VOLUMETRIC LANDMARK DETECTION WITH A MULTI-SCALE SHIFT EQUIVARIANT NEURAL NETWORK

Tianyu Ma[*], **Ajay Gupta**[†], **Mert R. Sabuncu**[*,†]

[*]School of Electrical and Computer Engineering; and Meinig School of Biomedical Engineering, Cornell University

[†]Department of Radiology, Weill Cornell Medical College

Abstract

Deep neural networks yield promising results in a wide range of computer vision applications, including landmark detection. A major challenge for accurate anatomical landmark detection in volumetric images such as clinical CT scans is that large-scale data often constrain the capacity of the employed neural network architecture due to GPU memory limitations, which in turn can limit the precision of the output. We propose a multi-scale, end-to-end deep learning method that achieves fast and memory-efficient landmark detection in 3D images. Our architecture consists of blocks of shift-equivariant networks, each of which performs landmark detection at a different spatial scale. These blocks are connected from coarse to fine-scale, with differentiable resampling layers, so that all levels can be trained together. We also present a noise injection strategy that increases the robustness of the model and allows us to quantify uncertainty at test time. We evaluate our method for carotid artery bifurcations detection on 263 CT volumes and achieve a better than state-of-the-art accuracy with mean Euclidean distance error of 2.81mm.

Index Terms—

Convolutional Neural Networks; 3D landmark detection

1. INTRODUCTION

Recent advances in deep convolutional neural networks (CNNs) show great potential in various medical image computing applications, such as segmentation [1], registration [2], and landmark detection [3]. In this study, we consider the problem of identifying landmark points (sometimes called key-points) that correspond to certain anatomical and/or salient features in the image. This problem has been explored with traditional machine learning techniques [4, 5, 6], yet more recently deep learning methods have produced the best results [7, 8]. One common approach with deep learning is to use a full-convolutional architecture such as U-Net [9] to compute a heatmap image at the output that highlights the location of the landmark(s). Thus landmark localization is turned into an image-to-heatmap regression problem [3], where the ground truth coordinates are used to place Gaussian blobs (of often arbitrary size) to create training data.

An important challenge in applying deep neural networks to volumetric data is computational, since resources needed to train on large-scale volumes will be proportional to the size of data [10]. This is often addressed by breaking the problem down to multiple steps, each solved separately. For example, one can tile the volume into 3D patches and apply landmark detection in each patch separately. Path-level predictions will be limited by the field of view and aggregation of patch-level results into the whole volume can be challenging. One complementary multi-stage strategy is to first use a coarse-scale network that produces candidate patches based on a lower resolution volume [10, 11, 12]. In these approaches, each step's model is trained independently, which can compromise final accuracy due to error accumulation.

In this paper, we build on the multi-scale approach, yet propose a strategy to implement it in a single step. Our building block is a shift-equivariant network that combines a U-Net and a center-of-mass layer to produce predicted landmark coordinates in a volumetric image of a computationally manageable size. Blocks operating at different scales are then connected through differentiable resampling layers, which allows us to train the whole network end-to-end. Unlike heatmap based approaches, the proposed model directly computes spatial coordinates for the landmarks. Our multi-scale network is memory efficient. The high resolution image which cannot fit into our GPU memory in the single-scale implementation can be used as the input for our multi-scale approach Contrary to other methods that use a coarse-to-fine multi-scale design, we employ a single step training scheme which improves accuracy. We further implement noise injection between scales to increase the robustness of our model and achieve uncertainty in predictions at test time.

We apply our method to computed tomography angiography (CTA) scans of the neck to localize the bifurcation of the left and right carotid arteries. In our experiments, we compare the proposed method to several baseline methods, where our model yields superior accuracy with useful uncertainty measurements that correlate with localization error.

2. PROPOSED METHOD

Our model for landmark detection consists of sequentially connected blocks of localizer networks (Loc-Net) that operate at different spatial scales - from coarse to fine. Each Loc-Net is simply a 3D U-Net-like architecture [13] that has a center of mass (CoM) layer at the end (see Fig. 1). As was demonstrated recently [14], a CoM layer computes the weighted average of voxel grid coordinates, where the weights are the pixel values of the U-Net output, which we can interpret as a heatmap that highlights the landmark. The weighted average coordinate vector is the predicted location of the landmark, computed by the scale-specific Loc-Net. Since the U-Net is a fully convolutional architecture, and the CoM is shift equivariant, it is easy to show that the Loc-Net will be shift equivariant[1]. That is, if we shift the input image by a certain amount (along the three axes), the predicted landmark coordinates will be shifted by the same amount. This, we believe, should be a critical property of a landmark detection model.

---

[1]This is approximately true, since down-sampling and pooling operations compromise the shift equivariance. A recent paper proposes a way to fix this issue [15].

Another advantage of the CoM layer is that it circumvents the need for treating the problem as heatmap regression - a common approach for detecting landmarks with neural networks [3]. The main drawback of a heatmap regression approach is that the training annotations, which are often provided as coordinates, need to be converted into heatmaps, e.g., by placing Gaussian kernels. These kernels are usually of arbitrary size. If they are too small, training the networks can be hard. If they are too big, the precision of the landmark detection can be compromised.

We connect scale-specific Loc-Net modules sequentially, from low to high resolution, via an intermediate differentiable crop and resample layer. This layer creates a 3D patch of size $56 \times 56 \times 56$, centered around the predicted landmark location and sampled at the next spatial scale. Fig. 2 illustrates our framework. This model is then trained end-to-end via minimizing the loss between the predicted location of the landmark and the ground truth. Our loss is a weighted sum of the scale-specific squared loss functions. Note that each scale-specific Loc-Net computes a prediction of the landmark location, which we can evaluate with a squared loss with respect to the ground truth. In our experiments, we implemented a piecewise linearly scheduled weighting scheme in combining scale-specific losses. All weights of the scale-specific loss terms were constant during each epoch. The relative weight of the lowest-scale loss started at one and linearly decreased to zero by epoch 500. The relative weight of the highest-scale loss started at zero and linearly increased to 1 by epoch 500. The loss weights corresponding to the other two scales started at zero, linearly increased and peaked at epoch 250, and then linearly decreased to zero by epoch 500. Thus, initially, our model's training was focused on predicting the landmark coordinates at lower scales, eventually focusing on the accuracy of the highest resolution.

In this paper, we also explore the use of a novel stochastic regularization technique, which allows us to improve test time accuracy and quantify uncertainty in predictions. Instead of propagating the predicted landmark to the final scale, we add random shifts (see below for further details). This approach is reminiscent of stochastic dropout [16] and its variants. While dropout was originally proposed to be used during training (to improve robustness), it was recently shown that its use at test time can be viewed as an approximate Bayesian inference strategy that can yield well-calibrated uncertainty estimates - allowing researchers to combat the over-confidence issue of modern deep neural networks [17]. Inspired by this, we also implement our noise injection strategy at test time to compute the uncertainty in the landmark predictions. So at test time, we implement multiple forward passes through our network with different noise realizations injected at the input of the final Loc-Net module, thus computing multiple predictions. The standard deviations of these predictions can be used to quantify the uncertainty in the localization.

2.1. Implementation

We implement our code in tensorflow. Our mini-batch size is 1 3D CTA volume, and all methods are trained for 500 epochs using the Adam optimizer with a learning rate of 0.0005. For noise injection, a 5mm uniform random shift in each direction is applied before cropping the 0.5mm resolution image patch.

3. EMPIRICAL RESULTS

In this study approved by the Weill Cornell institutional review board (IRB), we analyze 263 anonymized CTA exams obtained in the course of clinical care from patients with uni-lateral > 50% extracranial carotid artery stenosis, who were imaged at Weill Cornell Medicine/New York Presbyterian Hospital. All the data involving human subjects used in this paper has obtained the corresponding ethics approval/waiver. All CTA volumes are interpolated to a high resolution grid of size $448 \times 448 \times 448$ and 0.5mm voxel spacing. All down-sampled images used by our model are then created from this set of high resolution images. The lower resolution images have 1mm, 2mm and 4mm isotropic voxels. The data are randomly split into three non-overlapping sets where 167 for training, 42 for validation and 54 for testing. All of our landmarks for left and right carotid artery bifurcations are manually annotated by a trained human (T.M.).

To demonstrate the efficacy of the CoM layer, as a baseline, we implement a single scale 3D U-Net with a similar architecture to our Loc-Net[2] that computes a heatmap output. This is trained with L2-loss on the heatmap images, where the ground truth heatmaps were created by centering a Gaussian kernel (with std = 6 mm) at the annotated landmark coordinates This is compared to a single-scale Loc-Net, which has the CoM layer at the end. Due to limited GPU memory, these single-scale models accept 1mm spacing $224 \times 224 \times 224$ sized volumes as input.

We consider other baselines as well. Our state-of-the-art (SOTA) baseline is a patch based technique that was recently proposed [18]. We also implement two alternative versions of our proposed multi-scale model. In the first alternative, we disconnect the Loc-Nets from the different scales and train them sequentially in separate steps. In this training scheme, the first scale Loc-Net, for example, receives no learning signal from the highest resolution. A second alternative is the full multi-scale model trained in an end-to-end fashion, but without noise injection.

Table 1 shows the average Euclidean distance error (± standard deviation) for all considered methods, computed on the test data. First, we observe that the center of mass layer we use in our model boosts the performance of the U-Net over a heatmamp regression method. Fig. 3 illustrates representative examples of heatmaps computed with and without the CoM. In addition to a more accurate prediction, the CoM layer yields heatmaps that correlate better with the underlying anatomy.

The results further illustrate that the proposed multi-scale model can outperform the state-of-the-art method [18]. Furthermore, both end-to-end training and noise injection improve the accuracy of the predictions.

Fig. 4 presents the violin plots for all the different methods we test, which allows us the appreciate the full distribution of errors on the test cases. We note again that the proposed method of end-to-end training the multi-scale model with noise injection yields the most

---

[2]We included an additional convolution plus down-sample layer

compact distribution of errors with the smallest mean and median values. The improvement in the average error is mainly due to the reduction of the failure cases.

Similar to dropout [17], we apply random shifts (noise injection) at test time to obtain multiple predictions via 50 forward passes through the model. We then compute the standard deviations of these predictions along the three axes. Note that a single forward pass through our model takes 615 milliseconds on a NVIDIA Titan Xp GPU. Assuming an uncorrelated multivariate Gaussian, we compute the 90% confidence volumes for the predictions computed by our model. Fig. 5 plots these values with respect to the localization errors on the test cases. We observe that the confidence volumes (or rather the localization uncertainty) are significantly correlated (Pearson's correlation 0.608, P-value 1.08e-06) with localization error, suggesting that one could use such an approach for automatic quality assurance and propagating uncertainty to downstream analyses.

4. CONCLUSION

In this paper, we explore the use of a novel multi-scale shift-invariant model, trained in an end-to-end fashion, for automatically localizing landmarks, particularly in large-scale 3D images. Our model is memory efficient and thus can be deployed on very large clinical scans at full-resolution, without compromising precision. In our experiments, we demonstrate that the employed center of mass layer can boost performance over a heatmap regression method, while simultaneously yielding more interpretable heatmaps. We also show that the end-to-end training of the connected multi-scale Loc-Nets can yield a significant increase in accuracy. Our method also compares favorably to a state-of-the-art patch-based landmark localization method. Finally, we demonstrate the use of noise injection at test time to obtain useful confidence/uncertainty estimates.

An important weakness of the proposed method is that we assume that the target landmark exists in every single query volume. This is the case for the application we consider, yet it might not hold in general. We plan to pursue this direction in the future.

Acknowledgments

REFERENCES

[1]. Badrinarayanan Vijay, Kendall Alex, and Cipolla Roberto, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 12, pp. 2481–2495, 2017. [PubMed: 28060704]

[2]. Dalca Adrian V, Balakrishnan Guha, Guttag John, and Sabuncu Mert R, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," Medical Image Analysis, 2019.

[3]. Payer Christian, Štern Darko, Bischof Horst, and Urschler Martin, "Regressing heatmaps for multiple landmark localization using cnns," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2016, pp. 230–238.

[4]. Dantone Matthias, Gall Juergen, Fanelli Gabriele, and Van Gool Luc, "Real-time facial feature detection using conditional regression forests," in 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 2578–2585.

[5]. Criminisi Antonio, Shotton Jamie, Robertson Duncan, and Konukoglu Ender, "Regression forests for efficient anatomy detection and localization in ct studies," in International MICCAI Workshop on Medical Computer Vision. Springer, 2010, pp. 106–117.

[6]. U i á Michal, Franc Vojt ch, and Hlavá Václav, "Detector of facial landmarks learned by the structured output svm," VIsAPP, vol. 12, pp. 547–556, 2012.

[7]. Ranjan Rajeev, Patel Vishal M, and Chellappa Rama, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 1, pp. 121–135, 2017. [PubMed: 29990235]

[8]. Zhang Kaipeng, Zhang Zhanpeng, Li Zhifeng, and Qiao Yu, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[9]. Ronneberger Olaf, Fischer Philipp, and Brox Thomas, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[10]. Zheng Yefeng, Liu David, Georgescu Bogdan, Nguyen Hien, and Comaniciu Dorin, "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.

[11]. Zhong Zhusi, Li Jie, Zhang Zhenxi, Jiao Zhicheng, and Gao Xinbo, "An attention-guided deep regression model for landmark detection in cephalograms," arXiv preprint arXiv:1906.07549, 2019.

[12]. Sun Yi, Wang Xiaogang, and Tang Xiaoou, "Deep convolutional network cascade for facial point detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3476–3483.

[13]. Çiçek Özgün, Abdulkadir Ahmed, Lienkamp Soeren S, Brox Thomas, and Ronneberger Olaf, "3d u-net: learning dense volumetric segmentation from sparse annotation," in International conference on medical image computing and computer-assisted intervention. Springer, 2016, pp. 424–432.

[14]. Sofka Michal, Milletari Fausto, Jia Jimmy, and Rothberg Alex, "Fully convolutional regression network for accurate detection of measurement points," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 258–266. Springer, 2017.

[15]. Zhang Richard, "Making convolutional networks shift-invariant again," arXiv preprint arXiv:1904.11486, 2019.

[16]. Srivastava Nitish, Hinton Geoffrey, Krizhevsky Alex, Sutskever Ilya, and Salakhutdinov Ruslan, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.

[17]. Gal Yarin and Ghahramani Zoubin, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in international conference on machine learning, 2016, pp. 1050–1059.

[18]. Noothout Julia M. H., de Vos Bob D., Wolterink Jelmer M., Leiner Tim, and Isgum Ivana, "Cnn-based landmark detection in cardiac CTA scans," CoRR, vol. abs/1804.04963, 2018.
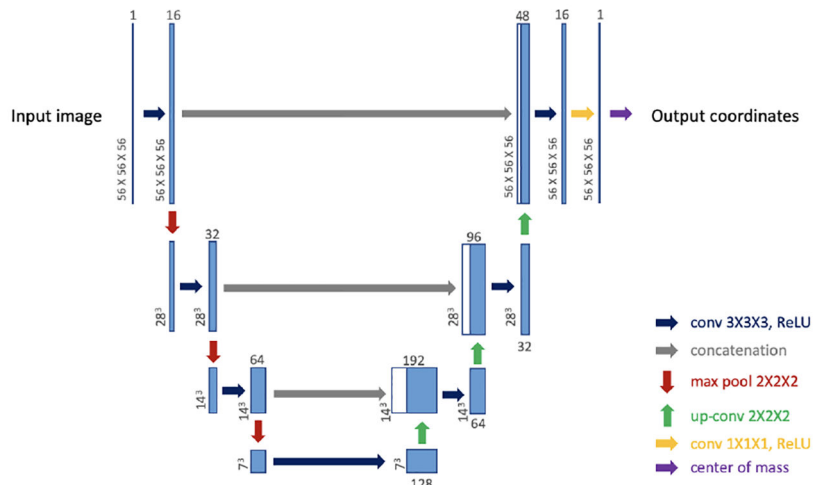
Fig. 1:
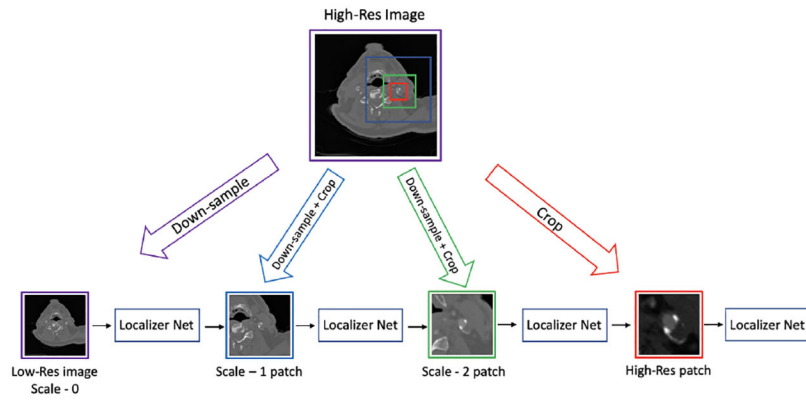Localizer Network (Loc-Net): our building block.

Fig. 2:
Proposed Multi-scale Landmark Localization. The architecture for all Localizer Nets is shown in Fig. 1
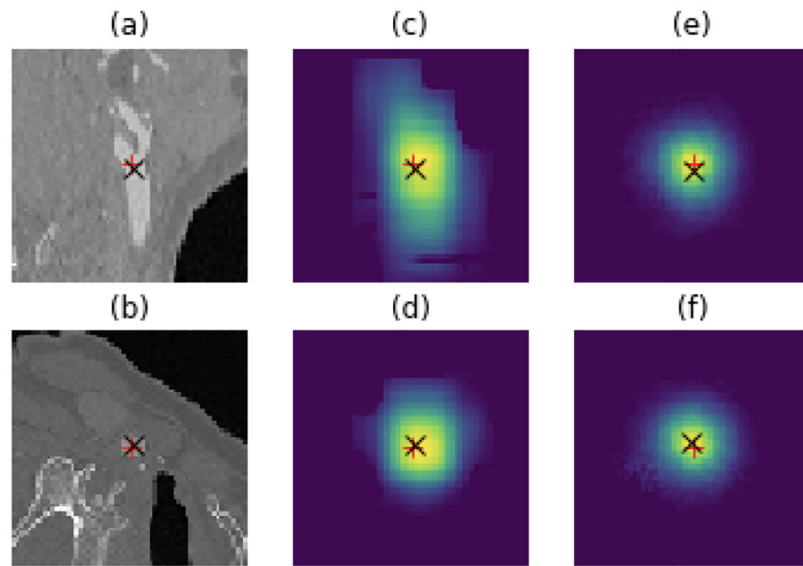
Fig. 3:
A representative test case. '+' shows the ground truth, and 'x' indicates the prediction of landmark. (a) and (b) are sagittal and axial views; (c) and (d) are the heatmaps obtained with single-scale Loc-Net with CoM layer; (e) and (f) are heatmaps computed with single-scale U-Net model.
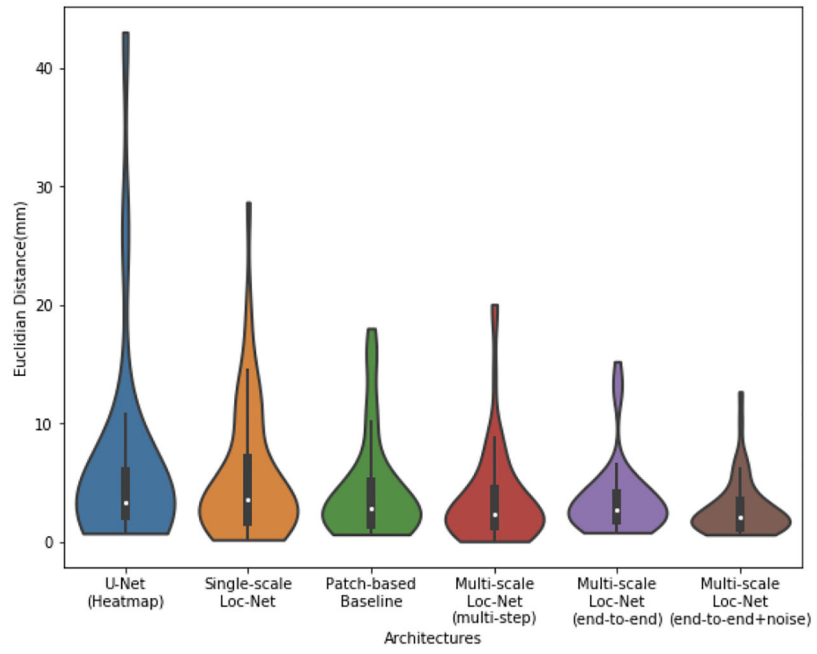
Fig. 4:
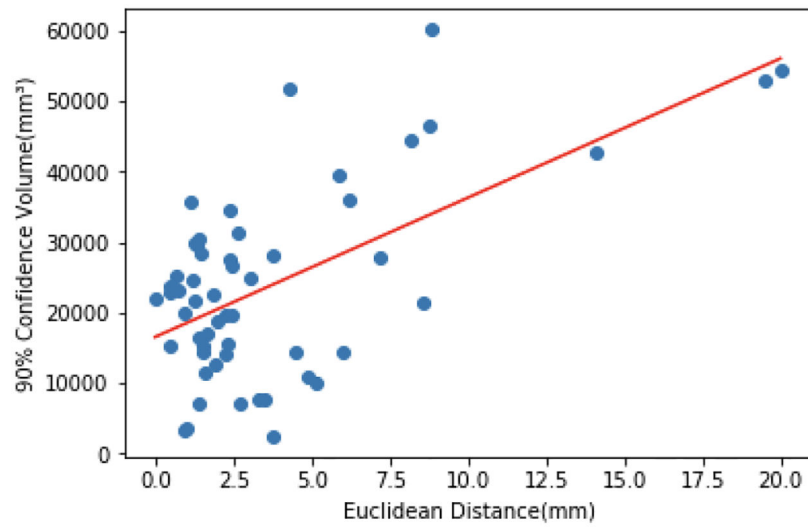Test error distribution for different methods

Fig. 5:
Quantifying localization uncertainty. We use random shifts at test time to obtain 90% confidence volumes for our predictions (see text for further details). These volumes are correlated with the localization errors with respect to the ground truth annotations.

Table 1:

Experimental results for different models

Architectures	Error
U-Net (Heatmap) (1mm resolution)	6.45 ± 9.18mm
Single-scale Loc-Net (1mm resolution)	5.52 ± 5.69mm
Patch-based SOTA Baseline [18]	4.35 ± 4.33mm
Multi-scale Loc-Net (multi-step)	3.72 ± 4.20mm
Multi-scale Loc-Net (end-to-end)	3.61 ± 3.17mm
Multi-scale Loc-Net (end-to-end + noise)	2.81 ± 2.37mm