One-shot Learning Landmarks Detection

Zihao Wang, Clair Vandersteen, Charles Raffaelli, Nicolas Guevara, Hervé Delingette

One-shot Learning Landmarks Detection

Wang Zihao[a,b,*], Vandersteen Clair[b,c], Raffaelli Charles[b,d], Guevara Nicolas[b], Delingette Hervé[a,b]

[a] *INRIA Sophia Antipolis Méditerranée, 2004 Route des Lucioles, 06902 Valbonne, FRANCE*
[b] *Université Côte d'Azur, 28 Avenue de Valrose, 06108 Nice, FRANCE*
[c] *Head and Neck University Institute, Nice University Hospital, 31 Avenue de Valombrose, 06100 Nice, FRANCE*
[d] *Department of Radiology, Nice University Hospital, 31 Avenue de Valombrose, 06100 Nice, FRANCE*

Abstract

Landmarks detection in a medical image is a mainstay for many clinical algorithms application. Learning-based landmarks detection is now a major successful methodology for many types of objects detection. However, learning-based approaches usually need a number of the annotated dataset for training the learning models. To tackle the lack of annotation issue, in this work, an automatic one-shot learning-based landmarks detection approach is proposed for identifying the landmarks in 3D volume images. A convolutional neural network-based iterative objects localization method in combine with a registration framework is applied for automatically target organ localization and landmarks matching. We investigated both qualitatively and quantitatively the performance of the proposed approach on clinical temporal bone CT volumes. The result shows that the proposed method is robust in convergence, effective in accuracy and reliable for clinical usage.

1. Introduction

Landmarks detection for target object localization plays a pivotal role in many imaging tasks. Automatic landmark detection can reduce the difficulty of annotation for human experts and save time for many image processing tasks such as the segmentation, registration, calibration etc. The difficult of landmarks detection in clinical images come from the imaging modality or body positions variability which leads to the different appearance of the imaging of the object.

The recent advance of learning based landmark detection in many automatic detection tasks outperforms than traditional manual feature extraction based landmarks detection methods. In this section, we first give a brief survey of current researches of automatic landmarks detection. Without loss of generality, we roughly summary the automatic landmarks detection approaches into two different branches: learning based landmarks detection and non-learning based landmarks detection.

1.1. Non-Learning based landmarks detection

Many works have been proposed for skeleton parts landmarks detection before. Cheung and Hamarneh [3] proposed to augment the scale-invariant feature transform (SIFT) to arbitrary n dimensions (n-SIFT) for 3D-MRI volumes matching thus could be applied for pairs images landmarks matching. The n-SIFT algorithm is robust due to that the SIFT feature points have the translation, rotation, and scales invariant attribution, however, 3D SIFT features computing is quite costly and the computational complexity increase cubic times with the volume size. Wörz and Rohr [24] leverage parametric intensity models for image landmarks detection. A gradient direction model applied for 3D region-of-interest (ROI) selection for initialization of the detection area. Further, a Gaussian error function and rigid transformation were proposed for

⋆This is an example for title footnote coding.
*Corresponding author:
Email address: zihao.wang@inria.fr (Wang Zihao)
[1]This is author footnote for second author.

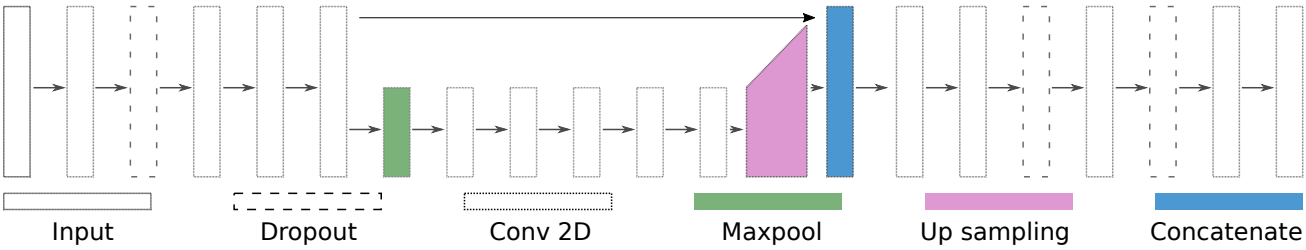

Figure 1: Neural network structure

modelling the target anatomical structures to get the landmarks of the target organ.

Literature research on the signal frequency domain for landmarks detection was proposed by Ferrari et al. [7]. The ideal is to use log-Gabor filters to extract the frequency features for constructing a 3D Phase Congruency (PC) measure. Then the 2-order PC moments of the frequency features were estimated for each of the feature points for the maxima detection. Their method was applied for head and neck landmarks detection. Another frequency analysis based method was proposed by Donner et al. [6]. applied random forest and Markov Random Field (MRF) for vertebral body landmarks detection. The methods consisted of the identification of a three-stage landmark: (1) geometric model fitting and random forest training based on the training dataset. (2) trained random forest for candidates landmarks classification and regression for prior generation. (3) Posterior result generation based on MRF and geometric priors. Their method was used for hand radiographs and full-body CT scans landmarks detection. Mothes and Denzler [16] proposed a one-shot SVM based landmarks tracking method for birds X-Ray 2D images landmarks detection. Their method relies on Histogram of oriented gradient (HOG) feature extractor for features points extraction.

1.2. Learning based landmarks detection

Probability models that can inference from parts prior are applied for bones landmark labelling were proposed in some prior works. Schmidt et al. [19] and Corso et al. [4] Both of those methods are based on a predefined population-mean prior of the parts and using probability models to personalize for a specific detection object. To introduce the spatial information of the landmarks, Potesil et al. [17] proposed to use joint spatial priors and parts based Graphical Model (GM) to improve the body organs landmarks detection accuracy. Two personalized priors (spatial and appearance) were learned from the training data. In the end, a Bayes classifier predicts the positive and negative landmarks labels. Hanaoka et al. [10] proposed a Bayesian inference based landmarks detection through construction a parametric stochastic landmark detector of the candidates. The detector was applied for spine landmarks detection. The drawback of their work is the inference time of the stochastic parametric model is not ideal since their model needs to use Markov chain Monte Carlo (MCMC) and simulated annealing algorithms for sampling the parameters which are quite a time consuming and need the burn-in for convergence.

Deep Learning-based landmarks detection algorithms are also attracting researchers due to the accuracy of identification and the generalization ability of neural networks. Suzani A. [21] proposed to train a convolutional neural network (CNN) with an annotated dataset for automatic vertebrae detection and localization. Liang et al. [15] proposed a two-step based residual neural network for inherent landmarks detection which the main idea is similar to the prior work except the network structure is optimized for landmarks detection. The reinforcement learning application for landmarks detection was investigated by Ghesu et al. [8]. In general, those agent-based landmarks localization methods treat the landmarks localization as a navigation problem and to train deep reinforcement agents who explore inside the anatomical volume space to find a state optimal trajectory for next step decision to approach the target landmarks. The main drawback of deep learning-based landmarks detection usually needs to have a massive dataset to train the neural networks. However, annotation dataset is expensive for collection and data lacking problem is quite com-

mon in any sense. To tackle this problem a deep learning-based landmarks detection method that can be used in limited medical image data circumstance. Zhang et al. [27] Their framework consists of two CNN: one used for regressing the patches and the target anatomical tags and the other one predict the coordinates of landmarks.

Nevertheless, those prior landmarks detection methods are still struggling in several aspects:(1) lack of annotation data. (2) small target landmarks detection. (3) hardware resource limitation for large scale volume.processing. When only a limit of annotation data available, or even worse, let's say only one case is annotated. Most of the prior methods will fail to generate reasonable results. The complexity of the problem is even higher when the target to be detected is smaller than the full searching domain. A typical circumstance is the cochlea landmarks detection where the human cochlea is a meticulous organ (width: 6.53 $\pm$0.35*mm*, height: 3.26 $\pm$0.24*mm* Zahara et al. [26].) Our research aims to tackle this difficulty.

1.3. Landmark Detection with Single Case

The proposed one-shot landmark detection has consisted of a CNN iteration based target centring model and a landmark matching model. First, the localization model will be centring the position of the target model through iterative target centring. In this stage will generate the FOI volume of the target organ. Second, the landmark matching model will generate the first matching of the landmarks through the rigid registration and then generate the precision landmarks through the inverse displacement vector field transform to get the final landmarks of the target image.

We organize the paper structure as follow. In the section 2 we present the implementation details of our work. In the section 3.3 we apply the proposed method on the cochlea landmarks detection problem to evaluate the landmark detection result. Section 4 discusses the effectiveness and limitation of the proposed method.

2. Method

In this section, we attempt to deal with the main concerns that most prior works failed to coup with. We first introduce in the subsection 2.2 the implementation details of the CNN based target organ interactive centring model which is realized through a 2D CNN trained on a single case of target segmentation. Further, in subsection 2.3 we introduce the details of landmarks detection model that consisted of moments of inertia rough alignment step, rigid registration step and none-rigid registration step. In each of those sub-steps, the inverse coordinates transformation was calculated to get the proper matching form the single annotated case.

2.1. Problem Definition

Given a collection of dataset $\{d_i \in D, i \in \mathcal{N}\}$, where d presents the volume image. For index $i = 1$ the volume d_1 is annotated with the segmentation $s \in d_1$ of the target volume d_1. Few landmarks points $m_i \in s_i$ inside the segmentation area s as coordinates that can be used for describing the spatial attribution of the target segmentation. The one-shot learning landmark detection task targets for learning a model $M(d_1, m_1)$ that only trained with one annotated paris: $d_1|m_1$. Fig:2 shows briefly the problem setting of landmark detection. Model M in most prior learning-based methods need to be trained on a large annotated dataset. In our case, we target to train model M that with only one annotated volume for training.

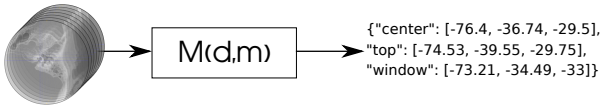

Figure 2: The landmark detection model M takes volumes as input and generate the corresponding landmarks information of target objects.

2.2. One-shot rough segmentation for target localization

Candidate area rough identification. Before applying the main method the raw volume needs to be preprocessed in order to reduce the hardware resources burden and ensure the robustness of the main algorithm. Directly processing the full 3D clinical volume is usually not amenable due to the memory limitation or algorithm computational complexity. Directly apply predictive models with a resampled volume is the reasonable way when the target object is relatively large enough. However, this is not practical for small objects. Another way is FOI detection, plenty of methods can be used for candidate FOI initialization. Deep learning approach can be an effective way once enough training data are collected. In our case,

we use a Bayesian model to get the candidate the FOI of the target object. Those candidate areas can be further selected by the one-shot trained CNN to get the final FOI. In our case, the target FOI appearance is modelling through a Gaussian mixture model. The components of the different class can be modelled by making each of the K components μ_k and Σ_k.

$$p_k(d_i|\theta_k) = \frac{1}{\sqrt{2\pi}^d |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(d_i-\mu_k)^t \Sigma_k^{-1}(d_i-\mu_k)}$$

The rough localization of the FOI can be realized with threshold the distribution histogram of the Gaussian mixture model.Greenspan et al. [9] For the cochlea area, the main components have consisted of dense bone that the average Hounsfield unit lies around 2000 HU. The target component is thus modeling by: $x \sim \mathcal{N}(\mu = 2000, \sigma = 50)$. We can then use the probability map to determinate the centre of rough FOI localization (x, y, z) with the mass centre of the probability map:

$$\begin{aligned} \bar{x} &= \frac{\iiint_{p_k(d|\theta_k)} x\rho(x,y,z)dXdYdZ}{\iiint \rho(x,y,z)dXdYdZ} \\ \bar{y} &= \frac{\iiint_{p_k(d|\theta_k)} y\rho(x,y,z)dXdYdZ}{\iiint \rho(x,y,z)dXdYdZ} \\ \bar{x} &= \frac{\iiint_{p_e dk(d|\theta_k)} z\rho(x,y,z)dXdYdZ}{\iiint \rho(x,y,z)dXdYdZ} \end{aligned} \tag{1}$$

where $\rho(x, y, z)$ represents the intensity of the volume, (x, y, z) are coordinates of the physical Cartesian space. All the volumes are processed to a proper size that is adaptable for the CNN for further processing.

CNN based iterative target localization. Learning-based approaches not limited to deep learning mostly have the risk of overfitting thus need at least hundreds of samples for training. In contrast, nonparametric methods that target to modelling the intrinsic properties of the task adapted well for one-shot learning tasks Wu et al. [23]. Learning with highly parameterized convolutional neural network from one volume image without any similar task knowledge transfer is not possible in our issue. The ideal that directly builds an end2end model with only one trainable sample as shown in Fig:2 is a mirage for our problem. However, it is still possible that realize the model $M(d_1, m_1)$ with only one image case by constructing a semi-learning framework.

Algorithm 1 Train a one-shot CNN

Inputs: image: d_1, d_v, segmentation: s_1, s_v
Output: CNN parameters ω
Initialize: f_{cnn}, R, T, Δ
while L_{val} *not converged* **do**
  $T \leftarrow N(0,1)\Delta_T$; `// Translation vector`
  $R \leftarrow N(0,1)\Delta_R$; `// Rotation matrix`
  $d|s \xleftarrow{R,T} d_1|s_1$; `// Random transform` d_1, s_1
  $K \leftarrow len(d)$; `// K is the slices number of d`
  for $i = 1;\ i < K;\ i++$ **do**
    `// Train the cnn with sliced pairs`
    $f_{cnn} \xleftarrow{\omega} d[i]|s[i]$
  end
  $L_{val} \leftarrow loss(s_v, f_{cnn}(d_v))$; `// Validation loss`
end

We firstly train a CNN f_{cnn} for learning the 'one-sided' segmentation from one paired case $d_1|s_1$. To let the CNN learn the stereo-invariant attribution of the volume image, we randomly rot and translate the volume in 3D space. The rotation matrix R and translation vector T is randomly computed through times the range control factors Δ_R, Δ_T respectively. The GPU memory requirement is fair large for 3D CNN to process the volume data. We thus train a 2D CNN with Binary Cross-Entropy (BCE) loss function. The detailed training procedures are described in algorithm 1.

The learnt CNN is not accurate enough for segmenting the target object, yet, the incomplete segmentation could provide ragged location information for navigation. To utilize the noisy location information we propose an iterative mass centre relocation method. Not, strictly speaking, the f_{cnn} can also be treated as a probability measure of target objects segmentation once the final layer is suitable with a proper activation function is chosen.

Algorithm 2 Iterative DL based mass center localization

Inputs: image: d, cnn: $f_{cnn}(\cdot)$
Output: Target object area centralized: d_{out}
Initialize: $ed_0, i, \epsilon \;\; dx, dy, dz \leftarrow d(x, y, z)$; `// slices of 3 dimensions`
while $|ed_i - ed_{i-1}| < \epsilon$ **do**
  $Kx, Ky, Kz \leftarrow len(dx), len(dy), len(dz)$ `// K. is the slices number in three dimensions`
  for $i = 1; i < Kx; i++$ **do**
    `// probability map of slice i in dim x`
    $p_{objX}[i] \xleftarrow{f_{cnn}} dx[i]$
  end
  for $i = 1; i < Ky; i++$ **do**
    $p_{objY}[i] \xleftarrow{f_{cnn}} dy[i]$
  end
  for $i = 1; i < Kz; i++$ **do**
    $p_{objZ}[i] \xleftarrow{f_{cnn}} dz[i]$
  end
  `// Joint probability map of 3 dimensions`
  $p_{obj} \leftarrow p_{objX} \cdot p_{objY} \cdot p_{objZ}$ `// Compute current mass center with Eq:1`
  $(\bar{x}, \bar{y}, \bar{z}) = mass\_center(p_{obj})$ `// Euclidean distance`
  $ed_i \leftarrow \sqrt{(\bar{x} - x)^2 + (\bar{y} - y)^2 + (\bar{z} - z)^2}$
  $x, y, z \leftarrow \bar{x}, \bar{y}, \bar{z};$ `// Center relocation`
  `// update slices with new center`
  $dx, dy, dz \leftarrow d(x, y, z)$
end
`// Crop image to get object with the center`
$d_{out} \leftarrow crop(d(x, y, z))$

Algorithm 2 presents the full procedures for centring of the target object. The ideal is the one-shot learnt CNN $f_{cnn}(\cdot)$ can only give a noisy, spoiled probability map of the target object. The noisy signals may come from similar anatomical structures or close intensities etc. The noisy output can be first filtered via computing the joint probability p_{obj} of the outputs $p_{objX} \cdot p_{objY} \cdot p_{objZ}$ from three different views. Further, the mass centring can filter again the center information since the noisy signals that far away from the high probability cluster of p_{objX} will have a small influence on the mass centre computing. Since in this iteration those noisy signal is uneven in distribution space, the current mass centre is still influenced by those uneven signals. To eliminate the influence of the noise, we can repeat the procedures until the noise signal are more evenly distributed in the FOI which means the FOI is centralized by the target object mass. The iteration stopped once the Euclidean distance between two iterations is converged to a certain threshold $|ed_i - ed_{i-1}| < \epsilon$. The FOI image will be focused on the target object once the above algorithm converges. We now get a target FOI volume which registration based landmark detection can be then applied to match the landmarks.

2.3. Landmarks matching

Moments of inertia alignment. The volume of FOI cropped similar to the target sample FOI can be roughly aligned by computing the moments of inertia of the two volumes to be amenable for registration. The moments of the inertia of the volume simulate the physical attribution of the objects which can be computed with Jaklic and Solina [12] Crisco and McGovern [5]

$$\begin{aligned} \tilde{x} &= \iiint_{p_k(d|\theta_k)} (y^2 + x^2)\rho(x, y, z)dXdYdZ \\ \tilde{y} &= \iiint_{p_k(d|\theta_k)} (x^2 + z^2)\rho(x, y, z)dXdYdZ \\ \tilde{z} &= \iiint_{p_k(d|\theta_k)} (x^2 + y^2)\rho(x, y, z)dXdYdZ \end{aligned} \tag{2}$$

where $\tilde{x}, \tilde{y}, \tilde{z}$ are moments of inertia for the yz plane, xz plane and zy plane respectively. We note the volume processed by the previous procedures as $\otimes(d_1)$, $\otimes(d_i)$ for annotated case and test case respectively. Since the FOI volume is cropped into an area volume that mainly occupied by the target object, the moments of inertia can give a sufficiently precise matching for registration between the being matched volume $\otimes(d_i)$ and the annotated volume $\otimes(d_1)$. The single annotated case $\otimes(d_1)$ contains the landmarks information which can be used for computing the landmarks information of $\otimes(d_i)$ from the inverse transform T^{-1} of $T : \otimes(d_i) \xrightarrow{T(H,D)} \otimes(d_1)$, where transform T is consisted by a homogeneous coordinates transform $\{H = \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix}, R \in SO(3), t \in \mathbb{R}\}$ and a diffeomorphism transform registration D given by diffeomorphism registration we will discuss in next two sub-sections.

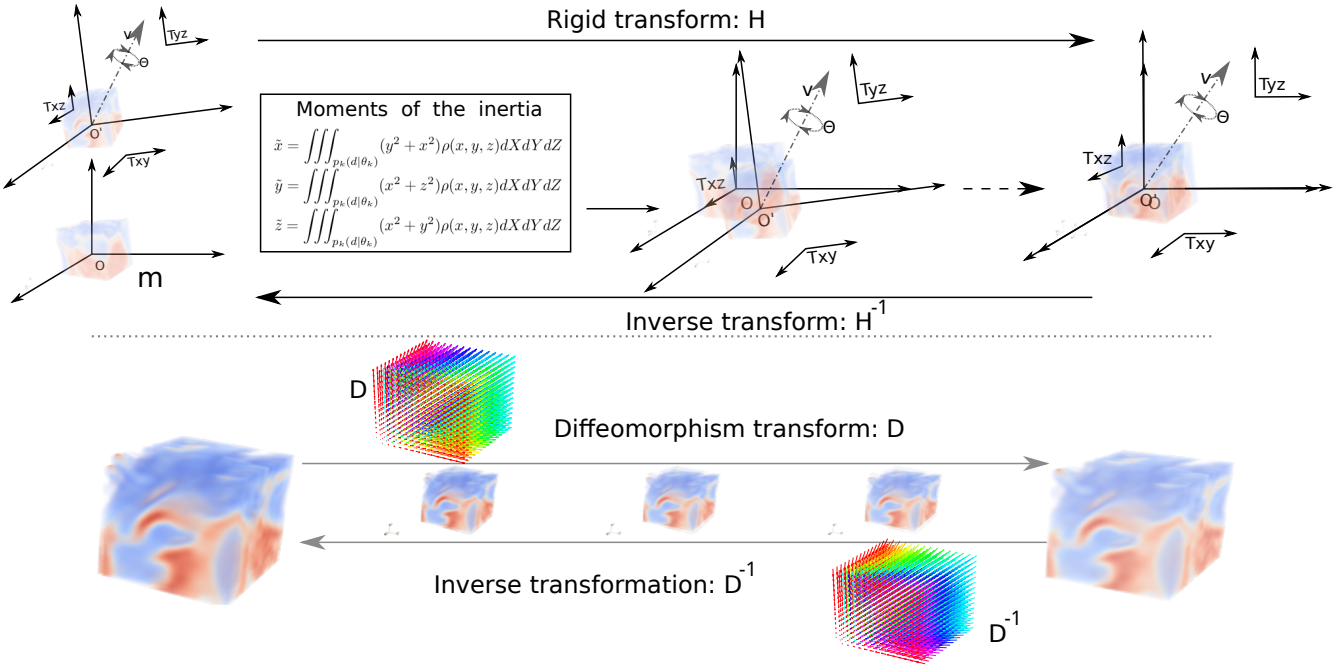

Figure 3: Landmarks matching based on inverse rigid and diffeomorphism transformation. Above subfigure shows the rigid transform H: (1) Compute the moments of the inertia of the two volumes. (2) Optimize the alignment position through optimizing the similarity measure metric. Below subfigure shows the diffeomorphism transform D: (1) Compute the diffeomorphism transformation D. (2) Compute the matched landmarks by inverse the displacement field D^{-1}.

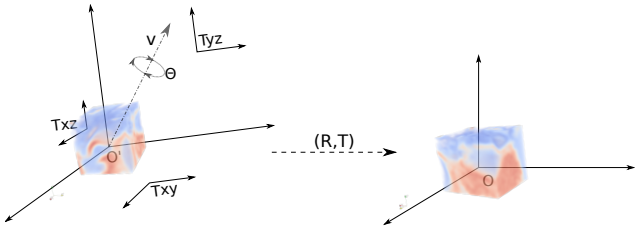

Figure 4: Rigid transform

Rigid transform matching. We note $\otimes(d_1)|_m$ as the aligned volume d_1 with landmarks m annotated. The aligned volume pair $\otimes(d_1)|_m$ and $\otimes(d_i)$ are first registered through the rigid transformation that the transform H can be got through computing the arguments of the maxima negative normalized cross-correlation Γ Yoo and Han [25] between the two volumes:

$$\arg\max_{H \in SE(3)} \Gamma(H) := \frac{\| \iiint_{xyz} (\otimes(d_1)|_m - \otimes(\bar{d}_1)|_m)(\otimes(d_i) - \otimes(\bar{d}_i))\|^2}{\iiint_{xyz} \|\otimes(d_1)|_m - \otimes(\bar{d}_1)|_m\|^2 \iiint_{xyz} \|\otimes(d_i) - \otimes(\bar{d}_i)\|^2} \tag{3}$$

We optimize Eq:7 with gradient descent algorithm that learning rate set as 0.99. The inverse transform : H^{-1} is thus given by: $H^{-1} = \begin{bmatrix} R^T & -R^T t \\ 0^T & 1 \end{bmatrix}$. The landmarks points m of $\otimes(d_1)|_m$ can be used for computing the landmarks m' of arbitrary object: $\otimes(d_i)|_{m'}$:

$$\begin{bmatrix} m' \\ 0 \end{bmatrix} = H^{-1} \cdot \begin{bmatrix} m \\ 0 \end{bmatrix} \tag{4}$$

Currently, we get the landmarks m' though computing an inverse transform of annotated landmarks m from a rigid transform H. Yet, when the accuracy requirements are relatively high the rigid transform-based landmarks matching might be not precise enough. In this circumstance, we

propose to apply diffeomorphism transform to get a more accuracy landmark matching.

Inverse displacement field mapping. We note the paired annotated volume and aligned volume after rigid transform as $\oslash(d_1)|_m$ and $\oslash(d_i)|_{m'}$ respectively. We can continue transform the $\oslash(d_i)|_{m'}$ with the B-spline transformation to match annotated volume $\oslash(d_1)|_m$.(Balci et al. [1] Kabus et al. [13]) Based on the trade-off between computational time and accuracy, a proper control points distance δ can be determined for creating a mesh with (n_x, n_y, n_z) control points for 3D volumes. For each control points with 3 directions freedom can be transformed with a displacement vector $\mathcal{V}$. We can construct a B-spline transformations mapping between for $\oslash(d_i)|$. in the control point $n : (x, y, z)$,

$$\Psi(\oslash(d_i; n)) = \iiint \mathcal{V}\psi_i(\oslash(d_i; x))\psi_j(\oslash(d_i; y)|)\psi_k(\oslash(d_i; z)) \tag{5}$$

where $\psi : \{\psi_k \in B - spline, k \in \mathbb{N}\}$ are a set of cubic B-spline functions. To maximize a suitable metric $\mathcal{M}$ which measure the similarity between the transformed volume $\Psi(\oslash(d_i; n))$ and the annotated volume $\oslash(d_1; n)$ we can get the final non-rigid transform D between the two volumes,

$$\arg\max_{\mathcal{V}\in D} \Gamma(H) := \mathcal{M}(\Psi(\oslash(d_i; n))\|(\oslash(d_1; n))) \tag{6}$$

We can then inverse the directions of transform vectors of D to get the inverse transform D^{-1} which can be applied similar to Eq:4 to get final landmarks $\hat{m}$ of $\oslash(d_i)$,

$$[\hat{m}, 0]^T = H^{-1} \cdot D^{-1}([{}^{m}_{0}]) \tag{7}$$

3. Experiment and Result

3.1. Network architecture and training details

Large scale CNN are easily trapped in overfitting for low data regimes (Vinyals et al. [22]). We use a 2D U-net like network (Ronneberger et al. [18]) for leaning the one-shot segmentation. The structures and layers consist is shown in Fig:1 where the network input size is $[\cdot, 100, 100, 1]$ with 4 convolutional layers (shape: $[\cdot, 100, 100, 64]$) following. Feature maps are then convoluted with a group of halved size layers but doubled in channels (shape: $[\cdot, 50, 50, 128]$). Up-sampling layer applied to recover the size of the feature maps to merged with the jump concatenates feature maps (shape:$[\cdot, 100, 100, 64 + 128]$). Finally, 5 convolutional layers (shape:$[\cdot, 100, 100, chn]$, $chn = 64$ for middle layers, $chn = 1$ for the last layer) process the feature maps for generating the analogous probability map. Some max-pooling layers dropout layers were applied for avoiding the overfitting and reducing the network size. Although we used the Unet in this experiment, we note that the structure of the CNN can be flexible by introducing more one-shot deep learning-based networks architectures design tricks, works on this direction could be found in the literature ofKoch [14], Vinyals et al. [22],Shaban et al. [20], Chen et al. [2], Jadon and Srinivasan [11]. An adam stochastic gradient optimizer with learning decay rate $lr = lr/epochs\_number$ initialized to $lr = 0.1$ is used for training the network with the binary cross entropy loss function. The neural network was implemented with Tensorflow 2.0 framework and trained on one NVIDIA 1080 Ti GPU.

3.2. Dataset

The dataset includes 200 3D volumes of temporal bones CT scan collected by a GE LightSpeed CT scanner at the Nice University Center Hospital. The dimensions of original volumes are $(512, 512, 160)$ in 3D with corresponding spacing of $(0.1875mm, 0.1875mm, 0.25mm)$. All the volumes are normalized based on the Eq: 8

$$\begin{cases} d_i(x,y,z) = \frac{d_i(x,y,z)-(-1024)}{3000-(-1024)} & d_i(x,y,z) \\ d_i(x,y,z) = 1 & d_i(x,y,z) > 1 \\ d_i(x,y,z) = 0 & d_i(x,y,z) < 0 \end{cases} \tag{8}$$

One volume was taken from the dataset and then segmented by an expert to get the volume for training the neural network, the landmarks of the cochlea were also marked. In our dataset we define the cochlea 3D landmarks as shown in Fig: 6 where three featured landmarks coordinates are defined. This training volume was also applied with the procedures of initialization localization (see steps (1)-(3) of Fig: 5) to crop into the same size as the neural network input data dimensions.

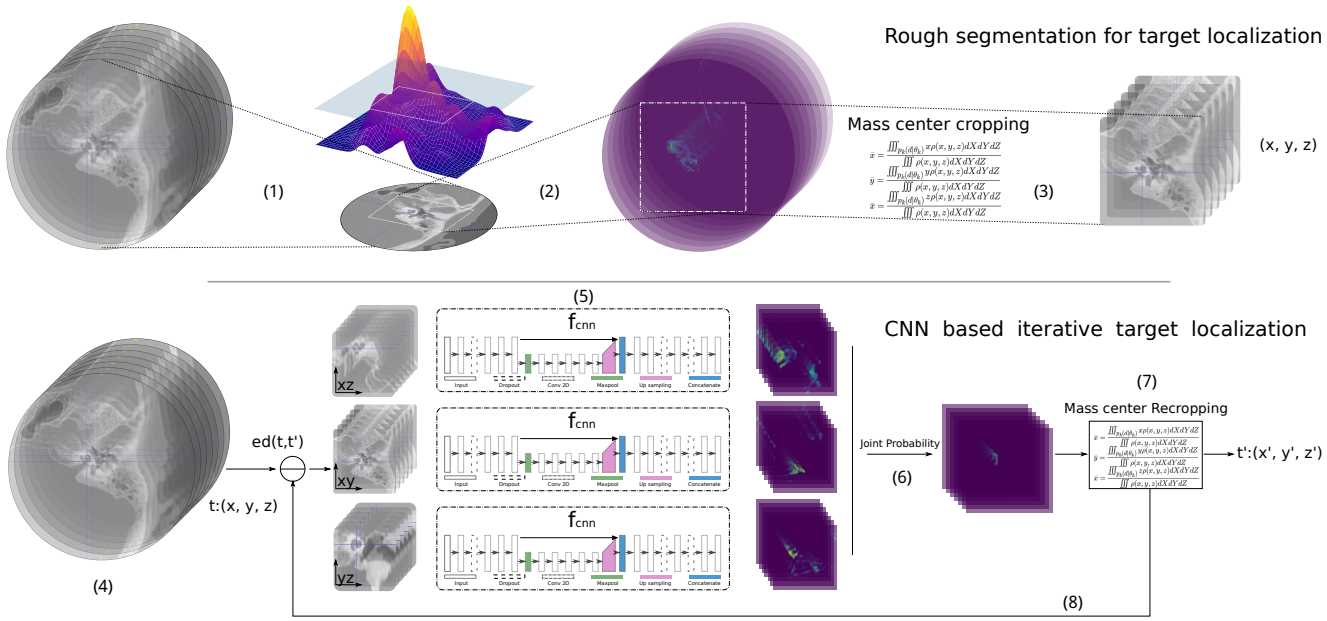

Figure 5: The schematics of the proposed one-shot interactive deep learning-based target localization. Steps (1)~(3) show the probability model-based area rough identification, the raw volume is cropped to a suitable size to be fed to CNN. Steps (4)~(5) show the CNN prediction of the high probability area of the target object from three different slicing views. (6): Computing the joint probability of the three views. (7) Get the centre of mass of the joint probability map. (8) Center of mass feedback for volumes cropping.

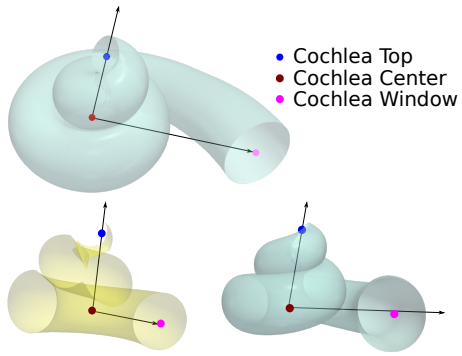

Figure 6: Cochlea landmarks shown with three import coordinates (cochlea top, cochlea center and cochlea round window points) of cochlea model.

3.3. Result

The result of the proposed approach is presented qualitatively and quantitatively. For qualitative evaluation, we show the coordinates of the detected target landmarks in three different views, each of the detection steps were shown for demonstration. To exam proposed method quantitative performance, we annotated another 20 volumes and then compute the mean square distance between the detected landmarks of automatic approach and the manual annotation.

3.3.1. Qualitative evaluation

Fig: 7 shows step-by-step outputs of landmarks detection for the cochlea volume. We see that the one-shot interactive localization for target centring converged with three steps (see sub-above figure (a) of 7). The coordinate positions is marked with cross marks in different colours represent different steps. We see that the centre is gradually localized to the target centre. The final landmarks positions are shown in column (b) for three different views. Columns (c)-(d) show the position detected by the steps we have demonstrated in Fig: 3. We see that the three landmarks are very approaching to the template landmarks as we defined in Fig: 6.

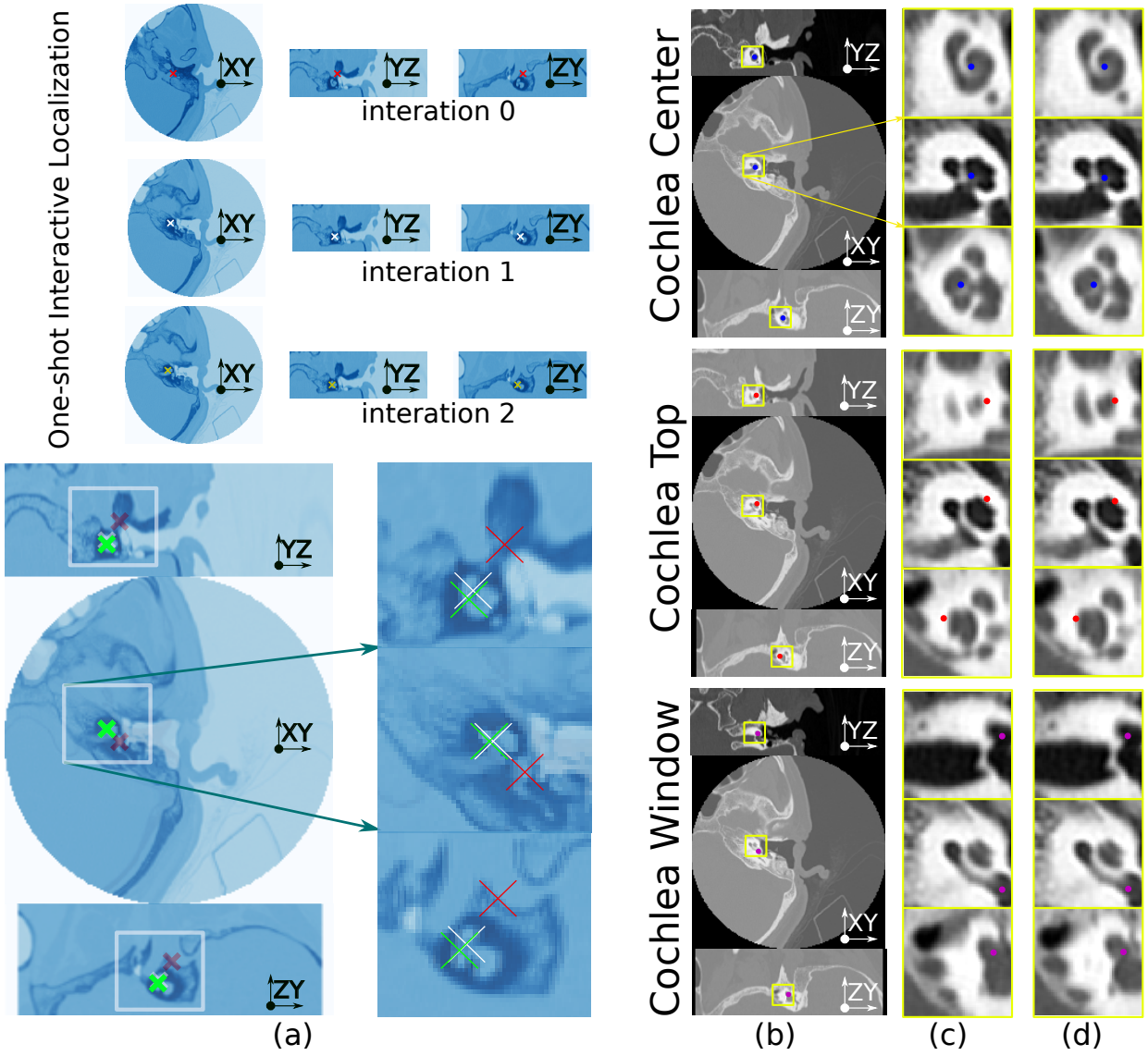

Figure 7: Qualitative evaluation of the proposed method. (a): Left sub-figure shows the convergence of the one-shot learning-based target volume localization. The above three sub-figures shows each step of the target localization: red, white, green corresponding to 1st, 2ed, 3rd respectively. The bottom sub-figure shows a transparent overlapping of those steps for interactively comparing the difference between the steps. (b) shows the result of the landmarks detection. (c) shows the amplified target area for the three landmarks position in different views of the rigid transformation. (d) shows the same but for diffeomorphism transform.

3.3.2. Quantitative evaluation

We compute the Euclidean distance between the expert annotated landmarks and proposed method detected landmarks.

Table 1: Euclidean distance for 20 volumes between manual and automatic detected landmarks

ID	CENTER	TOP	WINDOW
0	0.877104	0.702200	0.862706
1	0.281045	1.332490	0.648094
2	0.490064	0.564314	0.839740
3	0.695775	0.730758	0.554273
4	0.720785	0.720785	0.652519
5	0.565006	0.370034	1.117346
6	0.190016	0.309536	1.348332
7	0.488173	0.776744	0.309536
8	0.488857	0.348377	0.604708
9	0.385027	0.198213	0.494694
10	0.653242	1.632196	0.263451
11	0.837414	1.147350	1.064709
12	0.868734	1.002242	1.535283
13	0.673684	0.264812	0.718197
14	0.718201	1.037524	0.878216
15	0.722738	0.674657	0.812038
16	0.733301	0.797897	0.541678
17	0.536202	1.233090	0.335046
18	0.333654	0.549499	1.433840
19	0.389453	0.391017	0.880967
Mean	$0.58\pm0.19mm$	$0.73\pm0.39mm$	$0.79\pm0.36mm$

As shown in Tab: 1, the three landmarks (cochlea centre, cochlea top, cochlea round window) distance between human annotation and automatic detection across 20 volumes are average less than $0.8mm$. This result is quantitative satisfy by considering the spacing size of the detection target ($0.1875 \times 0.1875 \times 0.25$) and the relative target scale in comparison with the full CT volumes ($96mm \times 96mm \times 55mm$).

4. Discussion

We presented a novel one-shot learning approach for landmarks detection. The proposed method was validated with cochlea CT volumes of landmarks detection that annotated by three landmarks coordinates. We achieve landmarks detection with limited annotation dataset available (only one sample in our method) through integrating a neural network with an iterative localization and registration based landmarks matching framework.

The proposed method performance can be influenced significantly by the annotation accuracy of the template sample. There are two annotations to provide: (1) target object segmentation. (2) target landmarks coordinates. The former is fed to train the neural network for providing the gross location information to iterative localize the target, the later is provided to the role as the reference information for landmarks computing.

Moreover, it is worth to note that the uncertainty of the manual annotation is worth to be further studied. We see that the mean distance of the top landmarks and cochlea round window landmarks are relatively higher than the cochlea centre landmarks. This result might be due to the uncertainty of the manual annotation on YZ and ZY planes are higher than the XY plane. The uncertainty can come from the volume re-sampling error since the volumes are collected slice by slice along Z-axis (see subfigure (a) of Fig: 3), the human can be influenced by those errors when annotating the landmarks, or that error can come from the framework itself which could be ascribed to the error of the registration.

Although the proposed method is investigated to be robust enough with only single training data available, the performance of the method can be further improved by introducing more training data. This is because the accuracy of target object localization can influence the final detection accuracy through the centring of the inertia. A good localization initialization can improve the quality of alignment which is also important for landmarks matching steps.

5. Conclusion

To the best of our knowledge, our proposed method is the first one-shot learning approach for 3D volume landmarks detection, more specific for tiny object detection. we successfully solve the problem of 3D volumes landmarks detection by realizing a model as we showed in Fig: 2. Although the proposed method is not an end2end learning model, we can directly apply the framework for end2end landmarks detection once the neural network is well trained.

Quantitative and qualitative evaluation of the framework shows the effectiveness of the proposed method. We conclude that the proposed framework can be applied for

cochlea landmarks detection and could be potentially applied for other organs landmarks identification. The proposed method may be improved in future work by introducing more annotated data (few-shot learning) or modify the network architectures based on one-shot networks design technologies.

Acknowledgments

This work was partially funded by the regional council of Provence Alpes Côte d'Azur, by the French government through the UCA JEDI "Investments in the Future" project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01, and was supported by the grant AAP Santé 06 2017-260 DGA-DSH.

References

[1] Balci, S., Golland, P., Shenton, M., Wells, W., 2007. Free-form b-spline deformation model for groupwise registration. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention 10, 23–30. doi:10.1901/jaba.2007.10-23.

[2] Chen, Z., Fu, Y., Wang, Y.X., Ma, L., Liu, W., Hebert, M., 2019. Image deformation meta-networks for one-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[3] Cheung, W., Hamarneh, G., 2009. n-sift: n-dimensional scale invariant feature transform. Image Processing, IEEE Transactions on , 2012 – 2021doi:10.1109/TIP.2009.2024578.

[4] Corso, J.J., Alomari, R.S., Chaudhary, V., 2008. Lumbar disc localization and labeling with a probabilistic model on both pixel and object features, in: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 202–210.

[5] Crisco, J., McGovern, R., 1997. Efficient calculation of mass moments of inertia for segmented homogenous three-dimensional objects. Journal of Biomechanics 31, 97 – 101. URL: http://www.sciencedirect.com/science/article/pii/S0021929097001085, doi:https://doi.org/10.1016/S0021-9290(97)00108-5.

[6] Donner, R., Menze, B.H., Bischof, H., Langs, G., 2013. Global localization of 3d anatomical structures by pre-filtered hough forests and discrete optimization. Medical Image Analysis 17, 1304 – 1314. doi:https://doi.org/10.1016/j.media.2013.02.004.

[7] Ferrari, R.J., Allaire, S., Hope, A., Kim, J.J., Jaffray, D.A., Pekar, V., 2011. Detection of point landmarks in 3d medical images via phase congruency model. Journal of the Brazilian Computer Society 17, 117–132.

[8] Ghesu, F., Georgescu, B., Zheng, Y., Grbic, S., Maier, A., Hornegger, J., Comaniciu, D., 2019. Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 176–189. doi:10.1109/TPAMI.2017.2782687.

[9] Greenspan, H., Ruf, A., Goldberger, J., 2006. Constrained gaussian mixture model framework for automatic segmentation of mr brain images. IEEE Transactions on Medical Imaging 25, 1233–1245.

[10] Hanaoka, S., Shimizu, A., Nemoto, M., Nomura, Y., Miki, S., Yoshikawa, T., Hayashi, N., Ohtomo, K., Masutani, Y., 2017. Automatic detection of over 100 anatomical landmarks in medical ct images: A framework with independent detectors and combinatorial optimization. Medical Image Analysis 35, 192 – 214. doi:https://doi.org/10.1016/j.media.2016.04.001.

[11] Jadon, S., Srinivasan, A., 2019. Improving siamese networks for one shot learning using kernel based activation functions. ArXiv abs/1910.09798.

[12] Jaklic, A., Solina, F., 2003. Moments of superellipsoids and their application to range image registration. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 33, 648–657.

[13] Kabus, S., Netsch, T., Fischer, B., Modersitzki, J., 2004. B-spline registration of 3D images with Levenberg-Marquardt optimization, in: Fitzpatrick, J.M., Sonka, M. (Eds.), Medical Imaging 2004: Image Processing, International Society for Optics and Photonics. SPIE. pp. 304 – 313. URL: https://doi.org/10.1117/12.533976, doi:10.1117/12.533976.

[14] Koch, G.R., 2015. Siamese neural networks for one-shot image recognition.

[15] Liang, X., Zhao, W., Hristov, D.H., Buyyounouski, M.K., Hancock, S.L., Bagshaw, H., Zhang, Q., Xie, Y., Xing, L., . A deep learning framework for prostate localization in cone beam ct-guided radiotherapy. Medical Physics n/a. doi:10.1002/mp.14355.

[16] Mothes, O., Denzler, J., 2017. Anatomical landmark tracking by one-shot learned priors for augmented active appearance models, pp. 246–254. doi:10.5220/0006133302460254.

[17] Potesil, V., Kadir, T., Platsch, G., Brady, M., 2011. Personalization of pictorial structures for anatomical landmark localization, in: Székely, G., Hahn, H.K. (Eds.), Information Processing in Medical Imaging, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 333–345.

[18] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham. pp. 234–241.

[19] Schmidt, S., Kappes, J., Bergtholdt, M., Pekar, V., Dries, S., Bystrov, D., Schnörr, C., 2007. Spine detection and labeling using a parts-based graphical model, in: Karssemeijer, N., Lelieveldt, B. (Eds.), Information Processing in Medical Imaging, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 122–133.

[20] Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B., 2017. One-shot learning for semantic segmentation.

[21] Suzani A., Seitel A., L.Y.F.S.R.R.A.P., 2015. Fast automatic vertebrae detection and localization in pathological ct scans - a deep learning approach., in: In: Navab N., Hornegger J., Wells W., Frangi A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science.

[22] Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., Wierstra, D., 2016. Matching networks for one shot learning, in: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 29. Curran Associates, Inc., pp. 3630–3638. URL: http://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning.pdf.

[23] Wu, D., Zhu, F., Shao, L., 2012. One shot learning gesture recognition from rgbd images, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 7–12.

[24] Wörz, S., Rohr, K., 2006. Localization of anatomical point landmarks in 3d medical images by fitting 3d parametric intensity models. Medical Image Analysis 10, 41 – 58. doi:https://doi.org/10.1016/j.media.2005.02.003.

[25] Yoo, J.C., Han, T., 2009. Fast normalized cross-correlation. Circuits, Systems and Signal Processing 28, 819–843. doi:10.1007/s00034-009-9130-7.

[26] Zahara, D., Dewi, R.D., Aboet, A., Putranto, F.M., Lubis, N.D., Ashar, T., 2019. Variations in Cochlear Size of Cochlear Implant Candidates. International Archives of Otorhinolaryngology 23, 184 – 190.

[27] Zhang, J., Liu, M., Shen, D., 2017. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural

networks. IEEE Transactions on Image Processing
26, 4753–4764.