Supplementary material for the paper "Automatic assessment of Pectus Excavatum severity from CT images using deep learning"

Bruno Silva, Inês Pessanha, Jorge Correia-Pinto, Jaime C. Fonseca, and Sandro Queirós

APPENDIX A

Five alternative models were tested in the axial keypoint detection network, based on Unet [S1], VGG19 [S2], VGG16 [S2], MobileNetV1 [S3], and MobileNetV2 [S4] architectures. The Unet architecture is similar to the Unet++ model present in Fig. 4, but without the convolutional blocks present in the skip pathways. Figure S1 presents the implemented models for the other architectures, which are based on the model proposed in the *OpenPose* framework [S5]. Specifically, each architecture has a head (dashed in read) that regresses the keypoints' heatmaps from the features extracted by a backbone network (one of four, namely VGG19, VGG16, MobileNetV1 and MobileNetV2). The head was based on the CMU heads used in OpenPose [S5], being composed of 4 convolution blocks (in the case of MobileNetV1 and V2, depthwise separable convolutions and residual bottleneck blocks were used, respectively), in which the first 3 blocks contain a total of 128 filters and the last block has 512 filters. In the architecture with the MobileNetV1 backbone, features from three different levels were concatenated (with the respective up- or downsample to match the second level's size, *i.e.* 4 times lower than the input image) and used as feature maps for the network's head. Contrarily, in the MobileNetV2, the feature maps used in the head result from the concatenation of four levels of the backbone, all resampled to match the second level's size (also 4 times lower than the input image size).

To reduce the amount of overfit to the training set, for each network, the amount of l2-regularization was tuned. Thus, a weight of 0, 0, 5×10^{-5}, 5×10^{-5} and 1×10^{-4} was used for VGG19, VGG16, MobileNetV1, MobileNetV2 and UNet, respectively.

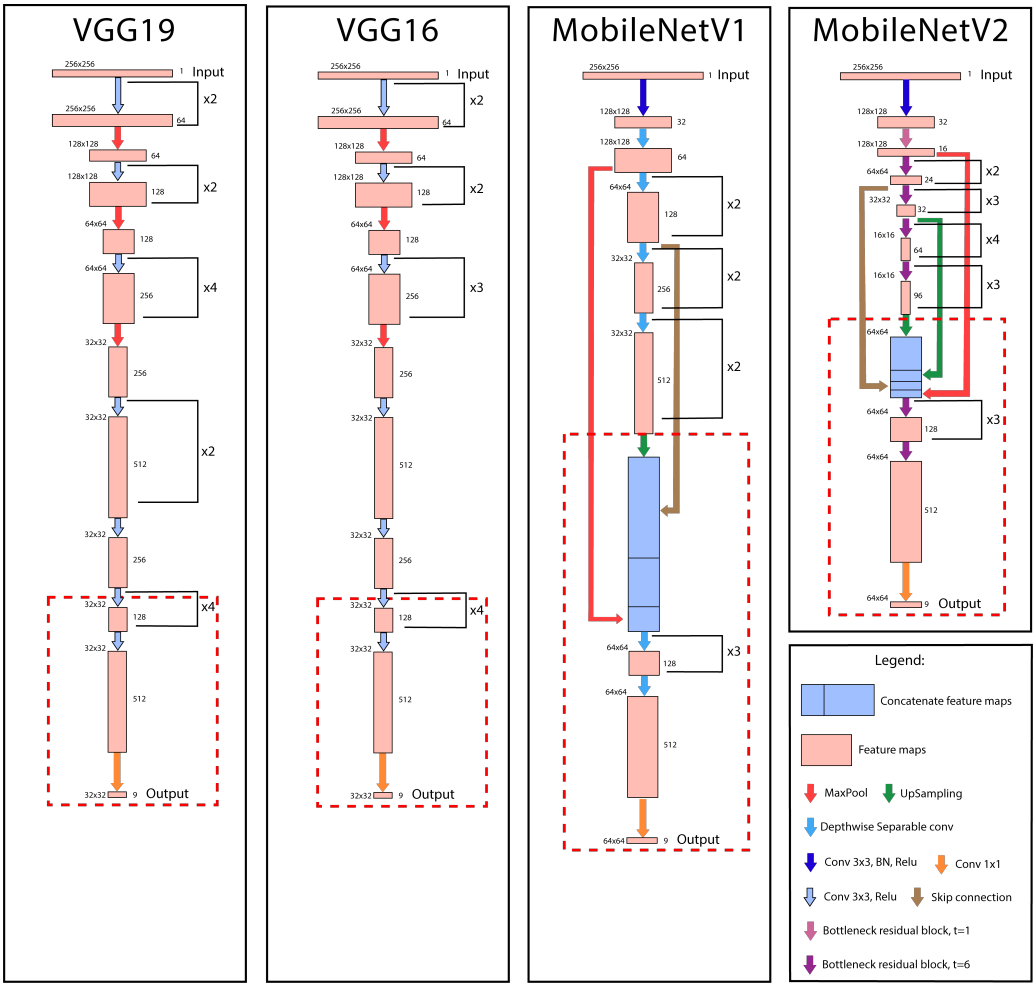


Fig. S1. Illustrations of four alternative architectures tested for the axial keypoint detection network, contemplating a backbone (based on VGG19, VGG16, MobileNetV1 and MobileNetV2) to extract a set of feature maps, and a head (dashed red rectangle) to regress the keypoints' heatmaps.

APPENDIX B

To test how each module influences the overall performance of the developed framework, an ablation study was carried out using the test set (section III-E). In short, one selectively enables or disables each of the framework's main stages, namely the sagittal module, the axial module, and the axial ROI-based processing and associated post-processing algorithm for optimal slice selection. Four combinations were thus tested: (1) enabling the axial module only, using the expert-select axial slice; (2) enabling both sagittal and axial modules, but without using the proposed axial ROI-based processing (*i.e.* only the axial slice chosen by the sagittal module is used as input for the axial module, and no post-processing selection of the optimal axial slice is thus performed); (3) the proposed strategy, in which all modules are enabled and a 2-cm wide ROI is used for optimal selection of the axial slice on which to extract the measurements; and (4) all modules enabled but with a wider axial ROI (4 cm), to evaluate whether increasing the search region for optimal axial slice selection improves the system's performance.

Table S1 summarizes the results (correlation coefficient, two-tailed paired *t*-test between indices, and bias and standard deviation from the Bland-Altman analysis). Overall, all tested combinations obtained a remarkable agreement with the expert's values, showing the robustness of the proposed modules with regard to each other. When analyzing the results of the isolated axial module (first row), one observes an excellent correlation with the expert's values and narrow limits of agreement (LOA). This finding is not unexpected, given that one is using the axial slice selected by the expert as the input for the axial module, therefore bypassing one of the major sources of potential disagreement (the selection of the correct slice on which to extract the measurements). This is in accordance with the results observed when enabling the sagittal module (second row), in which the agreement with the expert is slightly decreased (note the wider LOA and the lower correlation coefficient for all indices). Interestingly, this lower agreement is not observed when further enabling the axial ROI-based processing (third row; proposed), demonstrating the benefits of the proposed region-based analysis for optimal selection of the axial slice on which to extract the measurements (therefore reducing the sensitivity of the framework to the performance of the sagittal module). In fact, for both the Haller and asymmetry indices, the proposed strategy outperforms the isolated axial module, with only slightly lower agreement for the correction index. The latter might be explained by the increased sensitivity of this index with regard to the slice chosen (as the virtual correction distance is intimately linked to the ribs' position and, thus, to the chest curvature across the thorax) and the implicit difference in selection technique between the expert and the proposed framework. While the expert selects the axial slice by looking at the sagittal cuts and searching for the sternum's greatest depression point, the framework selects the optimal axial slice by searching for the minimum anteroposterior distance (*i.e.* distance between sternum and spine). Although this distance is supposed to be minimal in the region of the sternum's maximum depression, the spine's normal curvature may result in slight differences in the selected slice. Finally, when increasing the search region to 4 cm (fourth row), a lower agreement is observed for most parameters. Since the rib cage tends to become narrower at the ends, if the ROI is too large, this causes the network to select axial slices in these invalid regions (especially areas below the base of the heart). This explains the drastic rise in both bias and standard deviation for the correction index.

REFERENCES

[S1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[S2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[S3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[S4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[S5] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

TABLE S1
ABLATION STUDY ON THE FRAMEWORK'S MAIN MODULES

Sagittal module	**Axial module**	**ROI (cm)**	Haller index			Correction index			Asymmetry index		
			r	***p***	**BA**	r	***p***	**BA**	r	***p***	**BA**
✗	✓	0	0.984	0.027	-0.15 ± 0.46	0.972	0.082	-0.77 ± 3.12	0.972	0.028	0.52 ± 1.67
✓	✓	0	0.979	0.023	-0.16 ± 0.50	0.964	0.121	-0.81 ± 3.71	0.961	0.147	0.38 ± 1.84
✓*	✓*	2*	0.984	0.263	-0.06 ± 0.41	0.967	0.412	0.39 ± 3.40	0.969	0.080	0.41 ± 1.64
✓	✓	4	0.973	0.912	0.01 ± 0.47	0.940	0.051	1.20 ± 4.33	0.962	0.098	0.42 ± 1.80

*Proposed; r: Pearson's correlation coefficient; p: p-value on a two-tailed paired t-test against the expert's values; BA: Bland-Altman results, as bias ± standard deviation.