Supplementary Material: Contour-Hugging Heatmaps for Landmark Detection

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

1. Qualitative analysis of heatmaps for landmark prediction

In this section we examine localisation errors qualitatively by providing extra figures displaying the heatmaps our model outputted, as well as our model's **predicted points (shown as red dots)** and the **ground truth points (shown as green dots)**. We briefly examine examples of accurate predictions and then show examples where our model is incorrect because we believe these will be interesting for the research community. Figure 1 shows the ordering of the cephalometric landmarks for reference.

1.1. Accurate predictions

We show accurate predictions in Figure 2. These are examples where the radial error (euclidean distance between the predicted point and the ground truth point) for each landmark is less than 1.5mm. In most cases the heatmaps for these landmarks are small and condensed around the hottest point.

1.2. Moderately accurate predictions

We show moderately accurate predictions of landmark 19 in 3a, 16 in 3b, 6 in 3c and 4 in 3d. We define a moderately accurate prediction as a prediction with a radial error of between 2mm and 4mm. We can see that sometimes the heatmap for the prediction is spread out and covers the ground-truth point such as for landmark 19 in 3a. However this is not always the case as seen for landmark 6 in 3c. In this case the ground-truth point has been placed away from the edge of the skull and, as a consequence, the output heatmap has a very low value at that point. It is possible that this ground-truth point could have been placed incorrectly and should be right on the edge of the skull although we need an expert in cephalometry to confirm this.

1.3. Unsuccessful predictions

When our model is run over test set 1 and 2 it predicts a total of 10 landmarks with a radial error of over 10mm. A radial error that large signifies a landmark detection failure. After investigation we find that there are two causes for detection failures. The first is a phenomenon where an output

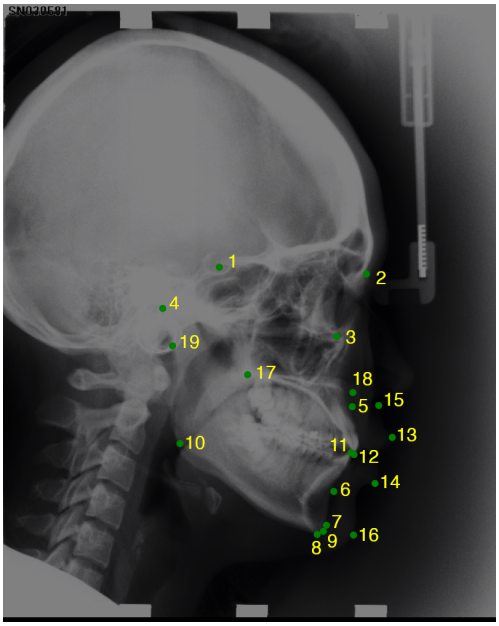


Figure 1. This figure shows the ordering of the landmarks and can be used to understand where on the head the cropped patches displayed in the other figures are taken from.

heatmap for a landmark contains high values around a landmark it is not meant to be localising. Examples of this can be seen in Figure 4 and Figure 5. Both images in Figure 4 show that landmark 19's heatmap contained higher values around landmark 4 than landmark 19 itself and as such the network has predicted landmark 4 and 19 to be very close to each other, which, in turn, means that landmark 19 has a very high localisation error. A similar situation is true in Figure 5 where landmark 16 has been erroneously placed very close to landmark 14. In addition, Figure 5a is slightly different to the others in that there are no pixels of significant value around the true position of landmark 16.

The second reason for a landmark detection failure is when the input image is significantly different to the images the model trained on, otherwise known as domain shift. The most obvious example of this is Figure 6 which shows an x-ray with an unusual appearance compared to others in the

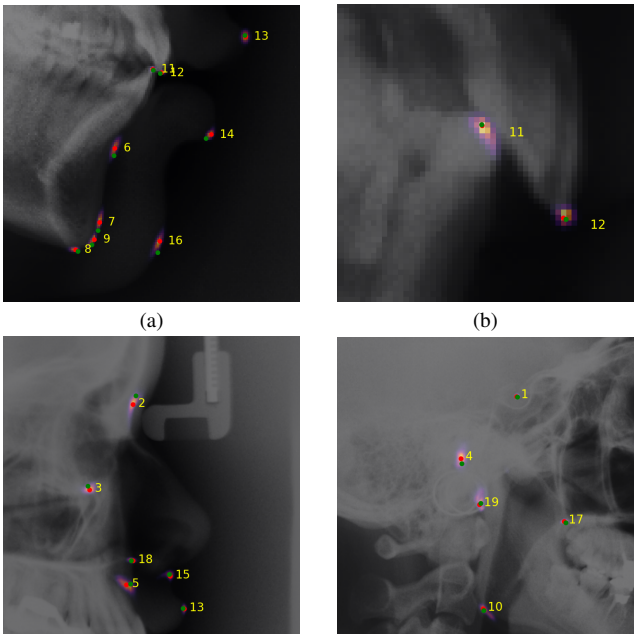


Figure 2. Examples of accurate localisations.

dataset. There seem to be metal artifacts present in the image and the person's teeth are aligned irregularly. As a result some of the landmarks are poorly localised, especially landmark 1 which has a localisation error of 23.1mm.

Future work should attempt to remedy the problems which cause prediction failures to improve the localisation success rate of the model. However our work goes some way towards this goal by flagging up predictions for which the Expected Radial Error (ERE) is too high (above a threshold) which is the case for the examples we see in Figures 4, 5 and 6.

2. Landmark 16 in test set 1 vs test set 2

It appears that landmark 16 has been consistently placed lower by the senior annotator in the training set and test set 1 than in test set 2. This means that our model, when trained on the training set, learns to place landmark 16 in a lower position. This can be seen when comparing Figure 7 to Figure 8. Figure 7 shows the ground-truth placement and our model's prediction for landmark 16 on images in test set 1. The ground-truth point is either very close or below our model's prediction. Figure 8, on the other hand, shows images from test set 2 and we can see that the ground-truth is consistently higher than the predicted point. This is reflected in the fact that the mean radial error our model achieves for landmark 16 over test set 1 is 1.358mm compared to 4.202mm for test set 2. In future it could be more scientifically sound to exclude landmark 16 from experiments involving test set 2.

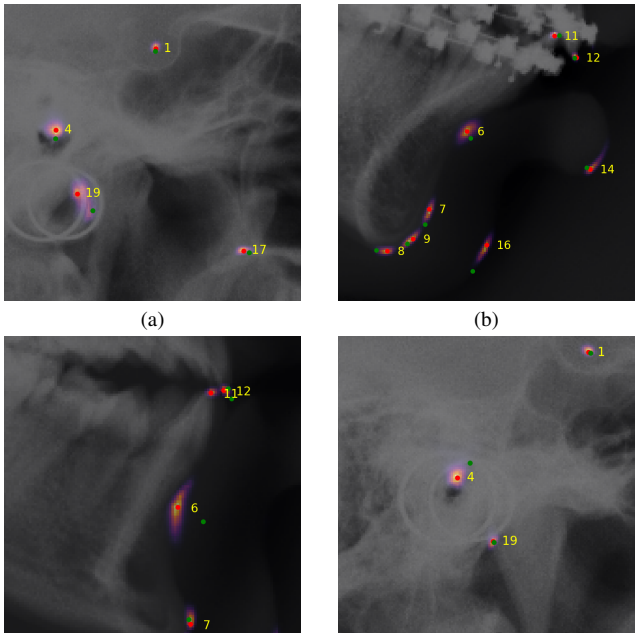


Figure 3. landmark 19 in 3a, 16 in 3b, 6 in 3c and 4 in 3d are examples of moderately accurate localisations.

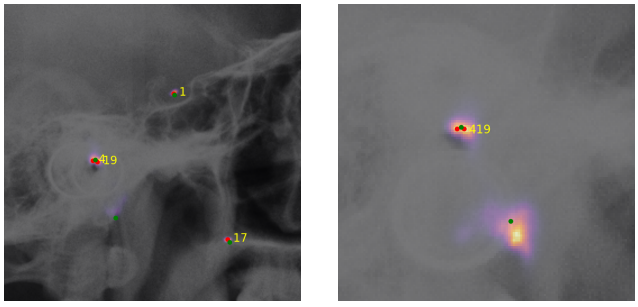


Figure 4. Examples where the heatmap for landmark 19 has high values around landmark 4.

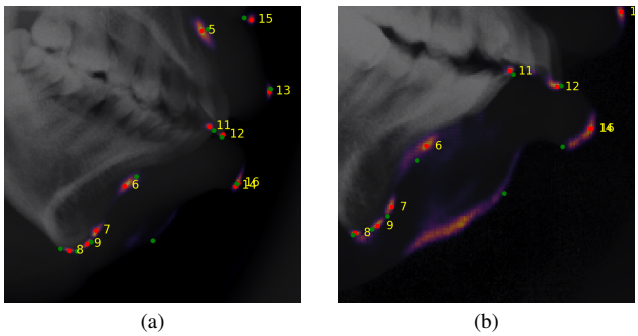


Figure 5. Examples where the heatmap of landmark 16 has high values around landmark 14.

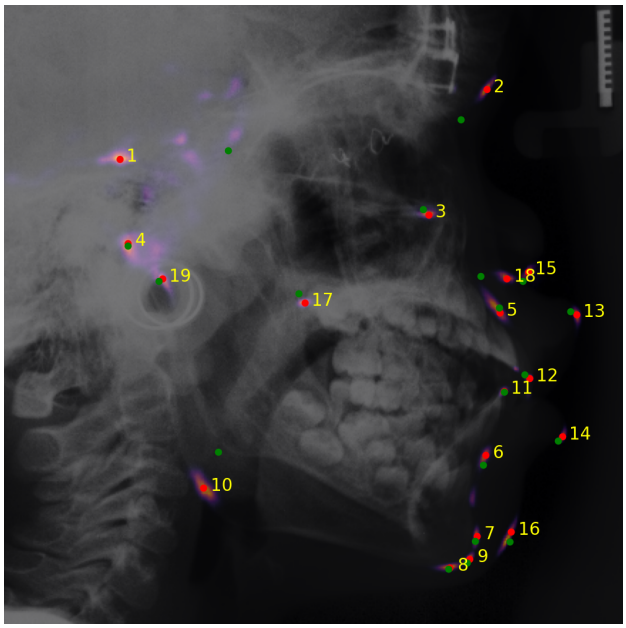
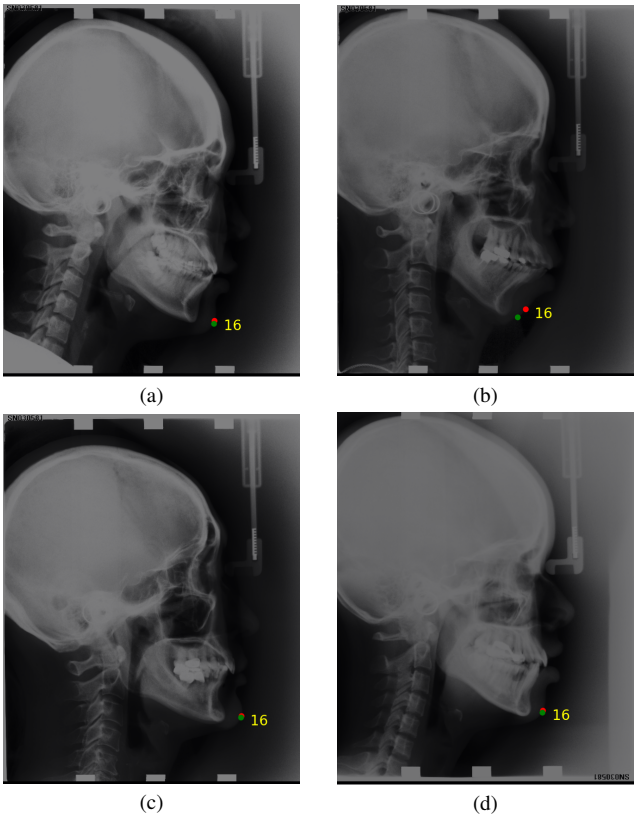
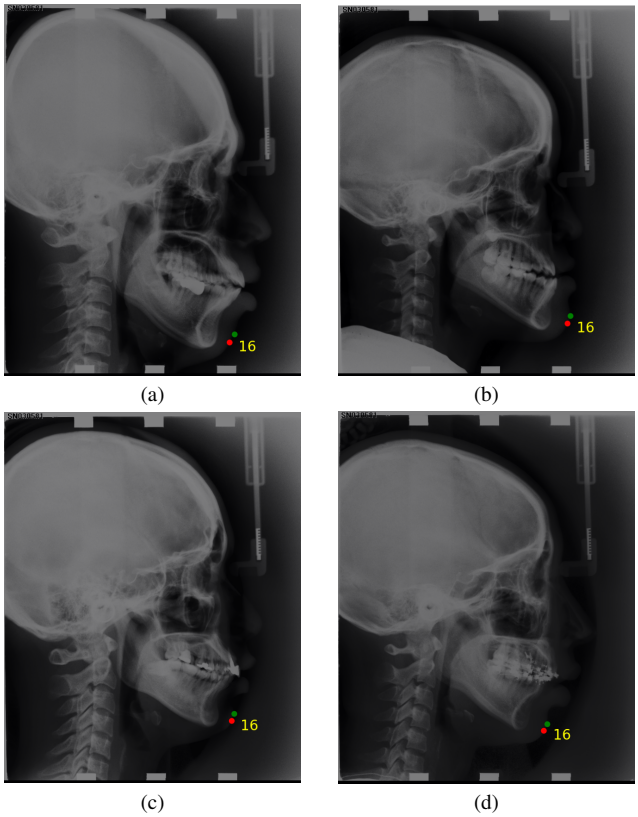


Figure 6. This figure shows an x-ray from test set 1 which contains metal artifacts and unusual presentation of the teeth. Our model fails to locate 4 landmarks on this image successfully.



(a) (b) (c) (d)

Figure 7. Shows the ground-truth placement of landmark 16 as the green dot (and our model's landmark prediction as the red dot) on images from test set 1.



(a) (b) (c) (d)

Figure 8. Shows the ground-truth placement of landmark 16 as the green dot (and our model's landmark prediction as the red dot) on images from test set 2.