Deep Learning-Based Regression and Classification for Automatic Landmark Localization in Medical Images

Julia M. H. Noothout, Bob D. de Vos, Jelmer M. Wolterink, Elbrich M. Postma, Paul A. M. Smeets, Richard A. P. Takx, Tim Leiner, Max A. Viergever and Ivana Išgum

Abstract—In this study, we propose a fast and accurate method to automatically localize anatomical landmarks in medical images. We employ a global-to-local localization approach using fully convolutional neural networks (FCNNs). First, a global FCNN localizes multiple landmarks through the analysis of image patches, performing regression and classification simultaneously. In regression, displacement vectors pointing from the center of image patches towards landmark locations are determined. In classification, presence of landmarks of interest in the patch is established. Global landmark locations are obtained by averaging the predicted displacement vectors, where the contribution of each displacement vector is weighted by the posterior classification probability of the patch that it is pointing from. Subsequently, for each landmark localized with global localization, local analysis is performed. Specialized FCNNs refine the global landmark locations by analyzing local sub-images in a similar manner, i.e. by performing regression and classification simultaneously and combining the results. Evaluation was performed through localization of 8 anatomical landmarks in CCTA scans, 2 landmarks in olfactory MR scans, and 19 landmarks in cephalometric X-rays. We demonstrate that the method performs similarly to a second observer and is able to localize landmarks in a diverse set of medical images, differing in image modality, image dimensionality, and anatomical coverage.

Index Terms—Landmark localization, Convolutional Neural Network, Deep Learning, Classification, Regression, Cardiac CT, Cephalometric X-ray, Olfactory MR

I. INTRODUCTION

IDENTIFICATION of anatomical reference points and landmarks is a prerequisite for numerous medical image analysis tasks [1]. These include image registration [2]–[6], initialization of segmentation methods [7], and computation of clinical measurements for patient diagnosis and treatment planning [8]–[13]. While manual identification of anatomical landmarks might be trivial, it is often tedious and cumbersome [8], [14]. Fast and accurate automatic landmark localization methods can replace manual identification and may be especially helpful when precise localization of multiple landmarks is required.

Several application-specific automatic landmark localization methods have been proposed previously, such as methods combining segmentation of specific structures containing the landmarks and subsequent local rule-based analysis of those structures [14]–[16]. More generic localization methods employ either multi-atlas image registration [6], [17] or machine learning. In multi-atlas image registration, multiple atlas images with annotated landmarks are registered to the image of interest. Subsequently, a voting scheme determines the location of landmarks. Such approaches are accurate and robust to limited diversity in the anatomy and image acquisition, but they are typically time-consuming [6], [17]. Machine learning provides a faster and more robust alternative.

Conventional machine learning approaches for landmark localization in medical images are often classification- [7], [12], [16], [18]–[23] or regression-based [5], [7], [9], [22]–[26]. Classification-based methods detect the presence of a landmark in image slices, patches, or voxels. Classification methods use a hard threshold: the landmark is either present or absent. Therefore, these methods usually rely on careful consideration of a final threshold value, which may be data and task specific. Regression-based methods circumvent the use of a hard threshold by outputting a continuous value [27]. Regression-based methods predict the displacement or distance to the landmark from image slices, patches, or voxels.

Similar to many other automatic image analysis tasks, automatic landmark localization methods have become primarily deep learning-based [28]–[30]. Deep learning methods outperform conventional machine learning methods in a wide range of applications [28]. The advantage of deep learning is that it does not require handcrafting of features.

Several deep learning methods have been proposed for landmark localization that employ classification. Yang et al. [31] classified image slices with a convolutional neural network (CNN) and predicted a landmark location based on intersecting the classification outputs from all axial, coronal and sagittal image slices. Zheng et al. [29] localized a landmark by classifying image voxels with multi-layer perceptrons, while Arik et al. [32] performed pixel classification with a CNN to
localize landmarks. Xu et al. [33] localized landmarks with a CNN that classified pixels based on their relative position (up, down, left or right) to the landmark of interest. Subsequently, landmarks were localized by using the obtained pixel-wise action steps.

Deep learning-based methods exploiting regression often predict heatmaps representing e.g. the distance between evaluated voxels and the landmark of interest [10], [30], [34]–[37]. Landmarks are identified as local or global minima in these heatmaps. Voxel labels in heatmaps can be seen as pseudo-probabilities, indicating how close a voxel is located to a landmark. This makes heatmap regression comparable with voxel classification without using a hard threshold. Wolterink et al. [34] employed a CNN containing dilated convolutions to predict heatmaps indicating landmark locations. Similar to Wolterink et al. [34], Payer et al. [35] and O’Neil et al. [30] also proposed methods to predict heatmaps for automatic landmark localization. Payer et al. [35] used a CNN that combined local appearance responses of a single landmark with the spatial configuration of that landmark to all other landmarks while O’Neil et al. [30] employed a CNN that analyzed low resolution images and subsequently used a second CNN for further refinement. Torosdaghi et al. [10] used a CNN to predict heatmaps representing the geodesic distance to a segmented organ containing landmarks and subsequently used a long short-term memory classification network to localize landmarks placed closely together. Unlike methods that performed a single task at a time, Meyer et al. [37] used a multi-task network to determine which landmark was closest to an analyzed pixel and subsequently predicted the normalized 2D distance towards that landmark.

Heatmap regression often requires combining a large number of predictions, for instance via a majority voting strategy, making it computationally expensive and often time-consuming. Therefore, a different approach was chosen by Zhang et al. [38], who used a CNN to predict displacement vectors indicating the distance and direction from an analyzed voxel towards the landmark of interest. Subsequently, the CNN-architecture was expanded with additional layers to model correlations between analyzed input patches and output predicted landmark coordinates. Even though good results were obtained, predicting landmark coordinates directly from the image might be prohibited to large and complex CNN-architectures that model the complex non-linear mappings from input image to landmark location.

Besides deep learning-based regression methods that directly predict landmark locations, regression has also been used to iteratively determine the landmark locations in an image [39]–[43]. Aubert et al. [39] used a network to regress the displacement from an initial input patch, chosen with a statistical shape model, to the reference landmark. The landmark position was obtained by iteratively moving the input patch, using the predicted displacements, until convergence was reached and the landmark was localized. Li et al. [40] localized landmarks in an iterative manner and employed a CNN that predicted the distance along each of the three coordinate axes from the center of 2.5D patches towards the landmark of interest while using classification to predict positive or negative movement along each coordinate axis. Ghesu et al. [41], Alansary et al. [42], and Al et al. [43] localized landmarks exploiting deep reinforcement learning to obtain the optimal search path from an initial starting location towards the landmark of interest.

In this study, we propose a global-to-local localization approach, where an initial FCNN predicts the global locations of multiple landmarks simultaneously, and subsequently specialized FCNNs refine the final location of each landmark. Global multi-landmark localization and subsequent local single landmark localization are performed in a similar manner. While previous landmark localization methods used one approach, either classification or regression, we propose a patch-based fully convolutional neural network (FCNN) that performs both classification and regression (shown in Fig. 1). A patch-based approach provides a computationally efficient alternative to voxel-based approaches. However, since patch-based classification is inherently less precise than its voxel-wise counterpart, we mitigate this by jointly regressing the displacement vectors that point to the location of the landmark. Conversely, using a regression-only localization approach might lead to sub-optimal localization results, because we postulate that displacement vectors predicted in image patches farther from the landmark of interest are less accurate than displacement vectors predicted in image patches closer to it. This can be mitigated by employing the posterior probabilities from the classification task as weights for weighted averaging of the displacement vectors. Combining regression and classification results in a landmark localization method that is both fast and highly accurate. We show that our method is generally applicable to a variety of landmark localization tasks: 8 landmarks in 3D coronary CT angiography (CCTA) scans, 2 landmarks in 3D olfactory MR scans, and 19 landmarks in 2D cephalometric X-rays. Additionally, we show that our method is able to localize single landmarks and multiple landmarks simultaneously.

II. Method

We propose an automatic landmark localization method that employs a global-to-local estimation of landmark locations (Fig. 2). During global landmark localization, a fully convolutional neural network analyzes full input images in
After the four ResNet-blocks, the network has two output heads: The first CNN provides global estimates of landmarks for the second specialized CNNs that predict final landmark locations. Since the CNNs are fully convolutional, they can handle input images of any size. This network contains a strided convolutional layer of one convolutional layer with 16 \((7 \times 7 \times 7)\) kernels. In contrast to the original ResNet34 [44], which contains a strided convolutional layer as first layer in every ResNet-block, our network contains a pooling layer before the first and second ResNet-block, which is an average pooling layer with a size and stride of \(2 \times 2 \times 2\) voxels. After the four ResNet-blocks, the network has two output heads: one for regression of displacement vectors, and another for classification of landmark presence. Both output heads are similar in design. Each head has two 256-node dense layers and an output-layer, implemented as \(1 \times 1 \times 1\) convolutions [45]. The classification head outputs scalars, one for each landmark, forced between 0 and 1 by a sigmoid function. The regression head outputs displacement vectors for each landmark.

The global FCNN is based on ResNet34 [44] and it consists of one convolutional layer with 16 \((7 \times 7 \times 7)\) kernels and a stride of 2, which is followed by 4 ResNet-blocks. One ResNet-block contains 3, 4, or 6 convolutional layer pairs, where each convolutional layer pair consist of two convolutional layers with 32, 64, 128, or 256 \((3 \times 3 \times 3)\) kernels. In contrast to the original ResNet34 [44], which contains a strided convolutional layer as first layer in every ResNet-block, our network contains a pooling layer before the first and second ResNet-block, which is an average pooling layer with a size and stride of \(2 \times 2 \times 2\) voxels. After the four ResNet-blocks, the network has two output heads: one for regression of displacement vectors, and another for classification of landmark presence. Both output heads are similar in design. Each head has two 256-node dense layers and an output-layer, implemented as \(1 \times 1 \times 1\) convolutions [45]. The classification head outputs scalars, one for each landmark, forced between 0 and 1 by a sigmoid function. The regression head outputs displacement vectors for each landmark.

Since a network is fully convolutional it can analyze input images of varying size. Depending on its input image the network outputs a varying number of displacement vectors and posterior probabilities during global landmark localization. Due to the network's average pooling layers and the first convolutional layer with a stride of two voxels, its outputs are distributed on a grid, where the grid spacing is defined by the sum of the number of pooling layers and strided convolutional
layers. With $n$ representing the sum of the number of pooling layers and strided convolutional layers for the global or local localization step, this leads to a down-sampling rate of $1/2^n$ and therefore, a patch size of $2^n$ voxels. Hence, for a given network, a grid with a grid spacing of $2^n$ voxels is used to sample patches from an input image.

III. DATA

We evaluated the method on three different datasets containing 3D CCTA scans, 3D olfactory MR scans, and 2D cephalometric X-rays (Fig. 3). The choice of these datasets was based on the diversity in image acquisition modality (CT, MR, and X-ray), image dimensionality (2D and 3D), and anatomical coverage (cardiac, brain, and head).

A. Coronary CT Angiography

The dataset consisted of 672 CCTA scans, which were acquired in the University Medical Center Utrecht (Utrecht, The Netherlands) as part of regular patient care. The need for informed consent was waived by the Institutional Medical Ethical Review Board. ECG-triggered scans were acquired with a 256-detector row scanner (Philips Brilliance iCT, Philips Medical, Best, The Netherlands). Tube voltage ranged from 80 to 140 kVp while tube current ranged from 210 to 300 mAs. Intravenous contrast was administered before acquisition. All acquired scans had a slice thickness of 0.9 mm with 0.45 mm spacing. In-plane resolution ranged between 0.29 and 0.49 mm. In all scans, an expert manually annotated eight clinically relevant cardiac landmarks: the aortic valve commissures between the non-coronary and right (NCRC), the non-coronary and left (NCLC), and the left and right aortic valve leaflets (LRC), the hinge points (most caudal attachments) of the left (LH), non-coronary (NCH), and right (RH) aortic valve leaflets, and the right (RO) and left coronary ostium (LO). These landmarks can be used to perform clinical measurements in patients undergoing transcatheter aortic valve implantation (TAVI) [11], [14]–[16]. An expert observer created manual annotations for the full dataset using the protocol by Kasel et al. [11]. To determine inter-observer variability, a second observer annotated 100 randomly selected scans from the test-set. This same set was used to determine the intra-observer variability: after one month, the first observer repeated annotations in the 100 scans. Variability was defined as the Euclidean distance between the landmark annotations.

B. Olfactory MR

The dataset contained 61 olfactory MR scans, which were acquired as part of clinical routine in Hospital Gelderse Vallei (Ede, The Netherlands). The local ethics committee approved the study where informed consent was obtained from all subjects. Scans were acquired with a 3T Magnetom Verio MRI scanner (Siemens, Erlangen, Germany). To visualize the olfactory bulbs, a coronal T2-weighted fast spin-echo sequence was performed (echo time: 153 ms, repetition time: 4630 ms, field of view: 120×120 mm). Each scan contained 28 coronal slices reconstructed to an isotropic in-plane pixel size of 0.47 mm and a slice thickness of 1 mm with no inter-slice gap. In each scan, an expert manually delineated the right and left olfactory bulb. The center of each manual delineation was taken as ground truth landmark location.

C. Cephalometric X-rays

The dataset consisted of 400 publicly available cephalometric X-rays (lateral cephalograms) from the ISBI 2015 Grand Challenge in Automatic Detection and Analysis for Diagnosis in Cephalometric X-ray Images [8]. X-rays were acquired with
a Soredex CRANEX® Excel Ceph machine (Tuusula, Finland) and Soredex SorCom software (3.1.5, version 2.0), and were obtained in TIFF format with a resolution of 1935 × 2400 pixels and an isotropic pixel size of 0.1 mm. In all X-rays, two experienced medical doctors both manually annotated 19 clinically relevant landmarks, which can be used for diagnosis and treatment planning in orthodontic patients [8], [10]. Following the challenge protocol, the average of the annotations provided by both experts was used as ground truth landmark location [3]. The landmarks were: the sella (L1), nasion (L2) orbitale (L3), porion (L4), subspinales (L5), supramentale (L6), pogion (L7), menton (L8), gnathion (L9), gonion (L10), lower incisal incision (L11) upper incisal incision (L12), upper lip (L13), lower lip (L14), subnasale (L15), soft tissue pogion (L16), posterior nasal spine (L17), anterior nasal spine (L18), and articulate (L19). The intra-observer and interobserver variability were determined within the challenge following Lindner et al. [26].

IV. Evaluation

Evaluation was performed by computing the median Euclidean distance and interquartile range (IQR) between manually defined reference and automatically predicted landmark locations.

In addition, following the ISBI 2015 Grand Challenge in Automatic Detection and Analysis for Diagnosis in Cephalometric X-ray Images [8], success detection rates (SDRs) were calculated. The detection of a landmark is considered successful when the Euclidean distance between the automatically localized landmark and its reference location is smaller than a predefined distance threshold. In our analysis we used 10 distance thresholds ranging from 0.5 mm to 5 mm. The maximal distance threshold was defined when 95% of the landmarks were successfully detected. Intra- and second observer SDRs were determined in a similar way by using the two annotated sets. When two annotations of a landmark were within the distance of the set threshold, the annotation was considered successful. Fig. 6 shows the SDRs obtained with the proposed method, while the intra-observer and second observer SDRs are shown as horizontal lines. Overall, the SDRs obtained with the method are similar or better than intra-observer and second-observer SDRs.

Fig. 5 shows vector fields visualizing the predicted displacement vectors for localization of the RO landmark in the axial viewing plane in six CCTA scans from the test set: three scans in which the localization error was below 2.5 mm (top row) and three scans in which the localization error was above 5.0 mm (bottom row). Larger errors were made in scans in which anatomical deviation was present, such as both coronary ostia being located on the left side in close proximity to each other (Fig. 5 bottom row). This anatomical deviation occurred in only 0.2% of the scans in the training set but in 2% of the scans in the test set, which might explain the error.

2) Olfactory MR: The available set of 61 olfactory MR scans was randomly divided into 36 training, 5 validation and 20 test scans. Scans were resampled to an isotropic voxel size of 0.47 mm³. Training settings were similar to those used in the CCTA experiment, described in Section V-A1. However, because scans contained only 60 coronal slices after resizing, scans were zero-padded in the z-direction. The size of the olfactory bulbs ranged between 1.4 and 5.2 mm in-plane, and between 5.0 and 13.0 mm in the z-direction. The median (IQR) Euclidean distance error between computed landmark locations and reference locations was 0.87 (1.36) mm and 0.90 (0.58) mm for the right and left bulb, respectively, and 0.90 (0.85) mm when taking both landmarks into account. Fig. 6 shows the SDRs obtained with the proposed model was used for subsequent analysis. A local FCNN was trained using similar settings. However, the size of the sub-images was chosen based on the distance errors obtained during localization of landmarks with the multi-landmark network in the validation set and was therefore set to 16 × 16 × 16 voxels. Moreover, sub-images were randomly sampled such that they always contained the landmark of interest.
Table I
Median (IQR) Euclidean distance errors (mm) between computed and reference landmark locations in CCTA scans obtained with networks used for multi-landmark localization. Different training settings are evaluated: regression of displacement vectors with (R_{log}) and without (R) log-transformation, employing the classification layer (C) or not, and performing only global localization (Global) or global-to-local localization (Global-to-local) of landmarks. The proposed method combined global-to-local landmark localization and employed regression of log-transformed displacement vectors and the classification output layer (Proposed ML). Additionally, distance errors obtained with the method adjusted for single landmark localization are listed as well (Proposed SL). The distance errors obtained by the intra-observer (Intra-observer) and the second observer (Second observer), computed as the distance between two annotations made by the same observer, and the distance between annotations made by two different observers, respectively, on a subset of the test set are also listed. Results are listed per landmark (NCRC, NCLC, LRC, LH, NCH, RH, RO, and LO) and for all landmarks together (All), with the smallest distance error shown in bold.

|             | Aortic valve commissures | Aortic valve hinges | Coronary ostia | All     |
|-------------|--------------------------|---------------------|----------------|---------|
|             | NRC | NCLC | LRC | LH | NCH | RH | RO | LO | All  |
| Intra-observer | 2.68 (2.25) | 1.93 (1.72) | 1.96 (2.07) | 2.04 (1.39)* | 2.54 (2.20) | 2.56 (2.28) | 1.43 (1.05) | 1.88 (1.40)* | 2.06 (1.84)* |
| Second observer | 3.00 (1.23)* | 3.46 (1.45)* | 2.96 (1.13)* | 1.73 (1.32)* | 1.96 (1.19) | 1.80 (1.62) | 1.78 (1.54)* | 2.31 (1.56)* | 2.50 (1.68)* |

Global
- R 3.20 (2.05)* 2.95 (1.95)* 3.08 (2.47)* 3.14 (2.10) 3.51 (2.04) 3.55 (2.14) 4.36 (2.92) 3.92 (2.71) 3.44 (2.36)
- R_{log} 3.12 (2.06)* 3.12 (2.02)* 3.04 (2.09) 2.49 (1.84) 3.36 (2.18) 3.18 (1.71) 4.08 (2.87) 4.05 (2.66) 3.23 (2.30)
- C 5.64 (4.13) 4.77 (2.10) 4.62 (2.37) 4.43 (2.23) 4.69 (3.01) 4.34 (2.73) 5.07 (2.67) 4.64 (2.61) 4.76 (2.72)
- R + C 2.93 (1.96) 2.60 (1.53) 2.62 (2.54) 2.73 (1.54) 3.31 (1.92) 3.14 (1.66) 4.23 (2.94) 3.72 (2.78) 3.09 (2.11)
- R_{log} + C 2.72 (1.71) 2.59 (1.84) 2.60 (1.98) 2.51 (1.74) 3.08 (1.79) 2.77 (1.70) 2.90 (1.93) 3.31 (2.23) 2.81 (1.88)

Global-to-local
- Proposed SL 1.99 (1.82) 1.94 (1.32) 1.69 (1.73) 2.29 (1.46) 2.68 (1.75) 3.09 (1.82) 1.48 (0.98) 1.55 (1.02) 2.03 (1.79)
- Proposed ML 1.85 (1.96) 1.80 (1.59) 1.76 (1.67) 2.40 (1.58) 2.48 (1.72) 2.23 (1.42) 1.45 (1.20) 1.55 (0.97) 1.87 (1.67)

Significance outcome by the Wilcoxon Signed Rank test compared to R_{log} + C for the upper part of the table, and the proposed method for multi-landmark localization (Proposed ML) and the Intra-observer, Second observer, and proposed method for single landmark localization (Proposed SL) is indicated with * for p<0.05, † for p<0.01, and ‡ for p<0.001.

Fig. 4. Success detection rates (SDRs) for landmark localization in CCTA scans. The results for the ablation study on the performance of global multi-landmark localization (Global) are shown for: FCNNs trained for one task, i.e. regression of displacement vectors (R), regression of a log-transformed displacement vectors (R_{log}), and classification of patches (C); FCNNs trained for joint regression and classification (R + C) and joint regression of log-transformed displacement and classification (R_{log} + C). Furthermore, the results for our proposed global-to-local FCNNs (Global-to-local) trained for single landmark localization (Proposed SL) and multi-landmark localization (Proposed ML) are also shown. Additionally, intra-observer and second observer SDRs are indicated by horizontal lines. Results are shown as % over all landmarks.
X-ray images before analysis. Since the X-ray is a 2D image,
the processing time per scan was on average 0.07 ± 0.009
seconds.

The 150 cephalometric X-rays were used for training (140 X-rays) and
validation (10 X-rays). The challenge provides two separate
test sets for evaluation: one set containing 150 images (Test1),
and one set containing 100 images (Test2). To mitigate varying
image contrast, histogram equalization was performed on the
images before analysis. Since the X-ray is a 2D image,
we used a 2D version of the network in Fig. 2. Furthermore,
because cephalometric X-rays are large (1935 × 2400 pixels),
we also added an average pooling layer before the third and
fourth ResNet-block. Adding average pooling layers allowed us
to enlarge the receptive field, while keeping a low computational
complexity. The network for global localization was again
trained for 300,000 iterations, using mini-batches containing
4 sub-images of 592 × 592 pixels. For local analysis, mini-
batches containing 4 sub-images of 16 × 16 pixels were used
during training.

Table II lists the median Euclidean distance errors obtained
with the proposed method. Errors range from 0.46 to 2.12
mm for different landmarks in Test1 and from 0.42 to 4.32
mm for different landmarks in Test2. For both test sets, the
best results were obtained for the localization of L12, which
is the upper incisal incision. As reported by Lindner et al.
[26], the mean intra-observer variability for the first and
second observer were 1.73 ± 1.35 mm and 0.90 ± 0.89 mm,
respectively, while the mean inter-observer variability was
1.58 ± 1.55 mm. When computing the mean distance error
obtained on all landmarks present in both test sets, we obtain
a distance error of 1.35 ± 1.19 mm, which is lower than the
intra-observer variability of the first observer and the inter-
observer variability. As defined by the challenge protocol[8],
we evaluated our method computing the SDRs using four
distance thresholds (2 mm, 2.5 mm, 3 mm, and 4 mm). These
results are shown in Fig. 7. On average, the processing time
per scan was 0.05 ± 0.009 seconds.

B. Ablation Study

To investigate whether the application of classification or the
log-transform during training is truly beneficial for accurate
landmark localization, we performed an ablation study with
CCTA scans only, assuming results generalize to other datasets.
For this, four additional networks for global multi-landmark
localization were trained. These networks were trained with
or without applying the log-transform with regression, and
with or without using the classification output layer. For the
classification-only network, a final landmark location was
obtained by computing a weighted average of all predicted
landmark locations. To obtain the final landmark location, the
centers of the analyzed patches served as predicted landmark
locations, while the posterior classification probabilities were
used as weights during averaging.

Table II shows that the proposed approach utilizing joint
classification and regression of log-transformed displacement
vectors achieved best performance (Rlog + C). Regression-
only networks obtained smaller distance errors compared
to classification-only networks. The addition of classification
improved both regression-only networks, one using log-
transform and one without it. The log-transform improved
localization performance in the networks performing regression,
with and without classification. When the approach for global
localization utilizing joint classification and regression of
log-transformed displacement vectors is combined with local
single landmark localization (Table II, Proposed ML), smaller
distance errors were obtained compared to utilizing only global

method. When a distance threshold of 4 mm was used, 95.0% of
all landmarks present in the test set were successfully detected.
The processing time per scan was on average 0.07 ± 0.01
seconds.

3) Cephalometric X-rays: The 150 cephalometric X-rays
from the training set of the ISBI 2015 Grand Challenge in
automatic Detection and Analysis for Diagnosis in Cephalomet-
ric X-ray Images[8] were used for training (140 X-rays) and
validation (10 X-rays). The challenge provides two separate
test sets for evaluation: one set containing 150 images (Test1),
and one set containing 100 images (Test2). To mitigate varying
image contrast, histogram equalization was performed on the
X-ray images before analysis. Since the X-ray is a 2D image,
localization. Fig. 3 shows the obtained SDRs. Better SDRs were obtained by networks performing joint classification and regression compared to regression-only and classification-only networks. However, the best results were obtained when joint classification and regression of log-transformed displacement vectors were used with a global-to-local approach.

C. Single Landmark Localization

The proposed method employing joint classification and regression of log-transformed displacement vectors was evaluated for single landmark localization in CCTA by training one network for each of the eight cardiac landmarks. Table I lists the obtained Euclidean distance errors (Proposed SL). With the exception of localization of the LRC and LH, the network trained for multi-landmark localization outperformed networks trained for single landmark localization. However, differences in performance were only significant for localization of the NCH and RH. Fig. 3 shows the obtained SDRs. For a distance threshold of 0.5 mm, the SDR obtained with networks trained for single landmark localization was slightly better than the SDR obtained with a network trained for multi-landmark localization. However, this difference was only 0.6%. For all other distance thresholds, better SDRs were obtained by the network trained for multi-landmark localization compared to networks trained for single landmark localization.

D. Comparison with State-of-the-art

A number of methods have previously been proposed to localize anatomical landmarks in medical images.

1) Coronary CT Angiography: Previous methods have been proposed to specifically localize the aortic valve hinges, the aortic valve commissures, and the coronary ostia in cardiac CT scans.

Waecher et al. [15] used pattern matching and reported distance errors of 1.0 ± 0.8 mm and 1.2 ± 0.6 mm for the right and left ostium, respectively. Wolterink et al. [34] used a CNN to localize the ostia and obtained a mean distance error of 1.8 ± 1.0 mm. However, both methods were tested on small sets containing only 20 [15] or 36 [34] scans that might not have contained the anatomical deviation which was present in our test set comprising of 200 CCTA scans. Removing eight scans from our test set that show severe anatomical deviation (the right ostium located on the left side, the left ostium located on the right side), or cases where a stent is present in the pulmonary arteries, improves results for localization of the coronary ostia from 2.03 ± 3.05 mm to 1.75 ± 1.84 mm.

For the aortic valve commissures and the aortic valve hinges, the proposed method obtained mean distance errors of 2.33 ± 1.90 mm and 2.61 ± 1.44 mm, respectively. Zheng et al. [16] exploited voxel classification with landmark specific probabilistic boosting trees and reported mean distance errors of 2.17 ± 1.31 mm, 2.09 ± 1.18 mm, and 2.07 ± 1.53 mm for the aortic valve commissures, the aortic valve hinges, and the coronary ostia, respectively. Landmarks were localized in in C-arm CT scans with a voxel size ranging between 0.70 and 0.84 mm.

Elattar et al. [14] applied a local rule-based approach and combined results obtained for localization of the aortic valve hinges and the coronary ostia in 40 CCTA scans with a voxel size varying from 0.44 to 0.9 mm. The analysis led to a mean distance error of 2.81 ± 2.08 mm. Al et al. [9] also combined results obtained for localization of all eight landmarks in 71 CCTA scans using cross-validation and obtained a mean distance error of 2.04 ± 1.11 mm. Voxel sizes of used CCTA scans were not reported. Computing the same measure, the mean distance error obtained for localization of the eight cardiac landmarks, we obtained a distance error of 2.36 ± 1.24 mm.

These aforementioned methods have been developed and tested on different CCTA datasets than used in our work. Hence, a comparison between the results should only be used as an indication of the performance. To enable a direct comparison of our methods with previous work, we have tested the very recently proposed methods by Alansary et al. [42], who employ reinforcement learning and localize a single landmark at the time, and Payer et al. [35], who employ heatmap regression to localize either a single landmark or multiple landmarks jointly, on our data. For this, we used code made publicly available by the authors1,2. The results are listed in Table III. The Wilcoxon Signed Rank test was used to test for significance. Results show that our method performing single landmark localization significantly outperformed the method proposed by Alansary et al. [42]. Furthermore, we significantly outperform the method proposed by Payer et al. [35] when trained for single landmark localization. When comparing our method for multi-landmark localization with the multi-landmark localization proposed by Payer et al. [35], differences in performance are not significant.

On average, the processing time per scan was 0.29 ± 0.56 seconds for the method proposed by Alansary et al. [42], and 0.42 ± 0.13 seconds and 0.49 ± 0.28 seconds for the method proposed by Payer et al. [35] for single landmark localization and multi-landmark localization, respectively. On average, the processing times per scan for our method were 0.04 ± 0.01 and 0.06 ± 0.05 seconds for single landmark localization and multi-landmark localization, respectively.

2) Olfactory MR: To the best of our knowledge, no landmark localization methods have been evaluated for localization of the olfactory bulbs in MRI. To compare our proposed method with state-of-the-art landmark localization methods, we have evaluated the publicly available methods by Alansary et al. [42] and Payer et al. [35] as for the landmarks in CCTA (see section V-D1). Results are listed in Table IV. Comparing our method for localization of one olfactory bulb per scan, we significantly outperform other methods performing single landmark localization. The difference between methods performing localization of both olfactory bulbs simultaneously was not significant. On average, the processing time per scan was 0.78 ± 1.53 seconds for the method proposed by Alansary et al. [42], and 0.38 ± 0.003 seconds and 0.44 ± 0.25 seconds for the method proposed by Payer et al. [35] for single landmark localization and multi-landmark localization, respectively. For 1https://github.com/amiralanasy
2https://www.github.com/christianpayer
the proposed method, the processing time per scan for single landmark localization and multi-landmark localization were 0.07 ± 0.008 and 0.07 ± 0.01 seconds per scan, respectively.

3) Cephalometric X-rays: Previous methods have been proposed to localize landmarks in cephalometric X-rays. Ibragimov et al. [18], Lindner et al. [26], and Urschler et al. [22] all employed conventional machine learning, while Arik et al. [32] and Payer et al. [35] both proposed a CNN to localize landmarks in cephalometric X-rays.

Fig. 7 shows a comparison between the SDRs obtained in previous studies and the SDRs obtained with the proposed method. Payer et al. [35] reported the percentage of outliers. Hence, for comparison with our results, we reformulated their results into SDRs. For all distance thresholds, our method obtained better SDRs compared to those obtained in previous studies.

Reported processing times for the method proposed by Lindner et al. [26], Urschler et al. [22], and Payer et al. [35] were 5, 56, and 2 seconds per scan, respectively. However, for the proposed method, the processing time for localization of all landmarks was on average 0.05 ± 0.009 seconds per scan.

VI. DISCUSSION

An automatic method for anatomical landmark localization in medical images has been proposed. The method employs global-to-local analysis where initially a fully convolutional neural network predicts the locations of multiple landmarks simultaneously. Subsequently, specialized FCNNs refine the global landmark locations. For global multi-landmark localization, an FCNN analyzes 2D or 3D images of arbitrary size in a patch-based manner. For every patch in an image, regression is used to predict displacement vectors that point from the center of the patch to landmarks of interest. Simultaneously, classification is used to predict the presence of landmarks of interest.
computing a weighted average of the predicted landmark
classification-only networks, landmarks were localized by
in mm. A higher resolution will probably lead to lower distance errors
addressing the hardware limitations and analyzing scans with
increasing the image resolution had a negative impact on
limitations prohibited analysis of complete images at the native
reduce the image resolution. It is worth noting that memory
comparisons between the results should only be used as an
interest in each image patch. The global landmark locations
obtained by a weighted average of the displacement vectors
predicted by regression, using posterior probabilities predicted
by classification as weights. Subsequently, specialized FCNNs
refine the global landmark locations by analyzing a local sub-
image around each landmark in a similar manner, performing
regression and classification simultaneously and combining the
results.

The method was evaluated using three different datasets,
namely 3D CCTA scans, 3D olfactory MR scans and 2D
cephalometric X-rays. Results demonstrate that the method
is able to localize landmarks with high accuracy in medical
images differing in modality, dimensionality and depicted
anatomy. Results obtained for the localization of the aortic valve
commissures, the aortic valve hinges, and the coronary ostia
in CCTA are comparable with the intra-observer variability
and second observer performance. Previous methods analyzed
images at higher resolution and reported slightly better results.
However, these methods have been developed and tested on
different CCTA datasets than used in our work and therefore,
a comparison between the results should only be used as an
indication of the performance. Ideally, landmarks would be
localized in the native resolution. However, due to hardware
limitation, we have resampled the scans prior to analysis to
reduce the image resolution. It is worth noting that memory
limitations prohibited analysis of complete images at the native
resolution. Hence, hardware limitations require partitioning
of images during inference for analysis at a resolution close
to the native resolution. Our preliminary experiments showed
that increasing the image resolution had a negative impact on
the performance when analyzing complete images. However,
addressing the hardware limitations and analyzing scans with
a higher resolution will probably lead to lower distance errors
in mm.

Using the CCTA dataset, we have shown that joint regression
and classification improves upon classification-only. For the
classification-only networks, landmarks were localized by
computing a weighted average of the predicted landmark
locations. For this, the center of analyzed patches were consid-
ered the predicted landmark locations while the classification
output, i.e. posterior probabilities, served as weights during
averaging. A different approach for the final decision could
also be considered. For example, only the center of the
patch with the highest classification probability could have
been taken into account or only patches with a posterior
probability higher than a threshold could have been used.
However, the ability of the classification-only network to
precisely localize a landmark will always be limited by its patch-
size. A deeper neural network that could perform voxel-based
classification could therefore be more precise compared to a
patch-based classification network. Nevertheless, a voxel-based
classification network would demand balancing of the data
during training due to a high class imbalance between landmark
and background voxels. Furthermore, voxel-based analysis
is more computationally demanding compared to the here
proposed patch-based classification network. Since landmark
localization is typically a prerequisite for subsequent, more
complex medical image analysis [11–13], localization speed
may be important.

Additionally, we have also shown that joint regression
and classification improves upon regression-only landmark
localization. For regression-only methods, the final landmark
location was predicted by computing the average over all
landmark locations obtained with the predicted displacement
vectors. Hence, independent of their distance to the landmark,
all patches contributed equally. Inspection of the results showed
that predictions from patches far from the landmark of interest
resulted in larger distance errors compared to those made
from patches close to the landmark of interest. Hence, equally
weighting all predictions resulted in larger distance errors
compared to joint classification and regression. With joint
regression and classification, such errors were mitigated by
weighting the displacement vectors using the posterior probabil-
ities obtained from classification. Namely, patches farther from
the landmark of interest received lower posterior probabilities,
thereby reducing the influence on the final landmark prediction.
Employing a log-transform for displacement further improved
localization. This is likely caused by the nature of the log-
transform under the influence of the mean absolute error loss
during training; i.e. during training, prediction errors from
patches close to the landmark of interest are more heavily
penalized than predictions from patches far from the landmark
of interest.

Networks trained for multi-landmark localization obtained
slightly better results compared to networks trained for single
landmark localization in CCTA. However, no statistically
significant difference between the two approaches was found,
indicating that our proposed method could be used for single-
as well as for multi-landmark localization.

Visual inspection of the results obtained with our proposed
method on the CCTA dataset showed that larger Euclidean
distance errors were obtained in images in which anatomical
abnormalities were present, such as the right ostium being
located on the left side of the patient. Training the network with
more images that depict these types of anatomical deviation
or modeling of these anatomical deviations by exploiting

| Method          | Right Bulb | Left Bulb | Both Bulbs |
|-----------------|------------|-----------|------------|
| Alansary et al. | SL         | 1.28 (1.46)| 1.44 (1.42)| 1.41 (1.42) |
| Payer et al.    | SL         | 1.67 (1.48)| 1.38 (2.22)| 1.55 (2.00) |
| Proposed        | SL         | 0.93 (0.74)| 0.99 (0.94)| 0.95 (0.94) |
| Payer et al.    | ML         | 0.91 (1.06)| 0.76 (1.01)| 0.89 (0.97) |
| Proposed        | ML         | 0.87 (1.36)| 0.90 (0.58)| 0.90 (0.85) |

Significance outcome by the Wilcoxon Signed Rank test compared to
the proposed method is indicated with * for p<0.05 and † for p<0.01.
Comparisons are made between methods performing single landmark
localization (SL) or methods performing multi-landmark localization
(ML).
data augmentation, such as elastic transformations of images, could be beneficial to increase the variation in the dataset and ultimately improve localization.

In contrast to previous work that required segmentation of the aorta [13][14][16], the here proposed method does not require any prior segmentation steps. Moreover, during inference, the method analyzes complete images and thus it is capable of localizing target landmarks in large 3D image volumes with high speed. As demonstrated by the results, direct learning from the data without preprocessing steps incorporating knowledge about the anatomy leads to accurate localization of the eight cardiac landmarks. When comparing with recent methods [35], [42], our method outperforms results obtained by Alansary et al. [42] and Payer et al. [35] in the localization of single landmarks. Furthermore, our method performs on par compared to the method proposed by Payer et al. [35] for multi-landmark localization.

Contrary to earlier approaches, the here proposed method can be used for both single and multi-landmark localization. Furthermore, our method is able to localize landmarks faster compared to competing methods. For pre-operative applications or offline tasks, such as the initialization of segmentation methods [7], localization speed might be less important but for real-time applications, such as intra-operative applications, speed may be crucial [6]. For landmark localization in both test sets of the cephalometric X-ray challenge, our proposed method outperformed previous methods, having an error close to the variability between the two observers.

VII. CONCLUSION

We have shown that the proposed method is able to localize landmarks in 2D and 3D medical images of arbitrary size, acquired with three different imaging modalities and depicting different anatomical coverage. The method localizes multiple or single landmarks with high accuracy and speed, making it suitable for application in studies including a large number of images or real-time localization.

ACKNOWLEDGMENTS

This work is part of the research program Deep Learning for Medical Image Analysis with project number P15-26, financed by the Dutch Technology Foundation with contribution by Philips Healthcare.

REFERENCES

[1] K. Rohr, Landmark-based image analysis: using geometric and intensity models. Springer Science & Business Media, 2001, vol. 21.
[2] S. Miao, J. Lucas, and R. Liao, “Automatic pose initialization for accurate 2D/3D registration applied to abdominal aortic aneurysm endovascular repair,” in Medical Imaging 2012: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 8316. International Society for Optics and Photonics, 2012, p. 83160Q.
[3] K. Murphy, B. van Ginneken, S. Klein, M. Staring, B. J. de Hoop, M. A. Viergever, and J. P. Pluim, “Semi-automatic construction of reference standards for evaluation of image registration,” Medical image analysis, vol. 15, no. 1, pp. 71–84, 2011.
[4] Z.-R. Wang, B.-C. Yin, J. Du, C. Liu, X. Tao, and G. Hu, “Fast and robust detection of anatomical landmarks using cascaded 3D convolutional networks guided by linear square regression,” in Chinese Conference on Biometric Recognition. Springer, 2018, pp. 599–608.
[5] D. Han, Y. Gao, G. Wu, P.-T. Yap, and D. Shen, “Robust anatomical landmark detection for MR brain image registration,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2014, pp. 186–193.
[6] F. Alam, S. U. Rahman, S. Ullah, and K. Gulati, “Medical image registration in image guided surgery: issues, challenges and research opportunities,” Biomechanics and Biomedical Engineering, vol. 38, no. 1, pp. 71–89, 2018.
[7] O. Oktay, W. Bai, R. Guererro, M. Rajchl, A. de Marvao, D. P. O’Regan, S. A. Cook, M. P. Heinrich, B. Gloeker, D. Rueckert et al., “Stratified decision forests for accurate anatomical landmark localization in cardiac images,” IEEE Transactions on Medical Imaging, vol. 36, no. 1, pp. 332–342, 2017.
[8] C.-W. Wang, C.-T. Huang, J.-H. Lee, C.-H. Li, S.-W. Chang, M.-J. Siao, T.-M. Laii, 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.
[9] W. A. Al, H. Y. Jung, I. D. Yin, Y. Jang, H.-B. Park, and H.-J. Chang, “Automatic aortic valve landmark localization in coronary CT angiography using colonial walk,” PloS one, vol. 13, no. 7, p. e0200317, 2018.
[10] N. Torosdaghi, D. K. Liberton, P. Verma, M. Sincan, J. S. Lee, and U. Bagci, “Deep geodesic learning for segmentation and anatomical landmarking,” IEEE Transactions on Medical Imaging, vol. 38, no. 4, pp. 919–931, 2018.
[11] A. M. Kasel, S. Cassese, S. Bleiziffer, M. Amaki, R. T. Hahn, A. Kastrati, and P. P. Sengupta, “Standardized imaging for aortic annular sizing: implications for transcatheter valve selection,” JACC: Cardiovascular Imaging, vol. 6, no. 2, pp. 249–262, 2013.
[12] R. I. Ionasec, B. Georgescu, E. Gassner, S. Vogt, O. Kutter, M. Scheuering, N. Navab, and D. Comaniciu, “Dynamic model-driven quantitative and visual evaluation of the aortic valve from 4D CT,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2008, pp. 686–694.
[13] Y. Zheng, M. John, R. Liao, J. Boese, U. Kirschstein, B. Georgescu, S. Zhou, J. Kempfert, T. Walther, G. Brockmann et al., “Automatic aorta segmentation and valve landmark detection in C-arm CT: application to aortic valve implantation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2010, pp. 476–483.
[14] M. Elattar, E. Wiegierinck, F. van Kesteren, L. Dubois, N. Planken, E. Vanbavel, J. Baan, and H. Marquering, “Automatic aortic root landmark detection in CTA images for preprocedural planning of transcatheter aortic valve implantation,” The international journal of cardiovascular imaging, vol. 32, no. 3, pp. 501–511, 2016.
[15] I. Waechter, R. Kneser, G. Korosoglou, J. Peters, N. Bakker, R. vd Boomen, and J. Weese, “Patient specific models for planning and guidance of minimally invasive aortic valve implantation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2010, pp. 526–533.
[16] Y. Zheng, M. John, R. Liao, A. Nottling, J. Boese, J. Kempfert, T. Walther, G. Brockmann, and D. Comaniciu, “Automatic aorta segmentation and valve landmark detection in C-arm CT for transcatheter aortic valve implantation,” IEEE Transactions on Medical Imaging, vol. 31, no. 12, pp. 2307–2321, 2012.
[17] I. Igum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. Van Ginneken, “Multi-atlas-based segmentation with local decision fusion—application to cardiac and aortic segmentation in CT scans,” IEEE Transactions on Medical Imaging, vol. 28, no. 7, pp. 1000–1010, 2009.
[18] B. Ibragimov, B. Likar, F. Pernus, and T. Vrtovec, “Computerized cephalometry by game theory with shape-and appearance-based landmark refinement,” in Proc. ISBI Int. Symp. Biomed. Imag. 2015. Automatic Cephalometric X-Ray Landmark Detection Challenge, 2015, pp. 1–8.
[19] M. A. Dabbah, S. Murphy, H. Pello, R. Bourbon, E. Beveridge, S. Wiseman, D. Wyeth, and I. Poole, “Detection and location of 127 anatomical landmarks in diverse CT datasets,” in Medical Imaging 2014: Image Processing, vol. 9034. International Society for Optics and Photonics, Mar. 2014, p. 903415.
[20] D. Mahapatra, “Landmark detection in cardiac MRI using learned local image statistics,” in Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Oct. 2012, pp. 115–124.
[21] X. Lu and M.-P. Jolly, “Discriminative context modeling using auxiliary maps for LV landmark detection from single MR image,” in Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Oct. 2012, pp. 105–114.
[22] M. Urschler, T. Ebner, and D. Štern, “Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization,” Medical image analysis, vol. 43, pp. 23–36, 2018.

[23] R. Donner, B. H. Menze, H. Bischof, and G. Langs, “Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization,” Medical image analysis, vol. 17, no. 8, pp. 1304–1314, 2013.

[24] D. Štern, T. Ebner, and M. Urschler, “From local to global random regression forests: exploring anatomical landmark localization,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2016, pp. 221–229.

[25] Y. Gao and D. Shen, “Collaborative regression-based anatomical landmark detection,” Physics in Medicine & Biology, vol. 60, no. 24, p. 9377, 2015.

[26] C. Lindner and T. F. Cootes, “Fully automatic cephalometric evaluation using random forest regression-voting,” in IEEE International Symposium on Biomedical Imaging, 2015.

[27] S. K. Zhou, “Discriminative anatomy detection: Classification vs regression,” Pattern Recognition Letters, vol. 43, pp. 25–38, Jul. 2014.

[28] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” Medical image analysis, vol. 42, pp. 60–88, 2017.

[29] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, and D. Comaniciu, “3D deep learning for efficient and robust landmark detection in volumetric data,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2015, pp. 565–572.

[30] A. Q. O’Neil, A. Kascenas, J. Henry, D. Wyeth, M. Shepherd, E. Beveridge, L. Clunie, C. Sansom, E. Sedukyte Keith Muir, and I. Poole, “Attaining human-level performance with atlas location autocontext for anatomical landmark detection in 3D CT data,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 0–0.

[31] D. Yang, S. Zhang, Z. Yan, C. Tan, K. Li, and D. Metaxas, “Automated anatomical landmark detection on distal femur surface using convolutional neural network,” in Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on Biomedical Imaging. IEEE, 2015, pp. 17–21.

[32] S. Ö. Arik, B. Ibragimov, and L. Xing, “Fully automated quantitative cephalometry using convolutional neural networks,” Journal of Medical Imaging, vol. 4, no. 1, p. 014501, 2017.

[33] Z. Xu, Q. Huang, J. Park, M. Chen, D. Xu, D. Yang, D. Liu, and S. K. Zhou, “Supervised action classifier: Approaching landmark detection as image partitioning,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2017, pp. 338–346.

[34] J. M. Wolterink, R. W. van Hamersvelt, M. A. Viergever, T. Leiner, and I. Isgum, “Coronary artery centerline extraction in cardiac CT angiography using a CNN-based orientation classifier,” Medical image analysis, vol. 51, pp. 46–60, 2019.

[35] C. Payer, D. Štern, H. Bischof, and M. Urschler, “Integrating spatial configuration into heatmap regression based CNNs for landmark localization,” Medical Image Analysis, vol. 54, pp. 207–219, 2019.

[36] D. Zhang, J. Wang, J. H. Noble, and B. M. Davant, “Headlocnet: Deep convolutional neural networks for accurate classification and multi-landmark localization of head CTs,” Medical Image Analysis, vol. 61, p. 101659, 2020.

[37] M. I. Meyer, A. Galdran, A. M. Mendonça, and A. Campilho, “A pixel-wise distance regression approach for joint retinal optical disc and fovea detection,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 39–47.

[38] J. Zhang, M. Liu, and D. Shen, “Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks,” IEEE Transactions on Image Processing, vol. 26, no. 10, pp. 4753–4764, 2017.

[39] B. Auberth, C. Vazquez, T. Cresson, S. Parent, and J. D. Guise, “Automatic spine and pelvis detection in frontal X-rays using deep neural networks for patch displacement learning,” in Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on Biomedical Imaging. IEEE, 2016, pp. 1426–1429.

[40] Y. Li, A. Alansary, J. J. Cerrolaza, B. Kahanal, M. Sinclair, J. Matthew, C. Gupta, C. Knight, B. Kainz, and D. Rueckert, “Fast multiple landmark localisation using a patch-based iterative network,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 563–571.

[41] F.-C. Ghesu, B. Georgescu, Y. Zheng, S. Gribic, A. Maier, J. Homegger, and D. Comaniciu, “Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans,” IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 1, pp. 176–189, 2019.

[42] A. Alansary, O. Oktay, Y. Li, L. Le Folgoc, B. Hou, V. Vaillant, K. Kamnitas, A. Vlontzos, B. Glöcker, B. Kainz et al., “Evaluating reinforcement learning agents for anatomical landmark detection,” Medical image analysis, vol. 53, pp. 156–164, 2019.

[43] W. A. Al and I. D. Yun, “Partial policy-based reinforcement learning for anatomical landmark localization in 3d medical images,” IEEE transactions on medical imaging, 2019.

[44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[45] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[46] S. Ioffe and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” in Proceedings of the 32nd International Conference on Machine Learning. ser. Proceedings of Machine Learning Research, vol. 37. PMLR, 2015, pp. 448–456.

[47] D. Kingma and J. Ba, “Adam: a method for stochastic optimization,” ICLR, 2015.

[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.