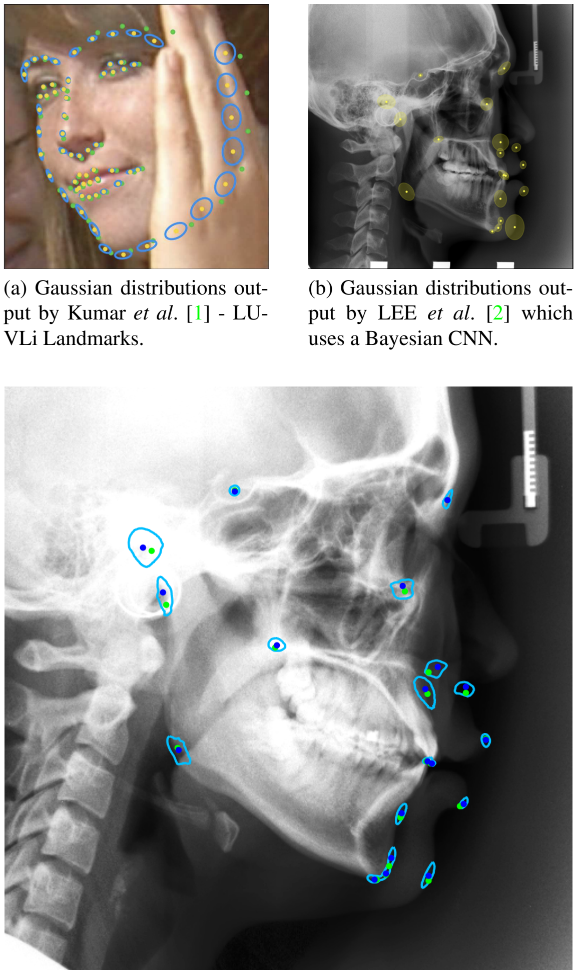
Contour-Hugging Heatmaps for Landmark Detection

James McCouat Irina Voiculescu
Department of Computer Science, University of Oxford
name.surname@cs.ox.ac.uk

Abstract

We propose an effective and easy-to-implement method for simultaneously performing landmark detection in images and obtaining an ingenious uncertainty measurement for each landmark. Uncertainty measurements for landmarks are particularly useful in medical imaging applications: rather than giving an erroneous reading, a landmark detection system is more useful when it flags its level of confidence in its prediction. When an automated system is unsure of its predictions, the accuracy of the results can be further improved manually by a human. In the medical domain, being able to review an automated system's level of certainty significantly improves a clinician's trust in it. This paper obtains landmark predictions with uncertainty measurements using a three stage method: 1) We train our network on one-hot heatmap images, 2) We calibrate the uncertainty of the network using temperature scaling, 3) We calculate a novel statistic called 'Expected Radial Error' to obtain uncertainty measurements. We find that this method not only achieves localization results on par with other state-of-the-art methods but also an uncertainty score which correlates with the true error for each landmark thereby bringing an overall step change in what a generic computer vision method for landmark detection should be capable of. In addition we show that our uncertainty measurement can be used to classify, with good accuracy, what landmark predictions are likely to be inaccurate. Code available at: `https://github.com/jfm15/ContourHuggingHeatmaps.git`



(a) Gaussian distributions output by Kumar *et al.* [1] - LUVLi Landmarks.

(b) Gaussian distributions output by LEE *et al.* [2] which uses a Bayesian CNN.

(c) Contours of the heatmaps output by our method. The dark blue dots are the predicted landmark points and the bright green dots are the ground truth. We call our heatmaps contour hugging because of the way they bend around the edges (in this case of the head). Our probability distributions are not restricted to being symmetrical and uni-modal.

Figure 1. Images demonstrating the difference in how our method quantifies the uncertainty of its landmark positions compared to previous approaches.

1. Introduction

Automatic landmark detection from images is an important task in a number of applications from monitoring a driver's vital signs [3] to medical imaging applications on numerous body parts including the knee, spine and lungs [4–6]. Most modern approaches to landmark detection use a deep learning pipeline and obtain impressive localization results. However these deep learning methods always detect some landmarks erroneously during testing. Take, for example, the task of cephalometric landmark detection from x-rays of the head. These landmarks are used to compute clinically useful angles and measurements from which clinicians can diagnose patients [7]. However, even the latest deep learning approaches detect at least 13% of landmarks outside the clinically accepted range (greater than $2mm$ error) [8, 9] so it could be dangerous to build these systems into safety-critical clinical workflows, especially if there was no human expert supervision. In this paper we address this problem by formulating the task of landmark detection as a classification task over all pixels in an image. This allows us to obtain more expressive and interpretable heatmaps as shown in Figure 1c. These heatmaps can be calibrated (Section 3.3) and then analysed using our novel statistic called Expected Radial Error ERE (Section 3.4). This statistic correlates well with the true localization error and can be used to flag potentially erroneous predictions (Section 5.3.1).

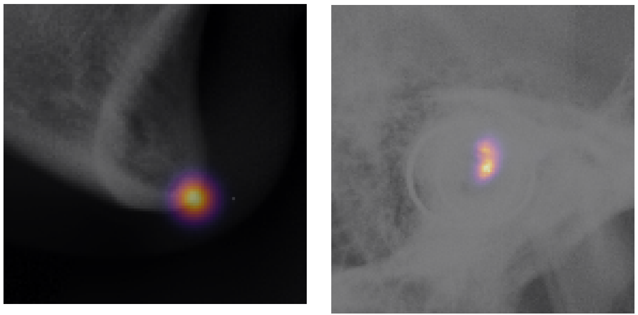
2. Background

In recent literature, fully convolutional neural networks (CNNs) have established themselves as the state of the art in landmark detection overtaking previous approaches such as random forests [10, 11]. This began with Tompson *et al*. [12] who used a CNN to regress target heat maps achieving state of the art performance on the human pose detection problem. Shortly afterwards, fully convolutional networks [13], including the U-Net [14], became very popular for segmentation tasks and its encoder-decoder architecture began to be applied to landmark detection as well, such as in Payer *et al*. [15]. More recent architectures in the area have stacked or cascaded models like this sequentially [16, 17] or in more complicated configurations [18]. However there is still evidence that the standard U-Net can perform at a high level when its hyper-parameters are tuned correctly in both segmentation [19] and landmark detection problems [20].

The works mentioned so far produce landmark predictions but do not produce any value of how ‘sure’ or how ‘uncertain’ their model is in that prediction. One reason for this is that the majority of existing approaches train networks on synthetically generated heatmaps created by a Gaussian distribution [9, 15, 21]. This has the disadvantage that the network is being trained on heatmaps which do not represent the uncertainty of where that landmark could realistically be placed, and so, the output of those models is not calibrated either [1]. This is illustrated in Fig 2a.

2.1. Uncertainty Estimation Methods

Recent works which aim to address this problem include Lee *et al*. [2], which uses a Bayesian CNN to output 2D Gaussian probability distributions for each landmark representing the probabilities of where that landmark could be placed and Kumar *et al*. [1], which regresses the position of the landmark as well as the values of a covariance matrix representing the uncertainty of its position. The problem with these approaches is that they restrict their output probability distribution to a Gaussian distribution, which is unrealistic for many real world tasks because it is uni-modal and symmetric.



(a) This heatmap is a Gaussian distribution like most recent methods. However it isn’t representative of the position of the landmark because this point at the end of a chin is unlikely to be placed within the chin itself or out in space. It is more likely to be placed along the contour of the chin by a human.

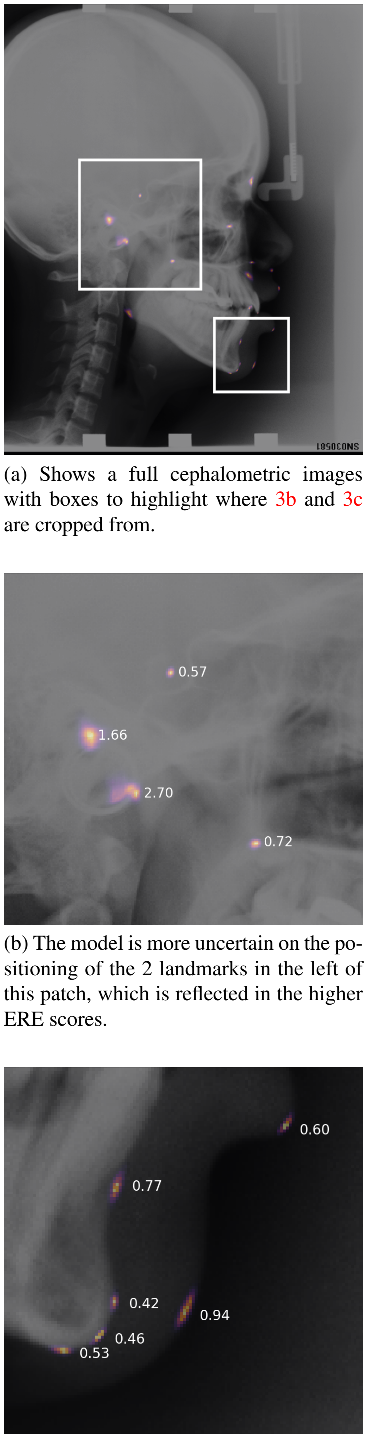
(b) Our network outputs a multi-modal distribution heatmap for the landmark in the center of this image. This nuance would not be captured by previous approaches which encode the uncertainty as Gaussian distributions.

Figure 2. (a) Shows a commonly used method for generating target heatmaps. (b) Shows how our network trained on one-hot heatmaps can express a multi-modal distribution in its output.

2.2. Our Work

We address this disadvantage by formulating the problem of landmark detection as a classification problem over all pixels in the image to obtain output heatmaps which are not restricted. We theorize that we can make these output heatmaps well calibrated using a temperature scaling method described in Guo *et al*. [22] and thus provide accurate heatmaps. We validate these heatmaps by assessing their calibration using reliability diagrams and measuring the Expected Calibration Error (ECE) in Section 5.2.

In addition, we propose a statistic called Expected Radial Error (ERE) to summarize how uncertain our model is based on the heatmap output. We firstly validate this statistic by showing there is a correlation between it and localization error. Then, secondly, we perform experiments to see whether we can filter landmarks, on an individual basis, using this statistic by applying a threshold to it to flag up when our model is likely to have made an inaccurate prediction (Section 5.3.1). This is relevant in real world applications: an AI system which can flag up when an error in its output is likely to occur is most valuable. Such functionality also leads to increasing users’ trust in the system.



(a) Shows a full cephalometric images with boxes to highlight where 3b and 3c are cropped from.

(b) The model is more uncertain on the positioning of the 2 landmarks in the left of this patch, which is reflected in the higher ERE scores.

(c) The model is reasonably certain on the positioning of these landmarks although the spread out heatmaps have slightly higher ERE scores.

Figure 3. Output heatmaps displayed with their Expected Radial Error (ERE) statistics.

2.3. Contributions

Our innovations are to:

1. Present a reproducible network which achieves performance comparable to the state-of-the-art on the cephalometric landmark detection task, and can run on a modest 8GB GPU. We share the code at: https://github.com/jfm15/ContourHuggingHeatmaps.git.
2. Show that it is possible to obtain near SOTA localization performance by formulating the landmark detection task as a classification task (Table 1).
3. Demonstrate that the probabilities in the output heatmaps can be calibrated using temperature scaling (Section 5.2.1).
4. Show that our novel Expected Radial Error (ERE) statistic correlates with the localization error (Section 5.3) and build a binary classifier based on ERE to flag up potential erroneous predictions with a good degree of accuracy (Section 5.3.1).

3. Method

The principal novelty in this work comes from how we have formulated the problem of landmark detection as a classification problem and how we have validated the utility of the output heatmaps in a qualitative and quantitative way. This work uses the well-established U-Net as the main network architecture, as described in the subsequent section.

3.1. Architecture

We perform all experiments using a U-Net [14] with a ResNet-34 encoder pretrained on ImageNet [23]. The U-Net architecture is chosen because it is easy to implement, reproducible and has evidence of obtaining good results on landmark detection problems [20]. Our decoder has 5 levels of upsampling with 256, 128, 64, 32 and 32 channels in each of the levels from the bottom level to the top.[1] After each convolution there is a batch normalisation layer and a ReLU activation function. We then have a final 1×1 convolutional layer to squash the 32 channels in the top layer into N channels each of which represents the heatmap for one of the landmarks, N being the number of landmarks to be detected. This is implemented using the pytorch segmentation models library.[2]

We then apply a 2D softmax activation function to each of these channels[3] to convert them into a probability distribution over all pixels in the image. Formally our network

[1] We ran our experiments on a 8GB GPU and were quite restricted in how many channel each layer could have.

[2] https://github.com/qubvel/segmentation_models.pytorch

[3] Unless we are performing temperature scaling, see Section 3.3, in which case we scale each channel by a temperature parameter first.

outputs a tensor comprised of n channels: $\{c_1, c_2, ..., c_n\}$, $c_l \in \mathbb{R}^{w \times h}$ where w and h are the width and height of the input images. The 2D softmax function works on each channel independently such that:

$$\sigma_l(i,j) = \frac{e^{c_l(i,j)}}{\sum_{s=1}^{w} \sum_{t=1}^{h} e^{c_l(s,t)}} \tag{1}$$

where tensor $\sigma_l(\cdot,\cdot)$ is calculated for each channel, $l{\in}[1..n]$. We use a negative log likelihood loss[4] to train the network. At test time we obtain predicted landmark points by selecting the hottest point in the heatmap or, in other words, the mode of the output distribution.

3.2. Heatmap Generation

As mentioned in Section 2.2 we formulate the landmark detection task as a classification problem. We do this by training our model on heatmaps which contain a single 1 spike at the ground-truth point, with 0s at every other position. Formally this is defined as:

$$H(i,j) = \begin{cases} 1, & \text{if } i = x \text{ and } j = y \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Where $H(i,j)$ denotes the pixel value of the heatmap at each point (i,j), and (x,y) denotes the coordinates of the ground truth landmark point.

3.3. Temperature Scaling

An advantage of formulating landmark detection as a classification problem is that we can perform temperature scaling to calibrate the heatmap probabilities. Temperature scaling is well described in Guo *et al.* [22]. After we train our model using the heatmaps described in Section 3.2 we add an additional parameter T for each landmark to the model such that each channel pixel $c_l(\cdot,\cdot)$ is divided by scalar T_l before being put through the softmax activation function. So the softmax output becomes:

$$\sigma_l(i,j) = \frac{e^{c_l(i,j)/T_l}}{\sum_{s=1}^{w} \sum_{t=1}^{h} e^{c_l(s,t)/T_l}} \tag{3}$$

We freeze all the other parameters in our trained network and fine tune our network to optimize the T_l parameters using the same the negative log likelihood loss. This does not change the localization accuracy of the network because the hottest points will remain the same no matter what values T_l takes. Section 5.2.1 shows how T_l calibrates the network.

3.4. Uncertainty Estimation

We obtain an uncertainty measurement for each landmark by calculating a statistic we call the Expected Radial Error (ERE). This statistic is calculated using the following equation:

$$ERE_{\sigma_l} = \sum_{i=1}^{w} \sum_{j=1}^{h} \sigma_l(i,j) \sqrt{(i-\tilde{x})^2 + (j-\tilde{y})^2} \tag{4}$$

Where w and h are the width and height of the image; $\sigma_l(\cdot,\cdot)$ is the probability distribution output by Eq 3 for each heatmap $l{\in}[1..n]$; and $(\tilde{x},\tilde{y})$ are the coordinates of the predicted landmark, or in other words, the hottest point in σ_l.

It is worth noting that σ_l should be pre-processed before ERE is calculated. This pre-processing consists of converting all values in σ_l which are below 5% of the hottest point to 0 and then re-normalizing σ_l by dividing by the sum of its values. This is done because pixels far away from the landmark have tiny (10^{-6}) heatmap values, thereby adding noise to the ERE calculation. Zeroing values less than 5% (chosen empirically) of the maximum helps increase the correlation between the ERE statistic and the true radial error in our experiments from 0.9 to 0.96 (see Figure 5).

In Section 5 we hypothesize that a high ERE score can be used to flag up erroneous predictions, which we validate in Section 5.3.1. Examples of ERE scores next to heatmaps are given in Figure 3.

4. Experiments

4.1. Dataset

We perform our experiments on a publicly available cephalometric dataset, originally released for a grand challenge at the IEEE ISBI conference in 2015 [24]. The dataset contains 400 x-rays, split into 150 training images and two test sets, Test Set 1 and Test Set 2, comprised of 150 and 100 images respectively. Each image in the dataset has a resolution of 1935×2400 where each pixel represents a $0.1mm$ square and each comes with two sets of ground truth annotations for 19 landmarks, one from a senior clinician and one from a junior clinician. These experts placed ground truth landmarks manually on each image according to strict medical definitions. As in previous works [24] we take the average of the landmark points given by the two clinicians as our ground truth landmark points which we train and test on. Before passing these images into our network for training or testing we resize them to 640×800 pixels.

4.2. Training The Network

We train our U-Net architecture by passing downsampled cephalometric x-ray images of size 640×800 through our network and 2D Softmax function (Eq 1) to

[4] https://pytorch.org/docs/stable/generated/torch.nn.NLLLoss.html

		Test Set 1					Test Set 2				
	Model	MRE	SDR (%)				MRE	SDR (%)			
		(mm)	2mm	2.5mm	3mm	4mm	(mm)	2mm	2.5mm	3mm	4mm
No uncertainties	Ibragimov *et al.* [10]	1.87	71.70	77.40	81.90	88.00	-	62.74	70.47	76.53	85.11
	Lindner *et al.* [11]	1.67	74.95	80.28	84.56	89.68	-	66.11	72.00	77.63	87.42
	Arik *et al.* [8]	-	75.37	80.91	84.32	88.25	-	67.68	74.16	79.11	84.63
	Yao *et al.* [20]	1.24	84.84	90.52	93.75	97.40	1.61	71.89	80.63	86.36	93.68
	Chen *et al.* [25]	1.17	86.67	92.67	95.54	98.53	1.48	75.05	82.84	88.53	95.05
	Zhong *et al.* [9]	1.12	86.91	91.82	94.88	97.90	1.42	76.00	82.90	88.74	94.32
	Ours (with uncertainty)	1.20	83.47	89.16	92.60	96.49	1.46	74.63	83.58	87.21	93.79

Table 1. Localization results for our method compared to existing methods (lower MRE is better and a higher SDR percentages are better). Our method reports results which improve on old methods and are on a par with recent SOTA methods whilst having the significant benefits of being a simpler architecture and outputting an uncertainty measurement. It is worth noting that Lee *et al.* [2] also localize landmarks on cephalometric data and produce uncertainty measurements; however those results do not belong in the table because their test set is a combination of Test Sets 1 and 2. When we performed the same experiment, our method obtained a MRE of 1.30 compared to their 1.54.

obtain predicted heatmaps. We then use a negative *log* likelihood function against our ground-truth heatmaps to generate a loss. We train the network using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 4. We step down the learning rate by a factor of 0.1 at epochs 4, 6 and 8.

A large amount of data augmentation is implemented using the `imgaug` library.[5] We implement the following augmentations: X and Y translation by a maximum of 10 pixels, intensity scaling of all pixels by a random factor between 1 and 0.5, scale up or down the image so that it is between 0.95 and 1.05 of its original scale, rotate the image either anti-clockwise or clockwise at most 3°, and finally elastically transform each image. As no validation set was given with the dataset we optimized our hyper-parameters by holding out 30 images from the training set to use as a validation set. Then, once we had found the best hyper-parameters, we trained on the whole of the training set for 15 epochs (15 epochs was chosen because localization results plateaued after epoch 15 on the validation set) to obtain our final models.

5. Evaluation

We evaluate our model not just on the accuracy of its predicted points but also on how well-calibrated its output heatmaps are and whether they can be used to flag up erroneous results. Given a particular computer vision application, these measurements can quantify its real-life utility.

5.1. Localization Results

Once an image is put through the trained network, scaled by the temperature parameters and passed through the 2D softmax layer, we select the hottest point on the heatmap as our final predicted heatmap position. To validate the accuracy of these positions we used statistics established since the cephalometric dataset was released [7]. These are the Mean Radial Error (MRE), which is the average Euclidean distance between the predicted landmark points and the ground-truth landmark points measured in mm, and the Success Detection Rate (SDR) for 4 thresholds: $2mm$, $2.5mm$, $3mm$ and $4mm$. The SDR is the percentage of points predicted with an MRE less than a given threshold. We compare our results both to methods proposed when the dataset was originally released [10,11] and to state of the art methods with novel deep learning architecures [9, 20, 25]. The results are presented in Table 1.

5.2. Validating Our Heatmaps

The qualitative analysis of our heatmaps (examples shown in Figure 3b and Figure 3c) is particularly appealing because the heatmaps hug the contours of the feature (the head, in this case); they offer realistic potential locations of where each landmark could be placed.

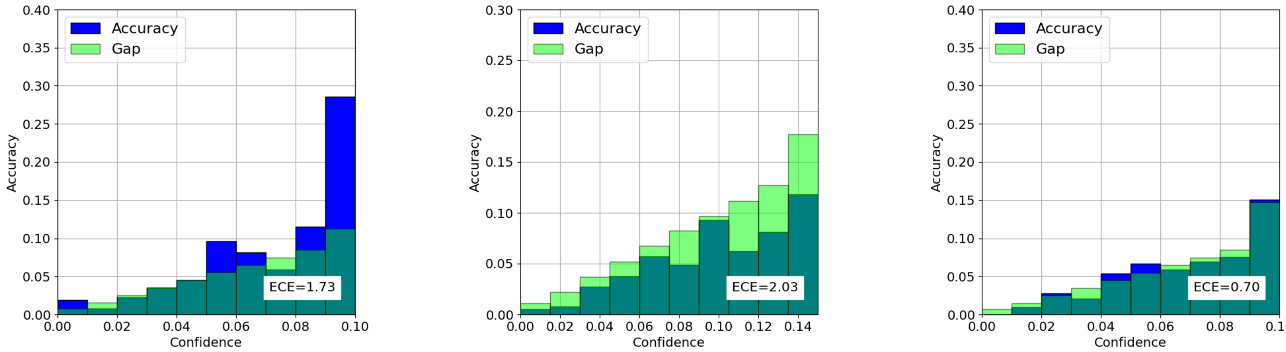
Additional to this geometric consideration, we also show numerically that these heatmaps are realistic. To do this we demonstrate that the probabilities in the heatmap are well calibrated by using a reliability diagram (Figure 4c).

5.2.1 Reliability Diagrams

To assess the calibration of our model after training we produce reliability diagrams like in Guo *et al.* [22]. The reliability diagrams are histograms which show accuracy as a function of confidence. In our case we define confidence to be the probability at the hottest point in the output heatmap, or, in other words, the value at our predicted landmark point.

To produce the histograms shown in Figure 4 we put each predicted landmark into a bin. There are 10 bins, each

[5] https://imgaug.readthedocs.io/en/latest/



(a) Reliability graph for a landmark detection model trained with Gaussian heatmaps, like in most existing approaches. We can see that the model is under-confident in the 0.09-0.1 bin.

(b) Reliability graph for the confidences of our model pre-temperature scaling. When compared to (c) we can see that temperature scaling improves the calibration of the model.

(c) The reliability graph for the confidences of our model after temperature scaling. The model's uncertainties are generally well calibrated.

Figure 4. (a) displays a poorly calibrated model trained on Gaussian heatmaps (σ=1). (b) displays another poorly calibrated model trained on one-hot heatmaps but with no temperature scaling. (c) is our model trained on one-hot heatmaps and temperature scaled (Sec. 3.2). Temperate scaling a model trained on Gaussian heatmaps does not make it more calibrated, unlike a model trained on one-hot heatmaps.

of equal width, such that the first bin will contain predicted points with confidences between 0 and $\frac{M}{10}$, the second will contain points with confidences between $\frac{M}{10}$ and $\frac{2M}{10}$ etc. Where M is the maximum confidence of any prediction in the validation set. The confidence for each bin is defined as the average confidence of predictions in that bin. We then define the accuracy of each bin as the percentage of correct predictions in that bin; in this case a prediction is correct when the hottest point on the output heatmap is at exactly the same coordinate as the ground-truth point (at the average position between our expert human annotations).

If the network is perfectly calibrated we would expect the confidence of each bin to be exactly the same as the accuracy for that bin. We display this information in reliability diagrams which overlays two histograms, one plotting the accuracy of each bin and the second plotting its confidence. These diagrams make it easy to see if the network is underestimating or overestimating the confidence in its predictions for each bin. If it is underestimating then the accuracy of a bin will be higher than its confidence so we will be able to see a solid blue piece at the top of the bar in that bin in the diagram. If it is overestimating we will see a solid lime green piece at that top of that bar for a bin the diagram.

From these diagrams we can calculate an Expected Calibration Error (ECE) which is the difference between the accuracy and confidence in each bin, weighted proportionally to how many landmarks are in each bin and then summed together. The lower the ECE the more calibrated the model. Figure 4c shows the reliability diagram for our model after temperature scaling and its ECE score of 0.7. For comparison we also show the reliability diagram for our model before temperature scaling in Figure 4b and a model trained on Gaussian heatmaps (like in Figure 2a) in Figure 4a. Both of which were also trained on the cephalometric dataset. We found that the model trained on Gaussian heatmaps was under confident in it's predictions, especially in the 0.09-0.1 bin as shown in Figure 4a.

5.3. Validating The ERE Statistic

We show that our 'Expected Radial Error' score correlates with the 'True Radial Error' in Figure 5 where the True Radial Error is the distance between the predicted landmark location and its ground truth location (average of the 2 clinicians placements) over Test Set 1. This graph is created by putting each predicted landmark into a bin depending on its ERE score. We chose each bin to contain 36 landmarks because this bin size was used in Kumar *et al*. [1]. The 36 landmarks with the lowest ERE scores are placed into one bin, then the next biggest 36 are put into another, etc. Once the landmarks are put into the bins we calculate the average ERE score (x axis value) and the average True Radial Error (y axis value) of all landmarks in the bin and plot that data point. We find there is a strong correlation between the ERE and True Radial Error which means ERE can be used as a good indicator for how accurate a prediction is likely to be.

5.3.1 Flagging Up Potential Bad Readings Using The ERE Score

The final part of this work is to show that the ERE score (Eq. 4) calculated from our heatmaps has a practical utility. We know that there is a high correlation between the ERE score and potential accuracy of the landmark position

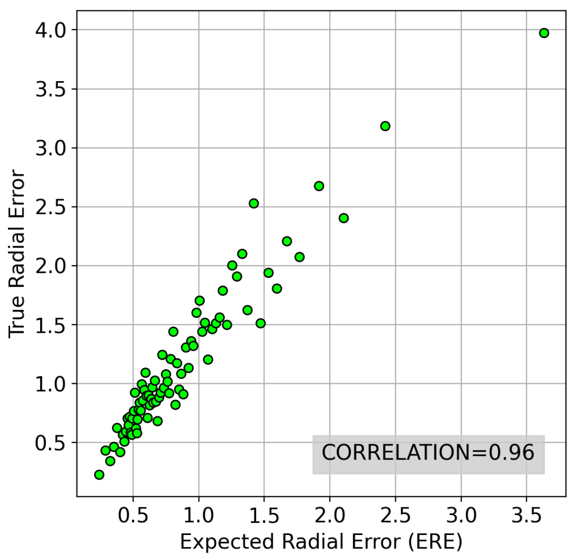


Figure 5. Correlation between the ERE statistic and the true radial error for bins of 36 landmarks over the cephalometric Test Set 1. The radial error is the Euclidean distance between the predicted and the ground truth point. The True Radial Error is slightly higher on average than the ERE score because of the pre-processing step discussed in Section 3.4.

(Figure 5) so we hypothesize that if we apply a threshold to the ERE score of a landmark when it is calculated we can filter out predictions which are likely to be inaccurate or 'erroneous'. This is valuable because the model can flag up when it is unsure of its prediction this way – a particularly desirable feature in medical applications where it is better to be cautious.

The experiment we conduct to test this hypothesis consists of the following steps:

1. Apply the model to all images in Test Set 1 to obtain heatmaps representing landmark locations.
2. Take the hottest points of these heatmaps to obtain predicted landmark points and calculate ERE scores for each landmark like in Figures 3b and 3c.
3. Compare the predicted landmark points to the ground truth points and classify each point as 'good' if its localization error is $< 2mm$, or 'erroneous' if its localization error is $> 2mm$. We chose $2mm$ as the threshold because this is the clinically accepted successful detection range [8].
4. Plot a Receiver Operating Characteristic (ROC) curve which describes how well different threshold values applied to the ERE scores can discriminate between 'good' or 'erroneous' predictions.

Our ROC curve is shown in Figure 6. A true positive is when the classifier (thresholding the value of the ERE for a landmark) predicted a localization error of over $2mm$ and the predicted point was incorrectly placed by at least $2mm$. A false positive is when the classifier predicted a localization error of over $2mm$ but the predicted point was within $2mm$ of the ground truth.

Once we have obtained the ROC curve we can choose what True Positive Rate (TPR) we would like for our application; in this work we choose a rate of 0.5 to demonstrate the technique of ERE thresholding. This TPR corresponds to a threshold of 1.414. In other words, if our model outputs a heatmap for a landmark which has a ERE score of over 1.414 (such as two landmarks shown in Fig. 3b) we classify it as erroneous and 'flag it up'.

When we apply this threshold to Test Set 2 we find that it has classified 1610 landmarks (85%) as 'good' and 290 landmarks (15%) as 'erroneous'. The MRE and SDRs statistics for each group are in Table 2. The MREs of the 'good' and 'erroneous' groups are 1.30 and 2.43 respectively which makes it clear that the 'erroneous' group contains landmarks which have been detected significantly less accurately that the other group thus proving that thresholding the ERE value to discriminate between accurately and inaccurately predicted landmarks is reasonably effective.

Set	# land-marks	MRE (mm)	SDR (%)			
			2mm	2.5mm	3mm	4mm
Overall	1900	1.46	74.63	83.58	87.21	93.79
Good	1610	1.30	78.26	85.59	89.50	95.90
Erroneous	290	2.43	57.24	66.55	72.76	84.48

Table 2. Results of our model over Test Set 2, and over Test Set 2 again, after it has been split by thresholding the ERE value of each landmark's heatmap ($ERE > 1.414$ means a landmark is put into the Erroneous group). MRE is the Mean Radial Error and SDR is the Success Detection Rate over all the landmarks in the group as described in Section 5.1. A lower MRE score, and higher SDR percentages, signify that the 'Good' set of landmarks are localized more accurately than the 'Erroneous' set.

6. Conclusion

We have shown that formulating the problem of landmark detection as a classification task, by using one-hot heatmaps during training, has several practical advantages. The heatmaps output by the model are more easily interpretable and visually intuitive because they hug the contours of objects; they are more expressive than previous approaches because they can be multi-modal or asymmetric. We have shown that near state of the art localization performance can be achieved when formulating the problem this way, even with a U-Net architecture.

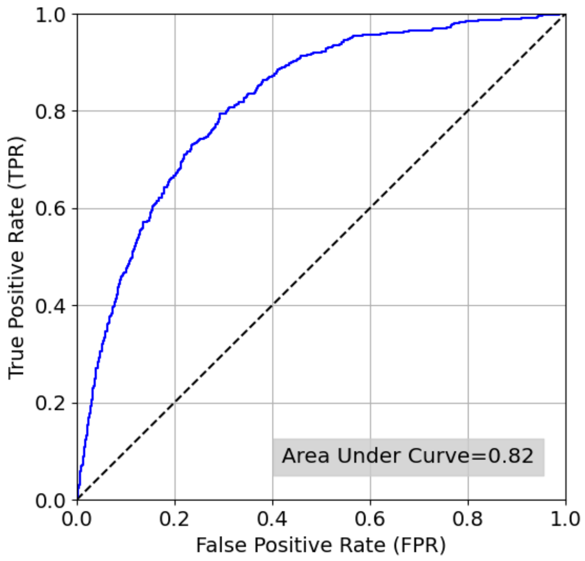


Figure 6. The ROC curve created by measuring the TPR and FPR as we increase the threshold for the ERE score to be for its landmark to be classified as ‘erroneous’. Area Under the Curve measures how well our ERE statistic discriminates between ‘good’ and ‘erroneous’ predictions, 1 being the best achievable score.

We then went on to validate the output heatmaps quantitatively, by showing that they were well calibrated using reliability diagrams, and that an Expected Radial Error (ERE) statistic could be calculated from them which is well correlated with the true radial error and thus can be thresholded to create a classifier which can ‘flag up’ potentially erroneous results.

Our approach to landmark detection should be valuable to the Computer Vision research community.

7. Future Work

This work can be extended in several ways. (1) Instead of simply having a one-hot value at the average ground truth position, should multiple-clinician ground truth be available, we could generate more expressive training heatmaps. For example, in the case where 2 sets of ground truth are available we could put a value of 0.5 in the training heatmap at each ground truth. $H(i, j)$ could be redefined for this purpose. (2) A more sophisticated calibration method could be used to validate the probabilities in the output heatmaps. (3) In addition, a current limitation of the approach is how the expressive heatmap must be condensed into a single statistic (in this case ERE) for its classification as a ‘good’ or ‘erroneous’ prediction. Instead of taking a single statistic we could take multiple statistics or even pass the heatmaps into another CNN which would classify more accurately. (4) More complicated deep learning architectures could also be used to improve the localization accuracy and uncertainty measurements. (5) In the short term, we plan to validate the heatmaps produced using this method more thoroughly by comparing them against measurements taken by multiple clinicians, as well as experimenting on more datasets.

8. Compliance with Ethical Standards

The aim of this research is to improve the safety and accuracy of existing landmark detection systems. However, when bringing this technology into contact with the general public it is always important to carefully check what data the system is handling and to make sure that the system has a suitable amount of human supervision. It is also important to consider how such a system could be tampered with, for example via an adversarial attack, and to take measures to stop that from happening.

This research study was conducted using human subject data from publicly available sources, which are known to have had ethical approval for using the data for research. There are no conflicts of interest to declare.

References

[1] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. LUVLi face alignment: Estimating landmarks’ location, uncertainty, and visibility likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8236–8246, 2020. 1, 2, 6

[2] Jeong-Hoon Lee, Hee-Jin Yu, Min-ji Kim, Jin-Woo Kim, and Jongeun Choi. Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks. *BMC oral health*, 20(1):1–10, 2020. 1, 2, 5

[3] Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) workshops*, pages 1272–1281, 2018. 1

[4] Florian Kordon, Peter Fischer, Maxim Privalov, Benedict Swartman, Marc Schnetzke, Jochen Franke, Ruxandra Lasowski, Andreas Maier, and Holger Kunze. Multi-task localization and segmentation for x-ray guided planning in knee surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 622–630. Springer, 2019. 1

[5] Jingru Yi, Pengxiang Wu, Qiaoying Huang, Hui Qu, and Dimitris N Metaxas. Vertebra-focused landmark detection for scoliosis assessment. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 736–740. IEEE, 2020. 1

[6] Martin Urschler, Christopher Zach, Hendrik Ditt, and Horst Bischof. Automatic point landmark matching for regularizing nonlinear intensity registration: Application to thoracic ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 710–717. Springer, 2006. 1

[7] Ching-Wei Wang, Cheng-Ta Huang, Meng-Che Hsieh, Chung-Hsing Li, Sheng-Wei Chang, Wei-Cheng Li, Rémy Vandaele, Raphaël Marée, Sébastien Jodogne, Pierre Geurts, et al. Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: a grand challenge. *IEEE transactions on medical imaging*, 34(9):1890–1900, 2015. 2, 5

[8] Sercan Ö Arik, Bulat Ibragimov, and Lei Xing. Fully automated quantitative cephalometry using convolutional neural networks. *SPIE Journal of Medical Imaging*, 4(1):014501, 2017. 2, 5, 7

[9] Zhusi Zhong, Jie Li, Zhenxi Zhang, Zhicheng Jiao, and Xinbo Gao. An attention-guided deep regression model for landmark detection in cephalograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 540–548. Springer, 2019. 2, 5

[10] Bulat Ibragimov, Boštjan Likar, F Pernus, and Tomaž Vrtovec. Automatic cephalometric x-ray landmark detection by applying game theory and random forests. In *Proc. ISBI Int. Symp. on Biomedical Imaging*, pages 1–8, 2014. 2, 5

[11] Claudia Lindner, Paul A Bromiley, Mircea C Ionita, and Tim F Cootes. Robust and accurate shape model matching using random forest regression-voting. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1862–1874, 2014. 2, 5

[12] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27:1799–1807, 2014. 2

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 2

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. 2, 3

[15] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Integrating spatial configuration into heatmap regression based cnns for landmark localization. *Medical image analysis*, 54:207–219, 2019. 2

[16] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision (ECCV)*, pages 483–499. Springer, 2016. 2

[17] Minmin Zeng, Zhenlei Yan, Shuai Liu, Yanheng Zhou, and Lixin Qiu. Cascaded convolutional networks for automatic cephalometric landmark detection. *Medical Image Analysis*, 68:101904, 2021. 2

[18] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019. 2

[19] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. In *Bildverarbeitung für die Medizin 2019*, pages 22–22. Springer, 2019. 2

[20] Qingsong Yao, Zecheng He, Hu Han, and S Kevin Zhou. Miss the point: targeted adversarial attack on multiple landmark detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 692–702. Springer, 2020. 2, 3, 5

[21] Heqin Zhu, Qingsong Yao, Li Xiao, and S Kevin Zhou. You only learn once: Universal anatomical landmark detection. *arXiv preprint arXiv:2103.04657*, 2021. 2

[22] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 2, 4, 5

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[24] Ching-Wei Wang, Cheng-Ta Huang, Jia-Hong Lee, Chung-Hsing Li, Sheng-Wei Chang, Ming-Jhih Siao, Tat-Ming Lai, Bulat Ibragimov, Tomaž Vrtovec, Olaf Ronneberger, et al. A benchmark for comparison of dental radiography analysis algorithms. *Medical image analysis*, 31:63–76, 2016. 4

[25] Runnan Chen, Yuexin Ma, Nenglun Chen, Daniel Lee, and Wenping Wang. Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 873–881. Springer, 2019. 5