Proceedings Paper:

Schobs, L., Zhou, S., Cogliano, M. et al. (2 more authors) (2021) Confidence-quantifying landmark localisation for cardiac MRI. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI 2021). IEEE International Symposium on Biomedical Imaging (ISBI 2021), 13-16 Apr 2021, Virtual conference. IEEE , pp. 985-988. ISBN 9781665429474

https://doi.org/10.1109/ISBI48211.2021.9433895

CONFIDENCE-QUANTIFYING LANDMARK LOCALISATION FOR CARDIAC MRI

Lawrence Schobs[⋆] *Shuo Zhou*[⋆] *Marcella Cogliano*[†] *Andrew J. Swift*[†] *Haiping Lu*[⋆]

[⋆] Department of Computer Science, the University of Sheffield, UK
[†]Department of Infection, Immunity & Cardiovascular Disease, the University of Sheffield, UK

ABSTRACT

Landmark localisation in medical imaging has achieved great success using deep encoder-decoder style networks to regress heatmap images centered around the target landmarks. However, these networks are large and computationally expensive. Moreover, their clinical use often requires human interaction, opening the door for manual correction of low confidence predictions. We propose **PHD-Net**: a lightweight, multi-task **P**atch-based network combining **H**eatmap and **D**isplacement regression. We design a simple **Candidate Smoothing** strategy to fuse its two-task outputs, generating the final prediction with quantified confidence. We evaluate PHD-Net on hundreds of Short Axis and Four Chamber cardiac MRIs, showing promising results.

***Index Terms*—** Landmark localisation, confidence, cardiac MRI, patch-based method, multi-task learning

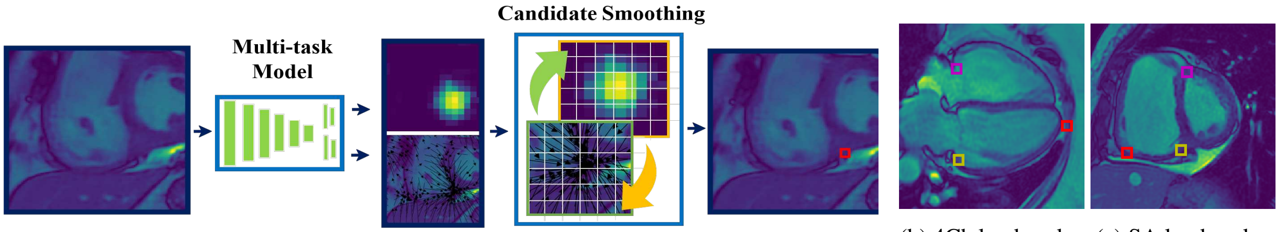
1. INTRODUCTION

Automated landmark localisation is an important step in medical image analysis, with convolutional neural networks (CNNs) dominating this task. For such deep learning methods, more data means better results [1]. However, in medical imaging, datasets are often limited in size due to high collection and annotation cost/difficulties. In order to extract as much value as possible from limited datasets, patch-based methods analyse an image patch-wise, generating a prediction of the coordinate for each small section in the image [2, 3]. While this approach can produce vast amounts of unique training samples, a clear drawback is that each prediction only takes into account a small local area.

This highlights another key challenge for landmark localisation in medical imaging: the prevalence of locally similar structures in an image. Patch-based approaches are particularly vulnerable to misidentifications from locally similar structures due to their strong local focus. Approaches pivoted to formulate the landmark localisation task as a heatmap estimation problem. For example, encoder-decoder style CNNs such as U-Net [4] or the Hourglass network [5] analyse the input image at several resolutions and output a Gaussian heatmap for each landmark [6, 7, 8]. Each point's activation on the heatmap can be seen as the pseudo-probability of it being the landmark. The network learns to generate a high response near the landmark, smoothly attenuating the responses in a small radius around it. Regressing heatmaps using the encoder-decoder style architecture facilitates both high and low level analysis of the image, mitigating the effect of locally similar but globally infeasible points in the image.

Less common are patch-based approaches that also regress heatmaps. Noothout *et al.* [9] comes close to this, applying multitask learning under a patch-based framework to jointly perform classification and regression on each patch. The classification task determines whether a patch contains the landmark, and the regression task estimates the 2D displacement from the patch to the landmark. Rather than regressing a heatmap, the classification task was formulated as a binary task: the patch containing the landmark was labelled 1, with the rest labelled 0. *Only* the patch classified as containing the landmark was used to determine the final coordinates. This multi-task, joint learning leads to a light-weight network and enhanced localisation performance, with the two tasks sharing a feature representation that improves the performance of both [10]. However, the resulting network has a strong local focus and is also susceptible to failure if the predicted containing patch is incorrect.

In order to increase robustness against misidentifications while being constrained by a small training set, we require a model that can learn rich feature representations while efficiently making use of the training data available. We also aim to produce a compact model, that is cheap to train. To this end, unlike encoder-decoder networks that train using full image samples, we opt to analyse images patch-wise, creating thousands of training samples from a single training image. To improve robustness while still being a lightweight model, we propose PHD-Net to perform two similar but distinctly separate tasks while sharing weights. Our main contributions are twofold: **(1)** A multi-task patch-based framework in which one branch of the network focuses on generating *locally accurate* candidate predictions, regularised by another branch focusing on the *globally likely* landmark location using heatmap regression. **(2)** A *Candidate Smoothing* strategy that combines the branch outputs to produce a locally accurate, globally feasible prediction, reducing misidentifications compared to the baseline approach [9]. We use this strategy to assign a confidence level to the prediction.



(a) The proposed PHD-Net framework. (b) 4Ch landmarks. (c) SA landmarks.

Fig. 1: **(a)** The *multi-task model* learns two regression tasks simultaneously and the *candidate smoothing* generates the final coordinate value and quantifies confidence. **(b)** Landmarks for 4 chamber (4Ch) CMR: Magenta = tricuspid valve; Yellow = mitral valve; Red = apex of left ventricle. **(c)** Landmarks for Short Axis (SA) CMR: Magenta = superior right ventricle insertion point valve; Yellow = inferior right ventricle insertion point; Red = inferior lateral reflection of right ventricle free wall.

2. METHOD

In the multitask network of [9], the regression and classification tasks share parameters in the convolutional layers. The network processes images patch-wise, with the regression task predicting the log 2D displacement from the centre of each patch to the landmark location. The classification task predicts whether the landmark is contained in the patch using a binary mapping. During training, subimages are randomly sampled from the image and used as training samples. In testing, the whole image is taken as input, and the displacement prediction from the patch with the highest classification score is used to calculate the landmark's predicted location [9].

PHD-Net has a similar formulation but two key differences: **(1) Heatmap regression**: Instead of considering the classification task as binary, we regress a Gaussian heatmap centered around the landmark-containing patch to provide smoother supervision. **(2) Coordinate calculation**: To improve the robustness and accuracy of the final coordinate prediction, we propose a *Candidate Smoothing* strategy. We consider each patch's prediction from the displacement output as a small Gaussian blob, producing *locally accurate* candidate predictions, and then regularise them by the predicted Gaussian heatmap from the *heatmap regression* branch.

Fig. 1a shows the framework for PHD-Net, illustrated on cardiac MRI (CMR). We adopt the architecture of [9] as the backbone of our network. In short, it composes three convolutional layers of 32 filters with 3×3 kernels, each followed by a maxpooling layer with 2×2 kernels. After these layers the input is broken into 8×8 pixel patches, each patch being represented by a single channel. Three convolutional layers with the same properties as follow, before branching into two sets of fully connected layers with 64 and 96 filters, modelled as 1×1 convolutional layers. One branch outputs the displacement prediction, and the other outputs the heatmap prediction. The model is compact with only 0.06M trainable parameters, enabling fast training.

2.1. Joint Displacement and Heatmap Regression

We make two predictions for each patch: the heatmap value and the displacement from the centre of the patch to the landmark. This provides two opportunities to discover the landmark: the *displacement regression* branch focuses on generating pixel-precise candidate coordinates, and the *heatmap regression* branch focuses on the more coarse object-detection task. Framing the task in this fashion facilitates predictions that are pixel-precise despite the output map's low resolution compared to the full image (due to patch-wise predictions, not pixel-wise). The total loss $\mathcal{L}_A$, consists of the displacement loss $\mathcal{L}_d$ and the heatmap loss $\mathcal{L}_h$:

$$\mathcal{L}_A = \mathcal{L}_d + \mathcal{L}_h. \tag{1}$$

The displacement loss $\mathcal{L}_d$ is a weighted sum of the mean squared error (MSE) between the predicted and annotated 2D displacement of each patch. The further the patch is from the landmark, the lower its predictive power. Thus, we dampen the effect of distant patches in two ways: (1) we apply the log function to the displacement labels [9] and (2) we weigh closer patches as more important than distant ones by multiplying the error of the patch-wise predictions by a Gaussian heatmap centered around the landmark.

The heatmap loss $\mathcal{L}_h$ is the MSE between the predicted patch-wise heatmap and the ground truth patch-wise heatmap. To generate the ground truth heatmap we define the mean as the patch containing the landmark, with a predefined standard deviation. For a landmark l_i contained in the patch (l_i^x, l_i^y), the 2D Gaussian heatmap image is defined as the 2D Gaussian function: $G_i(\mathbf{x} \mid \mu = (l_i^x, l_i^y); \sigma) : \mathbb{R}^d \to \mathbb{R}$.

The patch mapping's peak value is on the patch containing the landmark, with values smoothly attenuating with distance. Each patch's heatmap value now represents a psuedo-probability of the landmark being contained in it.

2.2. Candidate Smoothing

The next challenge is to calculate the final coordinate values from the model's outputs. We propose a strategy to combine the outputs from both branches into a final coordinate prediction value, increasing robustness against misidentifications and assigning a confidence level to the prediction. The key idea behind this strategy is to use a large number of patches to produce *locally precise* but ambiguous candidate predictions, which are then regularised to filter out the *globally unlikely* locations.

First, we find the 128×128 area section of the Gaussian heatmap with the highest summed activations. Second, for every patch contained in this area, we plot the prediction from the displacement branch as a small Gaussian blob with a standard deviation of 1. The mapping is additive, meaning if multiple patch's predictions overlap, the Gaussian values add on to each other. This produces a 128×128 mapping containing pixel-precise candidate locations for the landmark, $M_i^c(\mathbf{l_i})$:

$$M_i^c(\mathbf{l_i}) = \sum_{j=1}^{P} G_i(c_j^i + d_j^i; \sigma = 1), \tag{2}$$

where for each of P patches in the 128×128 subimage, c_j^i is the center of the patch and d_j^i is the inverse log predicted displacement. The candidate points are precise to a local degree, but since each patch predicts a location blind to its surroundings, it can fail due to locally similar structures.

To solve this we smooth the mapping by multiplying it with the up-sampled and smoothed Gaussian heatmap predicted by the heatmap branch $G_i(\overline{\mathbf{x}}; \mathbf{w}, \mathbf{b})$ to create a smoothed map:

$$M_i^s(\mathbf{x}) = G_i(\overline{\mathbf{x}}; \mathbf{w}, \mathbf{b}) \odot M_i^c(\mathbf{l_i}). \tag{3}$$

Multiplying the mapping by the predicted Gaussian heatmap suppresses the globally infeasible predictions determined by the heatmap regression branch, while retaining pixel-precise predictions from the displacement regression branch. To obtain the final coordinate value, we take the peak pixel of the new heatmap.

We assign a confidence level to each prediction. During validation, we determine a threshold by calculating a weighted average of the 10% least accurate predictions' peak values, weighted according to the magnitude of the error. In testing, if the final heatmap's peak value is below this threshold, we can infer that there was no clear consensus among the patches of the landmark's location, and consider it *low confidence* prediction. Otherwise, the prediction is considered *high confidence*.

3. EXPERIMENTS AND RESULTS

For all experiments we trained PHD-Net for 500 epochs using a batch size of 32 and a learning rate of 0.001, using the Adam Optimiser. Early stopping was employed if the validation set's loss was not improved for 75 epochs. The sizes of the sub-images used in training were 128×128 pixels. All landmark localisation experiments were conducted using a fixed 8-fold cross validation.

3.1. Data

We evaluate PHD-Net on a dataset from the ASPIRE Registry [11]. Each subject has a four chamber (4ch) view and/or a short axis view (SA). Each CMR sequence has a spatial resolution of 512×512 pixels, where each pixel represents 0.9375mm of the organ, and 20 frames (we use the first frame). There are 303 SA images, and 422 4ch images, each with three annotated landmarks (shown in Fig. 1b and Fig. 1c). The 4ch dataset represents a more challenging landmark localisation task as the images have much higher variability than the SA dataset.

Table 1: PHD-Net results for binary and Gaussian ($std = 2$) maps and different coordinate calculation strategies. Mean error and standard deviation (std) are in mm across all landmarks over a fixed 8-fold cross validation.

Mapping	Coordinate Calculation	Error ± std (mm)
Binary [9]	Simple [9]	22.76 ± 29.18
Gaussian	Simple [9]	6.08 ± 23.64
Gaussian	Candidate Smoothing	**4.73 ± 15.39**

3.2. Evaluation of Heatmaps and Candidate Smoothing

We perform an ablation study on the SA images to demonstrate the effectiveness of our proposed loss and coordinate calculation strategy compared to our baseline [9]. Table 1 shows the results comparing using a binary map to a Gaussian heatmap, where we experimentally select a standard deviation of 2. We find using a Gaussian heatmap noticeably outperforms a simple binary map, due to its smoother supervision and ability to encode some uncertainty in the prediction. The table also demonstrates that using our Candidate Smoothing strategy outperforms solely using the highest classifying patch [9]. Best performance was seen when using both heatmaps and Candidate Smoothing.

A common error with the naive coordinate resolution is landmark misidentification, causing gross errors. Since the simple strategy only considers the patch with the highest classification prediction value, the information from the surrounding patches is ignored, and a small error in the classification branch can lead to a complete misidentification. The *Candidate Smoothing* strategy ensures that even when the Gaussian activation is not particularly high on the correct patch, the additive consensus the patches achieved from the regression branch can overpower the suppression from the failed classification branch in Eq. (3).

3.3. Comparison

We evaluate PHD-Net on both SA and 4Ch images, comparing it to the baseline network [9] and two other approaches: **(1) Hourglass [5]**: We follow the authors' description to implement the model, downscaling the input images to 256×256 pixels, and learning a 64×64 heatmap for each landmark. We use a single stacked hourglass, leading to 6M trainable parameters. We train for a maximum of 1000 epochs using the Adam optimiser, employing early stopping. Through experimentation, we selected a learning rate of 0.001, batch size of 3 and standard deviation of 1 for the Gaussian labels. **(2) U-Net [4]**: We use the MONAI framework[1], designing the model with 5 encoding-decoding levels, creating 1.63M learnable parameters. Again, we downsample the input image to 256×256 pixels to create more capacity parity with

[1]Project MONAI, www.github.com/Project-MONAI

Table 2: Localisation error in mm. The *All* group represents all images in the dataset, and *HiC* represents the subset of images PHD-Net considered as *high confidence*.

Model	Short Axis Images		4 Chamber Images	
	All (100 %)	HiC (56 %)	All (100 %)	HiC (42 %)
Baseline [9]	24.79 ± 31.82	24.98 ± 33.98	52.90 ± 35.58	24.98 ± 33.98
Hourglass [5]	5.76 ± 8.48	4.54 ± 4.61	13.33 ± 21.63	8.40 ± 12.71
U-Net [4]	5.93 ± 12.75	4.22 ± 6.52	**7.78 ± 9.82**	5.72 ± 4.50
PHD-Net	**4.73 ± 15.39**	**2.97 ± 2.20**	9.51 ± 25.89	**4.40 ± 4.58**

PHD-Net. The output heatmaps are full size (256×256). We train for a maximum of 1000 epochs employing early stopping, with an experimentally selected batch size of 2, learning rate of 0.001 using the Adam Optimiser and standard deviation of 8 for the Gaussian labels.

Table 2 shows the results for both SA and 4ch images. For SA, PHD-Net performs the best by a significant margin, also discriminating well between *high confidence* and *low confidence* predictions. For 4ch, U-Net has the lowest error, followed by PHD-Net. However, when considering the predictions PHD-Net indicated as *high confidence*, PHD-Net performs the best. Fig. 2 shows the difference between all 4ch predictions and those labelled as *high confidence* by PHD-Net more clearly. The 4ch images were more challenging, with 14% less *high confidence* predictions. In addition, almost all models performed better on the high confidence subset, indicating PHD-Net is truly discriminating between difficult and easy images. Finally, Hourglass and U-Net models [5, 4] can predict multiple landmarks at once compared to PHD-Net's single landmark prediction, but they respectively have 100× and 27× more learnable parameters, making them significantly more expensive to train.

4. CONCLUSION

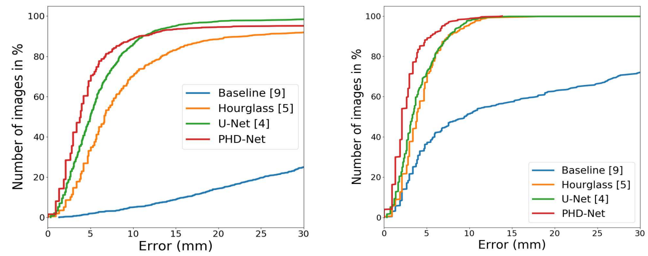
This paper proposed a lightweight, confidence-quantifying model for landmark localisation for cardiac MRI, named as PHD-Net. It takes a patch-based, multi-task approach with joint heatmap and displacement regression. It uses a candidate smoothing strategy to fuse multi-task outputs to generate the final prediction and quantify the confidence. We performed evaluation on a dataset covering two scanning protocols. PHD-Net achieved localisation error better or similar to more expensive comparison models and can accurately discriminate between high and low confidence predictions.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research was conducted retrospectively using the ASPIRE registry from the Sheffield Teaching Hospitals NHS Foundation Trust. Ethical approval was granted from the local ethics committee and institutional review board for this retrospective study (ref c06/Q2308/8).

6. ACKNOWLEDGMENTS

This work was supported by EPSRC (2274702) and the Wellcome Trust (215799/Z/19/Z and 205188/Z/16/Z).



(a) All. (b) High-confidence in PHD-Net.

Fig. 2: Cumulative distribution of localisation errors on 4Ch images.

7. REFERENCES

[1] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *ICCV*, 2017.

[2] Omar Emad, Inas A Yassine, and Ahmed S Fahmy, "Automatic localization of the left ventricle in cardiac mri images using deep learning," in *EMBC*, 2015.

[3] Yuanwei Li et al., "Fast multiple landmark localisation using a patch-based iterative network," in *MICCAI*, 2018, pp. 563–571.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[5] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016, pp. 483–499.

[6] Zhusi Zhong et al., "An attention-guided deep regression model for landmark detection in cephalograms," in *MICCAI*, 2019, pp. 540–548.

[7] Aleksei Tiulpin et al., "Kneel: Knee anatomical landmark localization using hourglass networks," in *ICCV Workshops*, 2019.

[8] Christian Payer et al., "Integrating spatial configuration into heatmap regression based cnns for landmark localization," *Medical Image Analysis*, pp. 207 – 219, 2019.

[9] Julia Noothout et al., "Cnn-based landmark detection in cardiac cta scans," in *Proc. Medical Imaging with Deep Learning*, 2018, pp. 1–11.

[10] Sebastian Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[11] J. Hurdman et al., "ASPIRE registry: assessing the spectrum of pulmonary hypertension identified at a REferral centre," *European Respiratory Journal*, 2012.