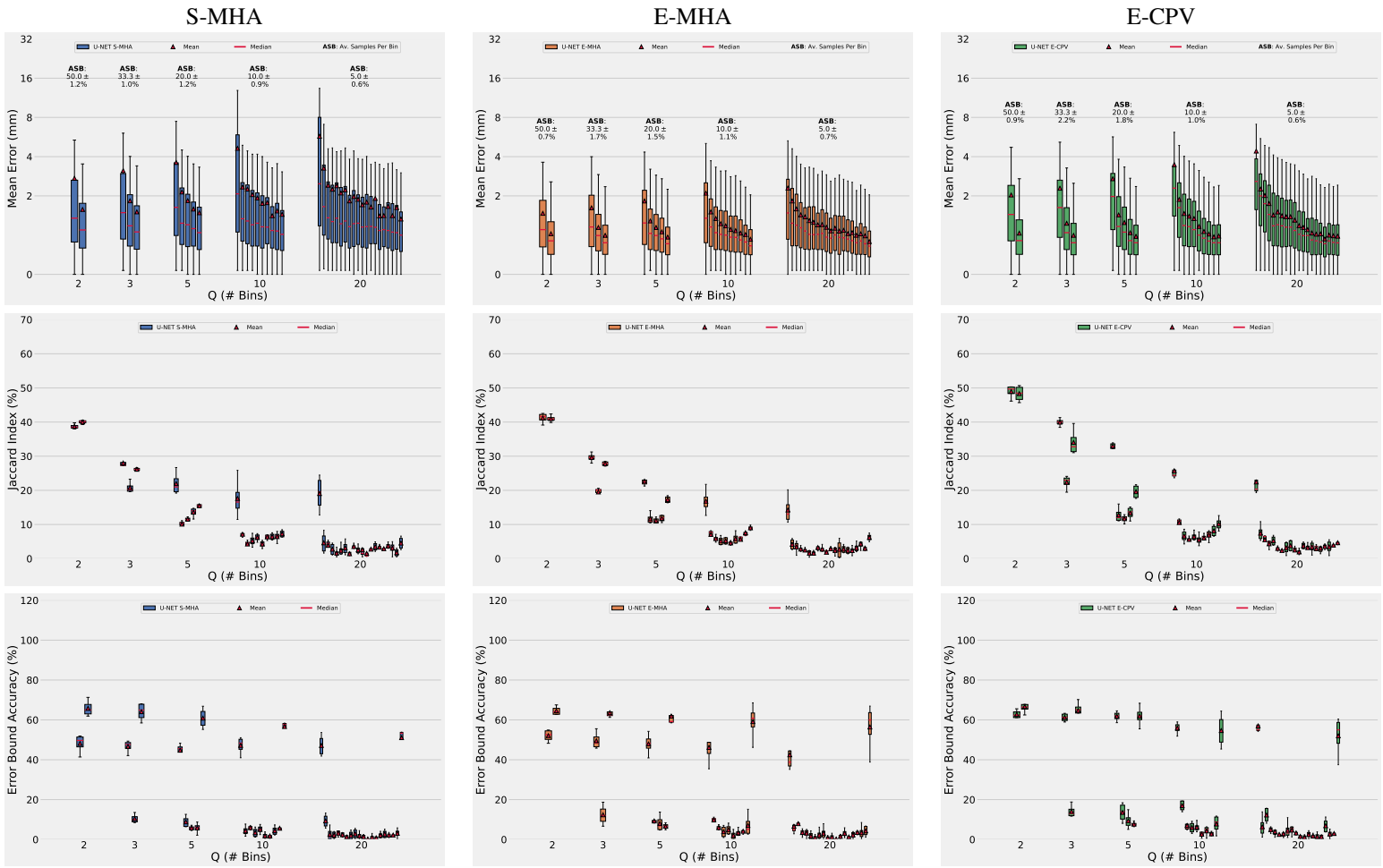


Fig. 8: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures S-MHA, E-MHA and E-CPV, over all landmarks from a 4-fold CV on the **Cephalometric dataset [1], trained on the U-Net model.**

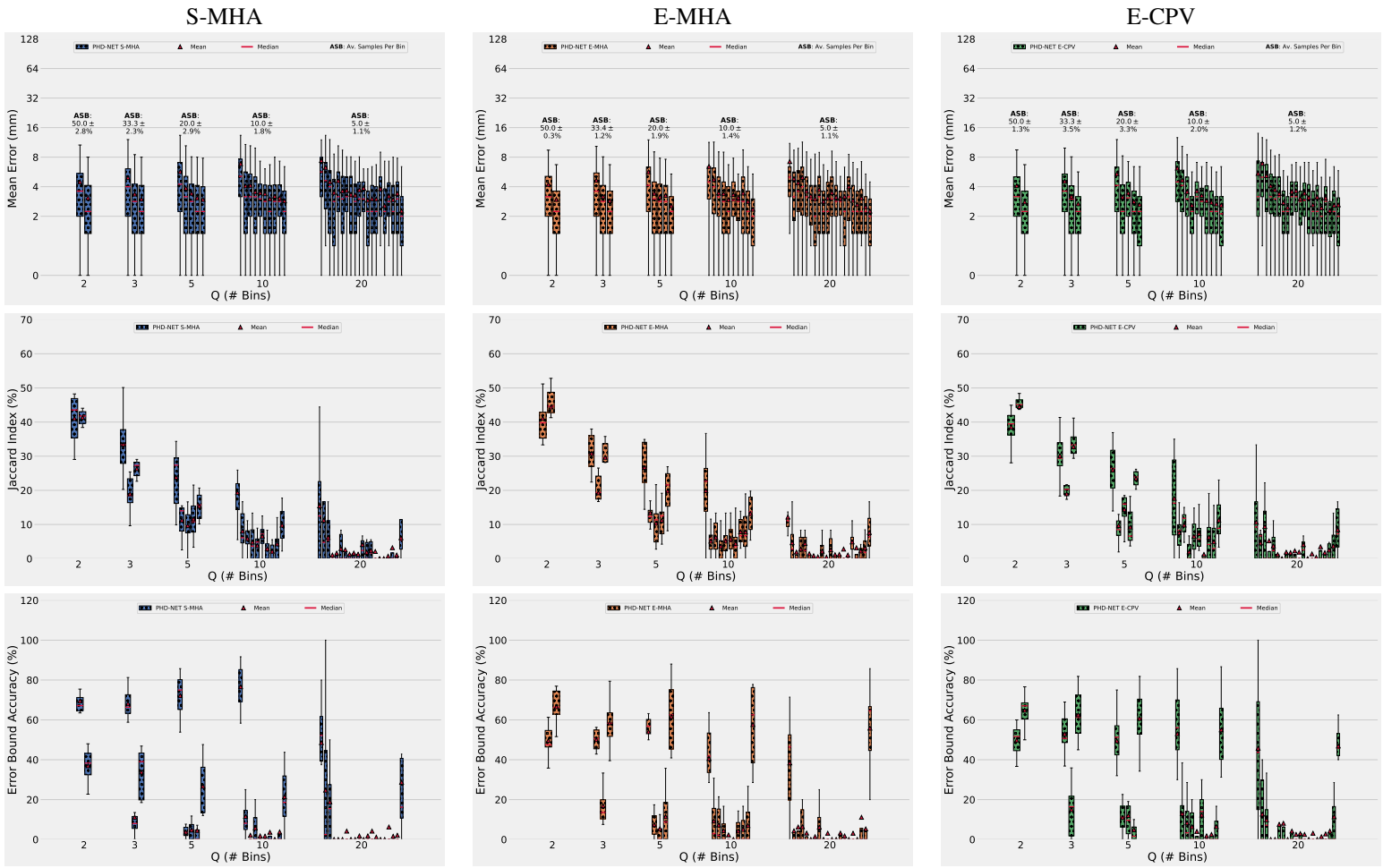


Fig. 9: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures S-MHA, E-MHA and E-CPV, over all landmarks from a 8-fold CV on the **SA dataset, trained on the PHD-Net model.**

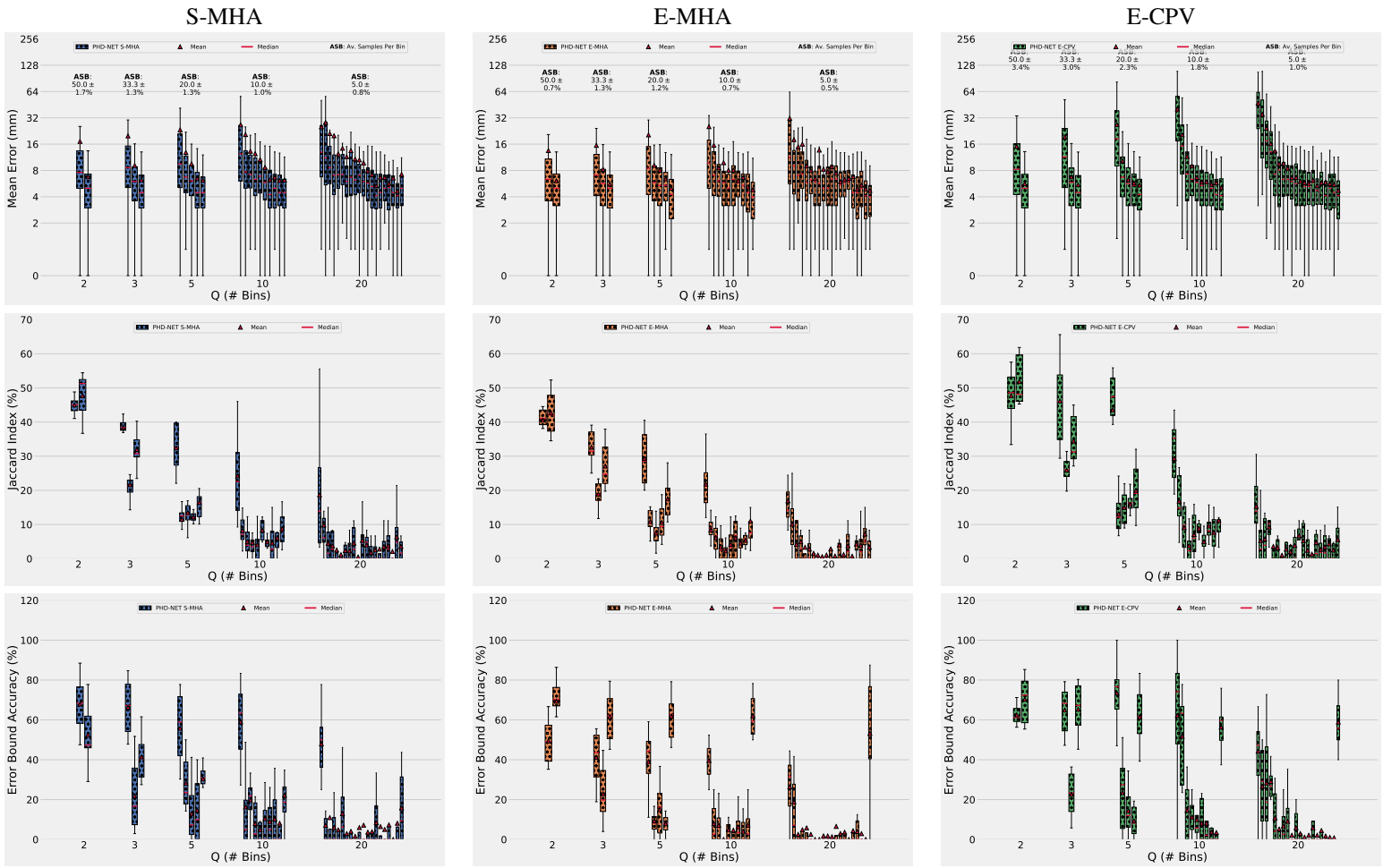


Fig. 10: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures S-MHA, E-MHA and E-CPV, over all landmarks from a 8-fold CV on the **4CH dataset, trained on the PHD-Net model.**

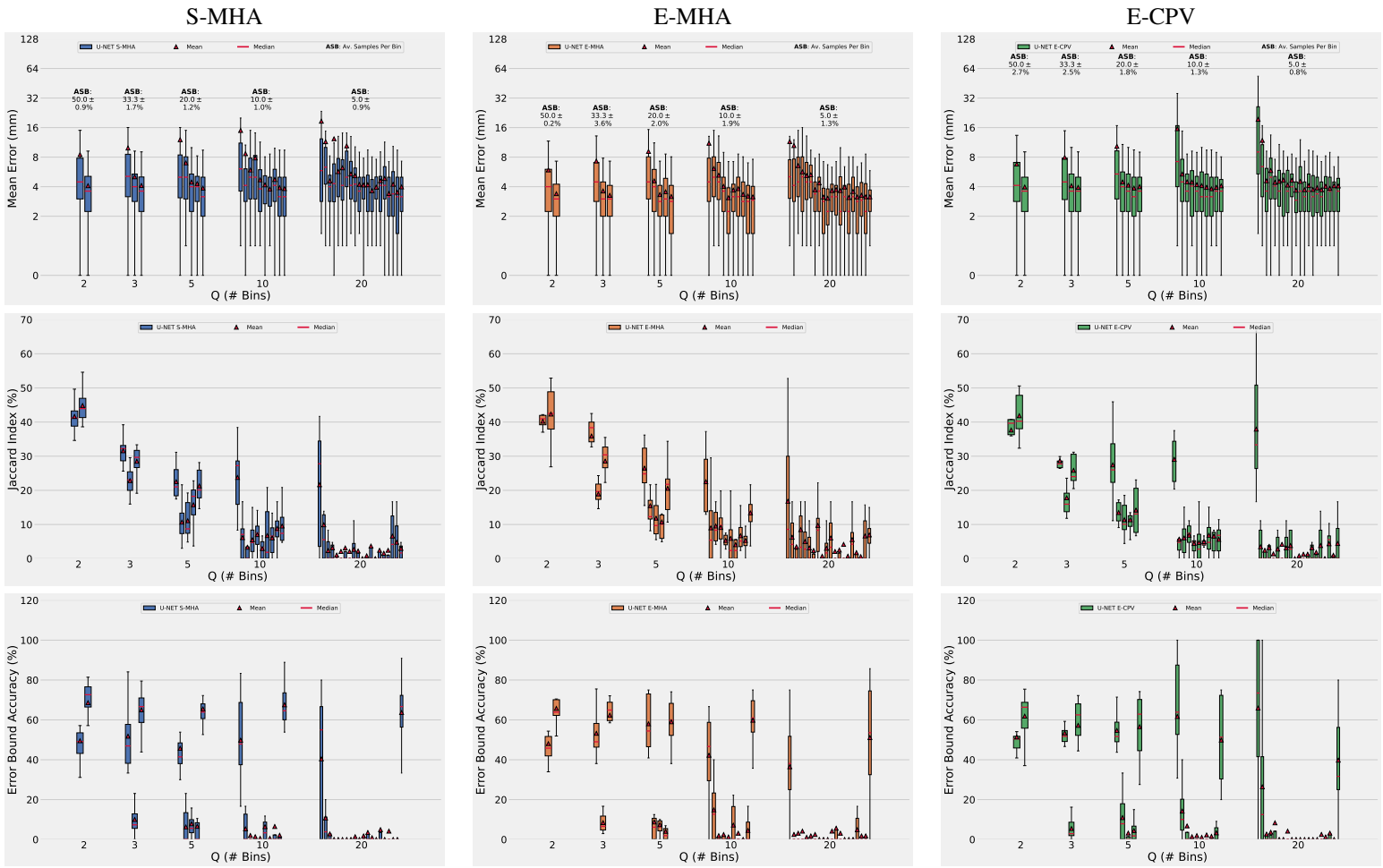


Fig. 11: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures S-MHA, E-MHA and E-CPV, over all landmarks from a 8-fold CV on the **SA dataset, trained on the U-Net model.**

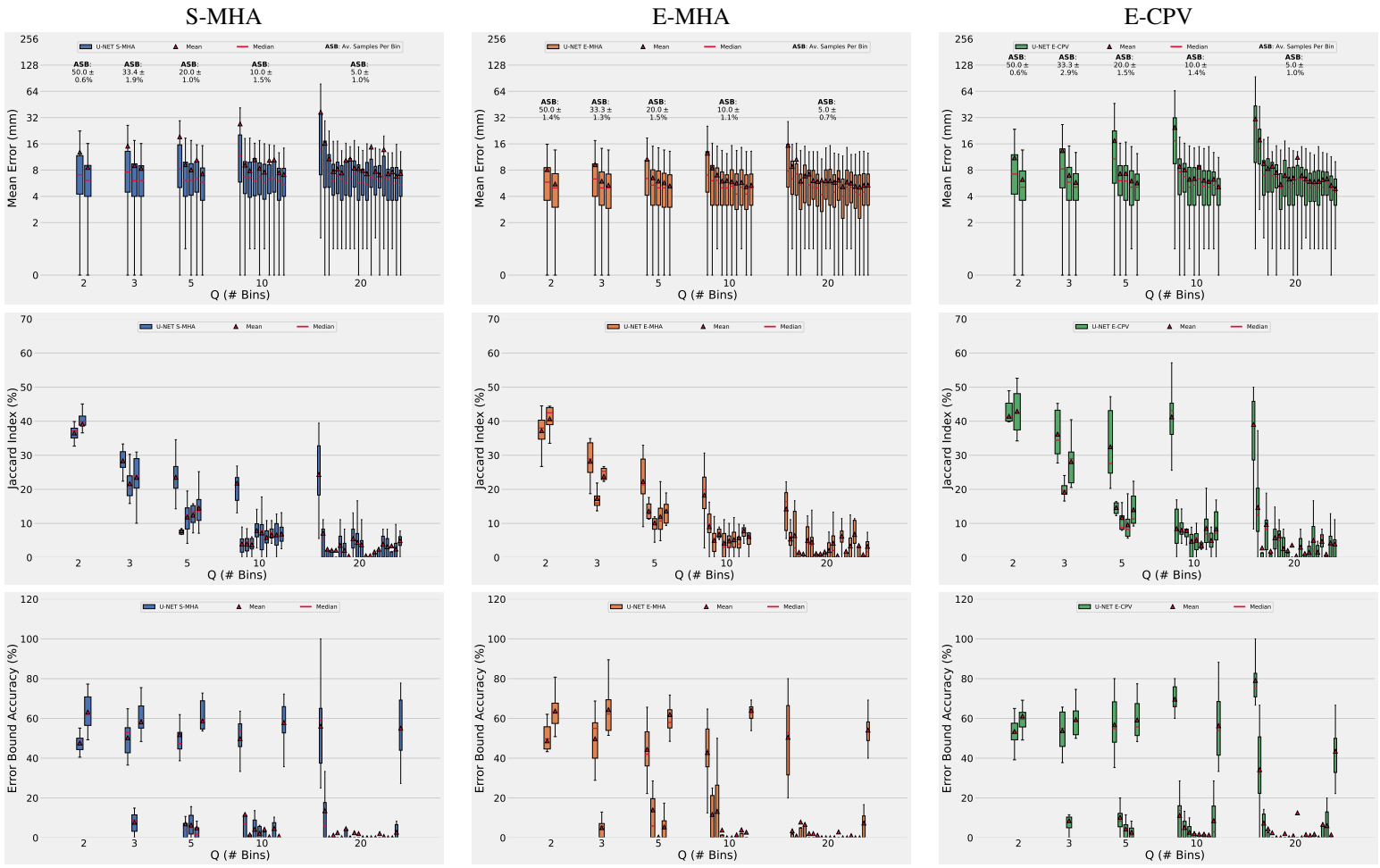


Fig. 12: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures S-MHA, E-MHA and E-CPV, over all landmarks from a 8-fold CV on the **4CH dataset, trained on the U-Net model.**

REFERENCES

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

[2] F. Thaler, C. Payer, M. Urschler, D. Štern *et al.*, "Modeling annotation uncertainty with gaussian heatmaps in landmark localization," *Machine Learning for Biomedical Imaging*, vol. 1, no. UNSURE2020 special issue, pp. 1–10, 2021.