Medical Image Analysis 54 (2019) 207–219

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Integrating spatial configuration into heatmap regression based CNNs for landmark localization

Christian Payer[a], Darko Štern[b], Horst Bischof[a], Martin Urschler[b,c,*]

[a] Institute of Computer Graphics and Vision, Graz University of Technology, Graz, Austria
[b] Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz, Austria
[c] Medical University of Graz, BioTechMed-Graz, Austria

ARTICLE INFO

Article history:
Received 22 August 2018
Revised 11 January 2019
Accepted 21 March 2019
Available online 25 March 2019

Keywords:
Anatomical landmarks
Localization
Fully convolutional networks
Heatmap regression

ABSTRACT

In many medical image analysis applications, only a limited amount of training data is available due to the costs of image acquisition and the large manual annotation effort required from experts. Training recent state-of-the-art machine learning methods like convolutional neural networks (CNNs) from small datasets is a challenging task. In this work on anatomical landmark localization, we propose a CNN architecture that learns to split the localization task into two simpler sub-problems, reducing the overall need for large training datasets. Our fully convolutional SpatialConfiguration-Net (SCN) learns this simplification due to multiplying the heatmap predictions of its two components and by training the network in an end-to-end manner. Thus, the SCN dedicates one component to locally accurate but ambiguous candidate predictions, while the other component improves robustness to ambiguities by incorporating the spatial configuration of landmarks. In our extensive experimental evaluation, we show that the proposed SCN outperforms related methods in terms of landmark localization error on a variety of size-limited 2D and 3D landmark localization datasets, i.e., hand radiographs, lateral cephalograms, hand MRIs, and spine CTs.

1. Introduction

Localization of anatomical landmarks is an important step in medical image analysis, e.g., to perform segmentation based on deformable statistical models (Beichel et al., 2005; Heimann and Meinzer, 2009), to initialize feature-based image registration (Johnson and Christensen, 2002; Urschler et al., 2006), or to parametrically model detected anatomical structures like vertebrae (Štern et al., 2011). Unfortunately, locally similar structures often introduce difficulties due to ambiguity into landmark localization. Such ambiguities make it hard to achieve low landmark localization error, defined as both high robustness towards landmark misidentification as well as high accuracy locally at each identified landmark. To deal with these difficulties, machine learning based approaches are predominantly used to automatically localize anatomical landmarks in images. These approaches often combine local landmark predictions with explicit handcrafted graphical models, aiming to restrict predictions to feasible spatial configurations as seen in the training data. Thus, the landmark localization problem is simplified by separating the task into two successive steps. The first step is dedicated to locally accurate but potentially ambiguous candidate predictions, while in the second step graphical models (Cootes et al., 1995; Felzenszwalb and Huttenlocher, 2005) eliminate ambiguities to improve robustness towards landmark misidentification.

Recent advances in computer vision have mainly been driven by deep convolutional neural networks (CNNs) due to their superior capabilities to automatically learn important features from images (LeCun et al., 1998, 2015). They have improved the state-of-the-art not only in many computer vision tasks (Krizhevsky et al., 2012; He et al., 2015), but also in a number of medical image analysis applications (Ciresan et al., 2013; Sirinukunwattana et al., 2017). Contrary to previous methods that use prior knowledge to build handcrafted models of landmark configuration, deep CNN architectures rely on large amounts of training data to automatically learn a model that restricts predictions to feasible ones. Especially in the medical image domain, the requirement for large amounts of training data is challenging to fulfill for two reasons. Firstly, ethical and financial concerns hinder large-scale acquisition of

[*] Corresponding author.
E-mail address: martin.urschler@cfi.lbg.ac.at (M. Urschler).

https://doi.org/10.1016/j.media.2019.03.007

medical images. Secondly, in supervised settings trained specialists are needed to create ground truth annotations, which is often difficult, costly and time consuming. Thus, compared to computer vision tasks, CNN-based approaches in medical image analysis have to be able to cope with significantly smaller quantities of annotated training data.

We hypothesize that the need for a large amount of training data can be reduced for CNNs by following the idea explored with handcrafted graphical models. Such models incorporate prior knowledge that the feasible location of a landmark is not distributed uniformly in image space, but is constrained by the locations of other anatomical landmarks. Thus, the hypothesis is that this prior knowledge on *spatial configuration* of landmarks could enable the network to simplify the modeling of the underlying distribution of feasible anatomical configurations and therefore less training data is required. Although constraints on spatial configuration have previously been used in computer vision, e.g., for semantic segmentation (Zheng et al., 2015) or to recognize poses (Tompson et al., 2014), they were only utilized to improve CNN-based predictions regarding robustness towards landmark misidentification, but not to simplify the prediction task based on prior knowledge. In medical image analysis tasks, where high local accuracy is considered equally as important as robustness towards landmark misidentification, incorporating spatial configuration of anatomical landmarks into CNNs to enable learning from limited amounts of training data has not yet been investigated.

In this paper, we show that our proposed two-component CNN is capable to achieve low landmark localization error by learning to benefit from spatial configuration, even with a limited amount of training data. Our SpatialConfiguration-Net (SCN) learns to dedicate one component to deliver locally accurate but potentially ambiguous candidate predictions, and the other component to focus on incorporating spatial configuration to improve robustness towards landmark misidentification by eliminating ambiguities. Thus, the localization task is split into two simpler sub-problems, reducing the overall need for large datasets. In our quantitative evaluation, we show that our proposed approach is widely applicable and outperforms the state-of-the-art on four size-limited 2D and 3D medical image analysis datasets from various modalities, i.e., hand radiographs, lateral cephalograms, hand MRIs, and spine CTs.

1.1. Related work

A prominent strategy for anatomical landmark localization combines local feature responses with handcrafted graphical models encoding the global spatial configuration of landmarks. Such approaches are extensively used in medical image analysis, since they efficiently capture anatomical variation without depending on the size of training datasets, as long as training data is representative. For example, Liu et al. (2010) successively restrict the search space of a landmark based on the relative positions of predictions for other landmarks. They show that this restriction greatly reduces runtime, while also delivering competitive performance. A widely used way of incorporating global shape information is to use point distribution models (Cootes et al., 1995). Lindner et al. (2015) extend upon this strategy by using a constrained local model that iteratively refines global landmark configuration on top of local feature responses generated from random forests (Breiman, 2001), showing state-of-the-art performance on 2D hand and skull radiographs. Other frequently used graphical models in medical image analysis are Markov Random Fields (MRFs). They have been used successfully to localize landmarks in many applications, like lung segmentation (Ibragimov et al., 2012), brain registration (Toews and Arbel, 2007), detection of the spinal column (Glocker et al., 2012), and analysis of whole body CT scans (Potesil et al., 2015), hand radiographs (Donner et al., 2013) or teeth MRI (Štern et al., 2016a). While many of these recent works use the random forest framework solely for generating local appearance feature responses (e.g., Glocker et al., 2012; Donner et al., 2013; Lindner et al., 2015), Ebner et al. (2014) have adapted the seminal work of Criminisi et al. (2013) on organ bounding box localization to first robustly restrict the predicted region based on global appearance features, followed by accurate localization based on local features. However, their performance is highly dependent on whether the first cascade stage delivers robust predictions, since in the second stage only accuracy can be improved. Later, to eliminate false positive predictions from local appearance feature responses, Urschler et al. (2018) integrated spatial configuration of landmarks into a random forest that uses global appearance as well as geometric features. Thus, they mimicked an MRF within a single, unified random regression forest framework.

With the help of large annotated training datasets like ImageNet (Russakovsky et al., 2015), computer vision methods have recently seen a disruptive shift towards deep learning and CNNs. Despite the usually limited amount of annotated training images in the medical imaging domain, some recent methods using CNNs showed success also in anatomical landmark localization. For vertebrae localization and identification, Chen et al. (2015) proposed a three-stage framework combining a random forest used for coarse landmark localization, a shape model incorporating the information of neighboring landmarks for refining their positions, and CNNs for identification of landmarks. A drawback of their method is that they solely use 2D CNNs, thus not benefitting from the full potential of volumetric information possible with 3D CNNs. Zhang et al. (2017) detect thousands of anatomical landmarks simultaneously from a limited amount of MR volumes of the brain, by first registering all volumes to a template volume, and subsequently regressing all landmark coordinates with a single CNN. However, as also observed by methods for human pose estimation (e.g., Toshev and Szegedy, 2014), regressing the coordinates directly involves a highly nonlinear mapping from input images to point coordinates (Pfister et al., 2015).

Instead of regressing coordinates, Tompson et al. (2014) proposed a simpler, image-to-image mapping based on regressing *heatmap* images, which encode the pseudo-probability of a landmark being located at a certain pixel position. Thus, their network for human pose estimation learns to generate high responses on locations close to the target landmarks, while responses on wrong locations are being suppressed. Their work also integrates the binary term of an MRF model inside the CNN architecture. However, Tompson et al. (2014) show shortcomings in landmark localization error due to their method being trained on image patches to predict heatmaps and their separately trained stages being fine-tuned together solely in the end. In our preliminary work (Payer et al., 2016) we also used the heatmap regression framework, however, we integrated spatial information of landmarks directly into an end-to-end trained, fully-convolutional network. There, we showed the potential to achieve good performance even in the presence of very limited amounts of training data. Building upon our proposed approach from Payer et al. (2016), Yang et al. (2017) used the heatmap regression framework to generate predictions for landmarks with missing responses, by incorporating a pretrained model of neighboring landmarks into their CNN. However, in contrast to our method, which directly reduces false positive responses on similar looking landmarks, their method needs an additional postprocessing step to remove false positive responses. Aiming for vertebrae identification and localization, Liao et al. (2018) proposed a three stage method. They pretrain a network to classify and localize vertebrae simultaneously, use the learned weights to generate responses with a fully convolutional network, and finally remove false positive responses with a bidirectional recurrent

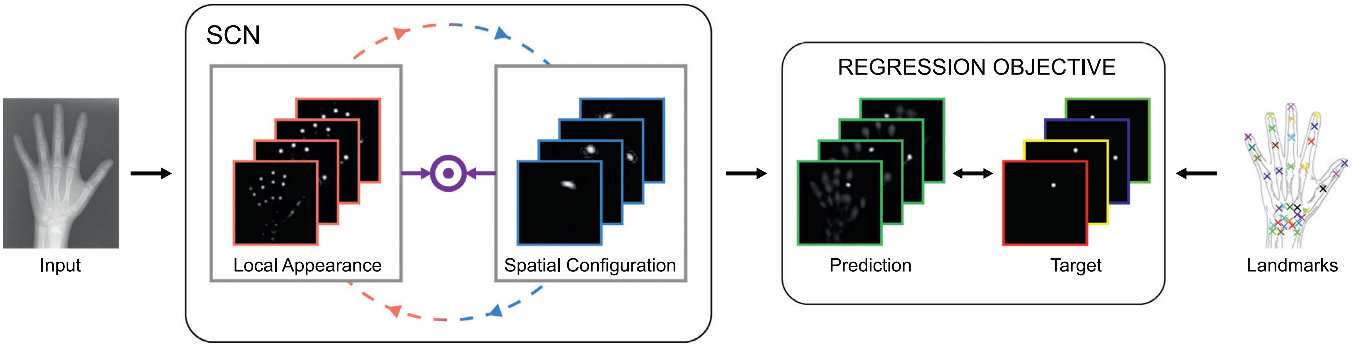


Fig. 1. Anatomical landmark localization by regressing a heatmap for each landmark in an end-to-end trained fully convolutional CNN framework. In our proposed SpatialConfiguration-Net (SCN), the two components interact via multiplication such that the local appearance component focuses on locally accurate but ambiguous candidate predictions, while the spatial configuration component focuses on reducing ambiguities to improve robustness towards landmark misidentification.

neural network. Very recently, a different strategy was investigated by Ghesu et al. (2018, 2019). They proposed the use of reinforcement learning to generate navigation trajectories that point towards the sought for landmarks. Limitations of their method are the computationally extensive training and the need for very large datasets.

The aforementioned CNN methods either need lots of training data, sophisticated postprocessing, or do not generalize to both 2D and 3D data. In this paper, we propose a single end-to-end trained fully convolutional network that works well in both 2D and 3D applications with limited amounts of training data, and does not need any dataset specific postprocessing.

1.2. Contributions

We propose to incorporate spatial configuration of anatomical landmarks into a CNN-based heatmap regression framework providing low landmark localization error even in the presence of limited training datasets. In our single stage approach, the network itself learns to dedicate one component to locally accurate but possibly ambiguous candidate predictions, while the other component solely has the task of eliminating ambiguities to improve robustness towards landmark misidentification. Trained with a single optimization process and in an end-to-end manner, both *local appearance* and *spatial configuration* components of our fully convolutional SCN focus on two simpler sub-problems that individually require less training data. Thus, the overall need to learn from large datasets is reduced.

The proposed method in this work extends our preliminary MICCAI paper (Payer et al., 2016) as follows:

- We modify the objective function to allow learning of the optimal heatmap target function for each landmark separately, depending on the prediction confidences of the network (see Section 2.1).
- We change the CNN structure of the *local appearance* component to better take into account image appearance features at different scales for regressing heatmaps that generate locally accurate candidate predictions (see Section 2.2.2).
- We improve the *spatial configuration* component to allow more variation among landmark configurations, showing benefits in disambiguating locally similar structures especially needed for challenging tasks like vertebrae identification and localization (see Section 2.2.3).
- We show the wide applicability of the SCN on multiple diverse medical imaging datasets in both 2D, i.e., hand radiographs and lateral cephalograms, as well as 3D, i.e., hand MRIs and spine CTs.
- We report new state-of-the-art results in terms of landmark localization error on all investigated datasets, including the ISBI 2015 Cephalometric X-ray Image Analysis Challenge (Wang et al., 2016) and the MICCAI CSI 2014 Vertebrae Localization and Identification Challenge involving pathological spine CTs from Glocker et al. (2013).

2. Method

Aiming for low landmark localization error in the presence of limited training datasets, our proposed SCN architecture directly combines the corresponding outputs of two interacting components. As illustrated in Fig. 1, the interaction of these components is made possible by multiplying the predictions from both components, when our fully convolutional network architecture based on heatmap regression (Section 2.1) is trained in an end-to-end manner. Due to this interaction, our SCN learns to dedicate its *local appearance* component to deliver locally accurate but potentially ambiguous candidate predictions, and its *spatial configuration* component to focus on the improvement of robustness towards landmark misidentification by eliminating ambiguities (see Section 2.2). When combined through multiplication, this interaction results in low landmark localization error, i.e. both high robustness towards landmark misidentification as well as high local accuracy at each identified landmark.

2.1. Using CNNs for regressing heatmaps

While CNNs that directly regress landmark coordinates (Toshev and Szegedy, 2014; Zhang et al., 2017) require dense layers with many network parameters to model the highly nonlinear and difficult to learn image to coordinate mapping, our method is based on regressing heatmap images (Tompson et al., 2014), which encode the pseudo-probability of a landmark being located at a certain pixel position. By enabling an image to image mapping, we therefore benefit from the use of fully convolutional networks (Shelhamer et al., 2017), since the number of network weights and thus computational complexity is reduced.

With N being the total number of landmarks, a d-dimensional heatmap image $g_i(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ of a target landmark L_i, $i = \{1, \ldots, N\}$ with coordinate $\overset{*}{\mathbf{x}}_i \in \mathbb{R}^d$ is defined as the Gaussian function

$$g_i(\mathbf{x}; \sigma_i) = \frac{\gamma}{(2\pi)^{d/2}\sigma_i^d} \exp\left(-\frac{\|\mathbf{x} - \overset{*}{\mathbf{x}}_i\|_2^2}{2\sigma_i^2} \right). \tag{1}$$

Thus, heatmap pixels near the target coordinate $\overset{*}{\mathbf{x}}_i \in \mathbb{R}^d$ have high values, which smoothly but rapidly decrease farther away from $\overset{*}{\mathbf{x}}_i$.

We introduce a scaling factor γ to avoid numerical instabilities during training due to otherwise very small values of the Gaussian function. Equal for each dimension d, the standard deviation σ_i defines the peak width of the Gaussian function in the heatmap image for landmark L_i.

Differently not only from our previous work (Payer et al., 2016), but also from other heatmap based methods (e.g., Tompson et al., 2014), we treat σ_i as an unknown parameter of g_i to allow it to be learned in addition to the network weights **w** and biases **b** during training the CNN. Thus, we enable learning of the optimal heatmap peak width separately for each landmark, depending on the prediction confidences of the network.

The network learns to regress N heatmaps *simultaneously* by minimizing the differences between the predicted heatmaps $h_i(\mathbf{x};\ \mathbf{w}, \mathbf{b})$ and the corresponding target heatmaps $g_i(\mathbf{x};\ \sigma_i)$ for all landmarks L_i:

$$\min_{\mathbf{w},\mathbf{b},\boldsymbol{\sigma}} \sum_{i=1}^{N} \sum_{\mathbf{x}} \| h_i(\mathbf{x}; \mathbf{w}, \mathbf{b}) - g_i(\mathbf{x}; \sigma_i) \|_2^2 + \alpha \|\boldsymbol{\sigma}\|_2^2 + \lambda \|\mathbf{w}\|_2^2. \tag{2}$$

This novel objective function extends upon (Payer et al., 2016) by a regularization term involving the L_2 norm of the heatmap peak widths $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)^T$. As $\boldsymbol{\sigma}$ are learnable network parameters, this term avoids the trivial solution when $\sigma_i \rightarrow \infty$ leading to $g_i(\mathbf{x};\ \sigma_i) \approx 0$. The factor α defines how strong the heatmap peak widths $\boldsymbol{\sigma}$ are being penalized, and λ controls the influence of the L_2 norm of the network weights **w**.

In (2), the L_2 distance of $h_i(\mathbf{x};\ \mathbf{w}, \mathbf{b})$ and $g_i(\mathbf{x};\ \sigma_i)$ and the term penalizing $\boldsymbol{\sigma}$ work against each other. To minimize (2), the former term prefers larger $\boldsymbol{\sigma}$, whereas the latter term urges $\boldsymbol{\sigma}$ to be as small as possible. The network predicts narrower heatmap peak widths for landmarks, where it is confident that the prediction is correct, and wider peak widths for more uncertain landmarks, originating e.g., from landmarks that are hard to specify exactly. Thus, as shown in Fig. 2, the network learns the optimal tradeoff between large σ_i generating oversmoothed, potentially inaccurate predictions, and small σ_i leading to potentially highly accurate responses but with multiple peaks in close proximity.

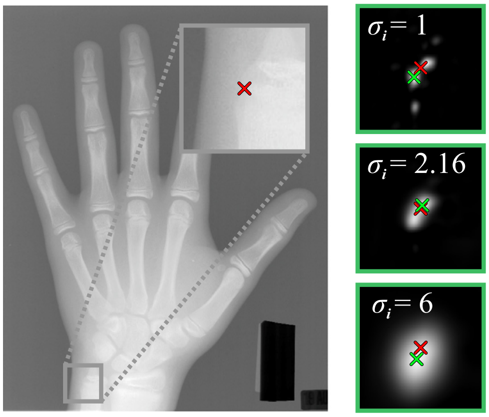


Fig. 2. Example heatmap output for different σ_i for the zoomed region of landmark L_i. Top image shows multiple peaks when choosing fixed σ_i too small; middle image shows responses for learned $\sigma_i = 2.16$; bottom image shows an oversmoothed response for too large fixed σ_i. The target coordinate $\overset{*}{\mathbf{x}}_i$ is depicted by ×, predicted coordinates $\hat{\mathbf{x}}_i$ by ×.

In network inference, we obtain the predicted coordinate $\hat{\mathbf{x}}_i \in \mathbb{R}^d$ of each landmark L_i by taking the coordinate where the heatmap has its highest value:

$$\hat{\mathbf{x}}_i = \arg\max_{\mathbf{x}} h_i(\mathbf{x}; \mathbf{w}, \mathbf{b}). \tag{3}$$

2.2. SpatialConfiguration-Net

The fundamental concept of the SCN for heatmap regression is the interaction between its two components, representing *local appearance* and *spatial configuration*, respectively (see Fig. 3). In Section 2.2.1, we explain how the two components interact in order to simplify the localization task into two sub-problems. Although the two interacting components are in principle flexible regarding their architectures, in Section 2.2.2 and 2.2.3, we describe our specifically proposed architectures for both components.

2.2.1. Local appearance ⇔ spatial configuration

For N landmarks, the set of predicted heatmaps $\mathbb{H} = \{h_i(\mathbf{x}) \,|\, i = 1 \ldots N\}$ is obtained by element-wise multiplication

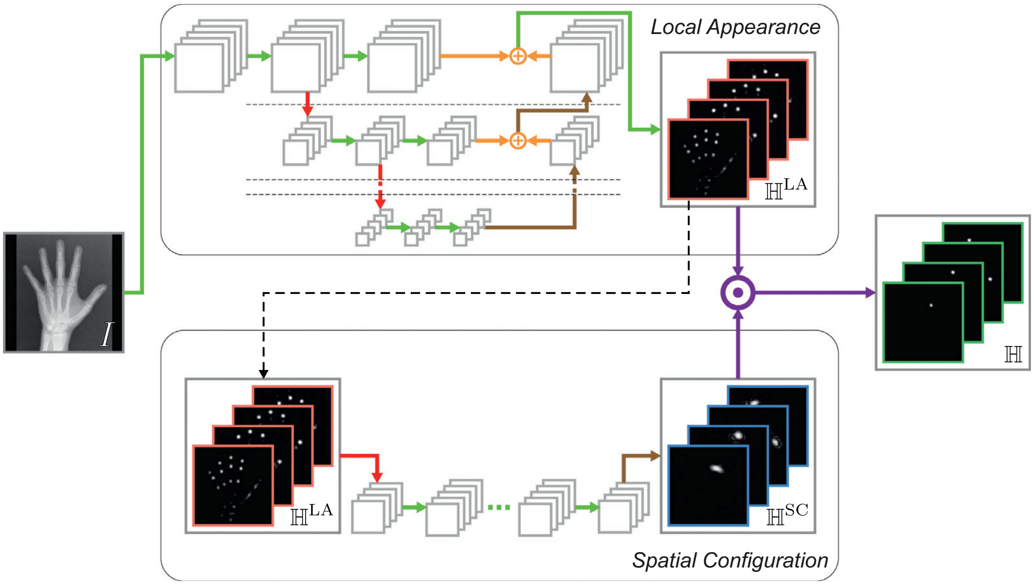


Fig. 3. Schematic representation of our proposed SCN. In the *local appearance* component, the input image I is transformed into $\mathbb{H}^{\text{LA}}$, representing *local appearance* heatmaps for each of the N landmarks. The dashed black line indicates that $\mathbb{H}^{\text{LA}}$ is used as an input for the *spatial configuration* component, where $\mathbb{H}^{\text{LA}}$ is transformed into the *spatial configuration* heatmaps $\mathbb{H}^{\text{SC}}$. A multiplication of $\mathbb{H}^{\text{LA}}$ and $\mathbb{H}^{\text{SC}}$ results in the final heatmaps $\mathbb{H}$. Empty boxes represent intermediate images; arrows represent connections, i.e., → convolution, → downsampling, → upsampling; ⊕ represents pixel-wise addition, ⊙ represents pixel-wise multiplication.

$\odot$ of the corresponding heatmap outputs of the two components:

$$h_i(\mathbf{x}) = h_i^{\mathrm{LA}}(\mathbf{x}) \odot h_i^{\mathrm{SC}}(\mathbf{x}), \tag{4}$$

where network weights $\mathbf{w}$ and biases $\mathbf{b}$ are omitted from here on due to ease of notation. The heatmaps $h_i^{\mathrm{LA}}(\mathbf{x})$ and $h_i^{\mathrm{SC}}(\mathbf{x})$ are the outputs of the *local appearance* and the *spatial configuration* components for each landmark L_i, respectively.

The two components interact through the multiplication in (4). This multiplication is crucial for the SCN to learn the simplification of the localization task, as it forces both components to generate a response on the location of the target landmark $\overset{*}{\mathbf{x}}_i$, i.e., both $h_i^{\mathrm{LA}}(\mathbf{x})$ and $h_i^{\mathrm{SC}}(\mathbf{x})$ deliver responses for $\mathbf{x}$ close to $\overset{*}{\mathbf{x}}_i$, while on all other locations one component can have a response as long as the other one does not have one. Thus, as long as the *spatial configuration* component does not have a response on locations of locally similar structures, the *local appearance* component can concentrate on transforming the input image to a locally highly accurate response at the location of the target landmark $\overset{*}{\mathbf{x}}_i$, without the need for suppressing locally similar structures. On the other hand, as long as the *local appearance* component generates a locally highly accurate response on the location of $\overset{*}{\mathbf{x}}_i$, the *spatial configuration* component can focus on discriminating locally similar structures by eliminating false positive responses from the outputs of the *local appearance* component, without the need to be highly accurate on the location of $\overset{*}{\mathbf{x}}_i$.

In comparison, previous works (Tompson et al., 2014; Pfister et al., 2015) aim for robustness towards landmark misidentification as well as local accuracy of the identified landmarks *in each network component simultaneously*, with the drawback of not achieving a simplification of the localization task. On the other hand, our architecture leads to a solution where both components are facing simpler tasks that can be learned from smaller amounts of training data.

2.2.2. Local appearance

Due to the multiplication in (4), the main focus of the *local appearance* component is to transform the input image I into a set of locally accurate but potentially ambiguous heatmaps $\mathbb{H}^{\mathrm{LA}} = \{h_i^{\mathrm{LA}}(\mathbf{x}) \mid i = 1 \ldots N\}$. Thus, for each landmark L_i the *local appearance* component generates the heatmap output $h_i^{\mathrm{LA}}(\mathbf{x})$, resembling the Gaussian target $g_i(\mathbf{x})$ solely in the close proximity of the landmark coordinate $\overset{*}{\mathbf{x}}_i$. This is achieved with a multi-level structure that is inspired by fully convolutional networks (Shelhamer et al., 2017; Ronneberger et al., 2015) and the residual network (He et al., 2016). As shown in the local appearance part of Fig. 3, each level consists of several consecutive convolution layers. In the multi-level structure, an average pooling layer connected before the last convolution layer of each level generates the input for the next lower level at half the resolution in our contracting path. In the expanding path, the outputs of each level are linearly upsampled to double the resolution. These upsampled outputs are added to the outputs of the final convolution layer from the next higher level. The outputs of each level represent a residual to the next lower levels, thus an intermediate heatmap is iteratively refined and at the same time resolution is increased until the original resolution is reached again. A last convolution layer at the highest level with the original resolution generates the set of local appearance heatmaps $\mathbb{H}^{\mathrm{LA}}$.

Compared to other fully convolutional network architectures like Ronneberger et al. (2015), the expanding path of our *local appearance* component uses less parameters making the SCN faster to train.

2.2.3. Spatial configuration

Due to the multiplication in (4), the main focus of the *spatial configuration* component is to disambiguate locally accurate but ambiguous heatmaps $\mathbb{H}^{\mathrm{LA}}$ from the *local appearance* component, thus providing robustness towards landmark misidentification for localization. Using only local appearance heatmaps $\mathbb{H}^{\mathrm{LA}}$ as its input, the *spatial configuration* component implicitly incorporates a geometric model of the spatial configuration of landmarks by learning how to robustly predict the position of a single landmark from local position predictions $\mathbb{H}^{\mathrm{LA}}$ of all landmarks. By transforming the whole set of local appearance heatmaps $\mathbb{H}^{\mathrm{LA}}$ into a single heatmap $h_i^{\mathrm{SC}}(\mathbf{x})$ for each landmark L_i, $h_i^{\mathrm{SC}}(\mathbf{x})$ delivers responses on coordinates $\mathbf{x}$ close to the target $\overset{*}{\mathbf{x}}_i$ and suppresses responses elsewhere. Thus, within our *spatial configuration* component, false positive responses in $h_i^{\mathrm{LA}}(\mathbf{x})$ are suppressed by constraining responses to feasible landmark configurations.

As shown in the spatial configuration part of Fig. 3, we model the transformations from $\mathbb{H}^{\mathrm{LA}}$ to $\mathbb{H}^{\mathrm{SC}} = \{h_i^{\mathrm{SC}}(\mathbf{x}) \mid i = 1 \ldots N\}$ with consecutive convolution layers. To cover the space among landmarks, these convolution layers need to have a large receptive field. As there is no need for high local accuracy in the *spatial configuration* component, the convolution layers can be calculated on a lower resolution compared to the *local appearance* component, which additionally enables keeping the convolution kernel sizes and the computational complexity reasonably small. After downsampling $\mathbb{H}^{\mathrm{LA}}$, the consecutive convolution layers generate the downsampled version of the heatmap $\mathbb{H}^{\mathrm{SC}}$, which is resized to $\mathbb{H}^{\mathrm{SC}}$ with an upsampling layer. Thus, $\mathbb{H}^{\mathrm{LA}}$ and $\mathbb{H}^{\mathrm{SC}}$ have the same size to enable the element-wise multiplication given in (4).

In contrast to our previous work (Payer et al., 2016) as well as (Tompson et al., 2014), which only allow modeling of pairwise relationships as in the MRF model, multiple convolution layers are able to model more complex relationships between landmarks. This greatly increases the variability of landmark configurations representable by the network. Additionally, multiple layers capture the same receptive field with smaller kernel sizes, thus reducing the number of parameters needed to capture the spatial configuration of landmarks. Therefore, our consecutive convolution layers increase potential representation capabilities of the network, while keeping network parameters and computational effort low.

3. Experimental setup and training details

Training and testing of the network was done in Tensorflow.[1] Preprocessing of the input images for the network is performed as follows. We keep the network input image and output heatmap size in pixels constant for every image of the dataset. Input images are rescaled with fixed aspect-ratio to fit the image size of the network inputs, while network outputs are rescaled back to the original input image size before using (3) to generate final landmark coordinate predictions $\hat{\mathbf{x}}_i$. Resampling uses bi/tricubic interpolation. Furthermore, the intensity range of each dataset is scaled to be within $[-1, 1]$ by shifting and scaling each pixel/voxel of an image with the same values derived from the minimum and maximum intensity of the dataset. In contrast to our previous work (Payer et al., 2016), we neither subtract the mean nor normalize the standard deviation of the input images.

We perform on-the-fly data augmentation using SimpleITK.[2] The intensity values are randomly multiplied with [0.75, 1.25] and shifted by $[-0.25, 0.25]$. The images are randomly translated by $[-20, 20]$ pixels, rotated by $[-15^\circ, 15^\circ]$, and scaled by $[-0.6, 1.4]$. We additionally employ elastic deformations by randomly moving

[1] https://www.tensorflow.org/.

[2] http://www.simpleitk.org/.

points on a regular 12×12 pixel grid by 5 pixels, and interpolating with 3rd order B-splines. All augmentation operations sample randomly from a uniform distribution within the specified intervals. The code of our on-the-fly training tool and the implemented CNN architectures are publicly available[3].

Unless otherwise stated, we set the parameters of the SCN as follows. The *local appearance* component has four levels, each consisting of three consecutive 3×3 (3D: $3 \times 3 \times 3$) convolution layers having 128 outputs, while a 2×2 (3D: $2 \times 2 \times 2$) average pooling after the second convolution layer generates the level below. We include dropout (Srivastava et al., 2014) of 0.5 after the first convolution layer at each level to improve generalization. The local appearance heatmaps $\mathbb{H}^{\text{LA}}$ are generated from a 3×3 (3D: $3 \times 3 \times 3$) convolution layer having a number of outputs equal to the number of landmarks. The *spatial configuration* component is calculated at $\frac{1}{16}$ (3D: $\frac{1}{4}$) of the input resolution. It consists of three consecutive 11×11 (3D: $7 \times 7 \times 7$) convolution layers with 128 outputs and an additional 11×11 (3D: $7 \times 7 \times 7$) convolution layer having a number of outputs equal to the number of landmarks. These outputs are upsampled back to the input resolution with bi-/tricubic interpolation to generate $\mathbb{H}^{\text{SC}}$. Each intermediate convolution layer of the whole SCN has a LeakyReLU (Maas et al., 2013) activation function with a negative slope of 0.1 to ease convergence at training. The convolution layer generating $\mathbb{H}^{\text{LA}}$ has a linear activation function; the convolution layer generating $\mathbb{H}^{\text{SC}}$ has a TanH activation function to restrict the outputs between -1 and 1. The biases of the convolution layers are initialized with 0, the weights with the method described in (He et al., 2015), except the layers generating heatmaps $\mathbb{H}^{\text{LA}}$ and $\mathbb{H}^{\text{SC}}$. Here, initial weights are drawn from a Gaussian distribution with standard deviation of 0.001. These small weights are needed to generate initial heatmap responses close to 0, as otherwise network training would not converge.

For training the networks, we use the objective function (2) with parameters $\gamma = 100$ (3D: $\gamma = 1000$), and $\lambda = 0.0005$ that were empirically determined during initial experiments. To determine parameter $\alpha = 20$ we perform an experiment on our hand radiograph dataset that is described in Section 4.1. For the 3D datasets, we set $\alpha = 1000$, which we determined empirically. We minimize the objective function with Nesterov's Accelerated Gradient (Nesterov, 1983) with learning rate 10^{-6} and momentum 0.99. We use a mini-batch size of 1, since with larger mini-batch sizes we have not seen improvements in results, but only an increase in training time and memory consumption. These parameters remain unchanged for all our tested datasets.

We determine the number of solver iterations for each evaluated dataset by training the network on 80% of the training images and using the remaining 20% of the training images as a validation set to assess when the validation error has reached a plateau. For the datasets evaluated with cross validation, we determine the number of training iterations with initial experiments by evaluating the validation loss on the first fold. All folds of the dataset are then trained for the determined number of iterations.

3.1. Localization U-Net

For a comparison of our proposed SCN to a fully convolutional architecture, we used a U-Net (Ronneberger et al., 2015) based architecture due to its state-of-the-art performance in various segmentation applications, e.g., Sirinukunwattana et al. (2017). Adapting its multi-scale architecture for the landmark localization task based on heatmap regression, we set up our Localization U-Netwith 5 levels and 128 outputs for all intermediate convolution layers, thus resembling a deeper architecture than the *local appearance* component of our SCN. Dropout is implemented in the deepest two levels, as proposed by the authors in Ronneberger et al. (2015). We use padded convolutions to keep input and output image size the same. Furthermore, we use average instead of max pooling and fixed linear upsampling instead of deconvolutions. In a number of parameter tuning experiments with our Localization U-Netimplementation, we found that these modifications of the original implementation of Ronneberger et al. (2015) lead to both faster training and improved results. Moreover, further increasing the number of levels and convolution outputs, i.e., the depth of the Localization U-Net, showed no improvements. We train the Localization U-Netwith the same loss function as the SCN, which also learns the optimal heatmap parameter σ.

3.2. Evaluation metrics

The performance of landmark localization methods was evaluated with several commonly used metrics from the literature, describing localization error in terms of both local accuracy and robustness towards landmark misidentification. The point-to-point error for each landmark L_i in image j is defined as the Euclidean distance between the target coordinate $\overset{*}{\mathbf{x}}{}_i^{(j)} \in \mathbb{R}^d$ and the predicted coordinate $\hat{\mathbf{x}}_i^{(j)} \in \mathbb{R}^d$. To compensate for unknown or varying scale in the datasets, the point-to-point error is multiplied with an image-specific normalization factor $s^{(j)}$, based on the Euclidean distances of specifically selected landmarks. For landmark L_i in an image j, the normalized point-to-point error is defined as

$$\text{PE}_i^{(j)} = s^{(j)} \left\| \overset{*}{\mathbf{x}}{}_i^{(j)} - \hat{\mathbf{x}}_i^{(j)} \right\|_2 . \tag{5}$$

This allows calculation of the median, mean and standard deviation (SD) of the point-to-point error for all images over all landmarks, i.e., PE_{all}. We also report the number of predicted landmarks L_i that are outside a certain point-to-point error radius r for all images, i.e., number of outliers

$$\#\text{O}_r = \left| \left\{ (j, i) | \text{PE}_i^{(j)} > r \right\} \right|. \tag{6}$$

Computed for all landmarks of a single image j, we define the image-specific point-to-point error $\text{IPE}^{(j)}$ as

$$\text{IPE}^{(j)} = \frac{1}{N} \sum_{i=1}^{N} \text{PE}_i^{(j)}, \tag{7}$$

where N is the number of landmarks. Differently from $\#\text{O}_r$ that gives general insights on total number of outliers for all landmarks in all images, IPE provides information on the number of images in which outliers of a certain error radius occur. We present plots of the cumulative IPE distributions, which give the proportion of tested images that achieve a certain IPE.

To compare to other methods on the dataset of spine CT scans, we calculate the number of correctly identified landmarks. As defined by Glocker et al. (2013), a predicted landmark is correctly identified, if the closest groundtruth landmark is the correct one, and the distance from predicted to groundtruth position is less than 20 mm.

The number of correctly identified landmarks is defined as

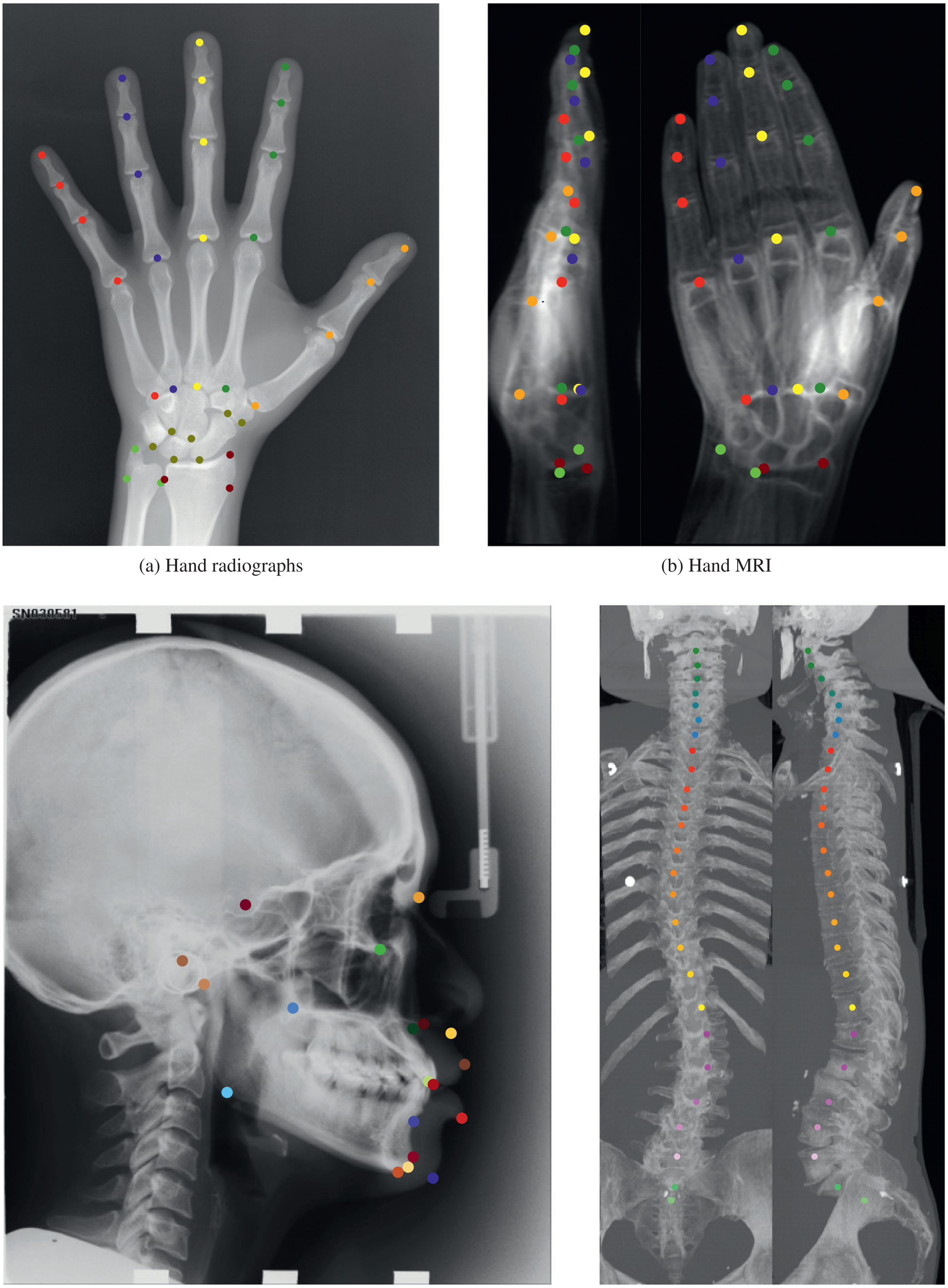
$$\#\text{ID} = \left| \left\{ (j, i) | \arg\min_k \left\| \overset{*}{\mathbf{x}}{}_k^{(j)} - \hat{\mathbf{x}}_i^{(j)} \right\|_2 = i \wedge \text{PE}_i^{(j)} \leq 20 \right\} \right|. \tag{8}$$

For comparison with other methods, the ID_{rate} is defined as the percentage of correctly identified landmarks over all landmarks.

4. Evaluation and quantitative results

We compare our proposed SCN architecture to various state-of-the-art localization methods on four datasets from the

[3] https://www.github.com/christianpayer/.



(a) Hand radiographs

(b) Hand MRI

(c) Lateral Cephalograms

(d) Spine CT

Fig. 4. Sample images of the evaluated datasets. Volumes are projected to 2D coronal and sagittal images for visualization. Circles indicate landmark annotations.

medical imaging domain, i.e., radiographs of left hands (2DHand), volumetric MR scans of left hands (3DHand), lateral cephalograms (2DSkull), and volumetric CT scans of the spine (3DSpine). Fig. 4 shows representative examples for all datasets.

We trained and tested our networks with an NVIDIA Geforce Titan Xp with 12 GB RAM. Training a single network took approximately 4 h for 2DSkull, 6 h for 2DHand and 3DHand, and 20 h for 3DSpine. Testing 2D images took approximately 2 s, testing 3D volumes approximately 5 s.

4.1. Hand radiographs (2DHand)

We use a publicly available dataset of hand radiographs[4] for comparison with state-of-the-art anatomical landmark localization algorithms and to investigate in detail a number of hyperparameters of our SCN. The dataset consists of 895 radiographs of left

[4] Digital Hand Atlas Database System, www.ipilab.org/BAAweb.

Table 1
Localization results from a three-fold cross validation on 895 images from the 2DHandFull dataset with 37 annotated landmarks.

Method	Input Size	PE_{all} (in mm)		$\#O_r$ (in %)		
	(in pix)	Median	Mean ± SD	$r = 2$ mm	$r = 4$ mm	$r = 10$ mm
SCN:	512 × 512	**0.43**	**0.66 ± 0.74**	**1659 (5.01%)**	**241 (0.73%)**	**3 (0.01%)**
Localization U-Net:	512 × 512	0.44	0.70 ± 2.18	1703 (5.14%)	270 (0.82%)	22 (0.07%)
Payer et al. (2016):	256 × 256	0.91	1.13 ± 0.98	4109 (12.4%)	444 (1.34%)	12 (0.04%)
Urschler et al. (2018):	1250 × 1250	0.51	0.80 ± 0.93	2586 (7.81%)	510 (1.54%)	18 (0.05%)
Štern et al. (2016a):	1250 × 1250	0.51	0.80 ± 0.91	2582 (7.80%)	512 (1.55%)	15 (0.05%)
Ebner et al. (2014):	1250 × 1250	0.51	0.97 ± 2.45	2781 (8.40%)	716 (2.16%)	228 (0.69%)
Lindner et al. (2015):	1250 × 1250	0.64	0.85 ± 1.01	2094 (6.32%)	347 (1.05%)	20 (0.06%)

hands with an average size of 1563 × 2169 pixels, acquired with different X-ray scanners. We performed a manual annotation of 37 characteristic landmarks on finger tips and bone joints. As the images lack information about physical pixel resolution, we calculate an image-specific normalization factor $s^{(j)}$. Using the same normalization as in (Lindner et al., 2015; Štern et al., 2016a; Payer et al., 2016), we assume a wrist width of 50 mm determined by two of the annotated landmarks at the wrist, i.e., $s^{(j)} = 50/\|\overset{*}{\mathbf{x}}^{(j)}_{\text{l\_wrist}} - \overset{*}{\mathbf{x}}^{(j)}_{\text{r\_wrist}}\|_2$. We use the same three-fold cross validation setup as described in (Štern et al., 2016a; Payer et al., 2016), who split the 895 input images into three folds with equal number of images, resulting in approximately 600 training and 300 testing images per fold. Moreover, we use the same preprocessing of the input images as (Štern et al., 2016a; Payer et al., 2016), by performing histogram matching to a reference image solely inside the outline of the hand as computed by Otsu thresholding. We train the network for 30,000 iterations.

4.1.1. *Full set of hand radiographs (2DHandFull)*

We show the results of comparing our SCN to our preliminary work (Payer et al., 2016) and other state-of-the-art methods on the 2DHandFull dataset in Table 1, which states the point-to-point error (PE_{all}) for landmark localization error. Additionally, we also report the number of outliers given different error radii ($\#O_r$) to assess robustness towards landmark misidentification. Local accuracy of the identified landmarks can be best seen from the cumulative distribution of the image specific point-to-point error (IPE), which is illustrated graphically in Fig. 5.

From Table 1 and Fig. 5, we see a significant improvement in terms of landmark localization error of our proposed SCN that uses an input/heatmap target image size of 512 × 512 pixels compared to our previous work in Payer et al. (2016). To enable a fair comparison to Payer et al. (2016) that used 256 × 256 pixels as input image size and to investigate the influence of this input size hyperparameter, we perform an experiment on the 2D hand dataset varying the input/heatmap target size between 128 and 512 pixels. The result of this experiment is shown as a cumulative IPE distribution in Fig. 6, demonstrating that we outperform our previous work even if the input/heatmap target size is reduced to 128 × 128 pixels. Thus, our observed improvement compared to Payer et al. (2016) does not come from the increased input/heatmap target size, but instead from the following two extensions.

Firstly, our modified *local appearance* component uses a wider receptive field due to the additional depth levels enabled by the residual architecture (He et al., 2016). We analyzed the effect of different numbers of convolution outputs for this component, obtaining only small differences when varying the number of outputs between 32 and 256, with the difference of 128 and 256 being indistinguishable. Therefore, we use 128 convolution outputs for all our further experiments, since this choice allows faster training.

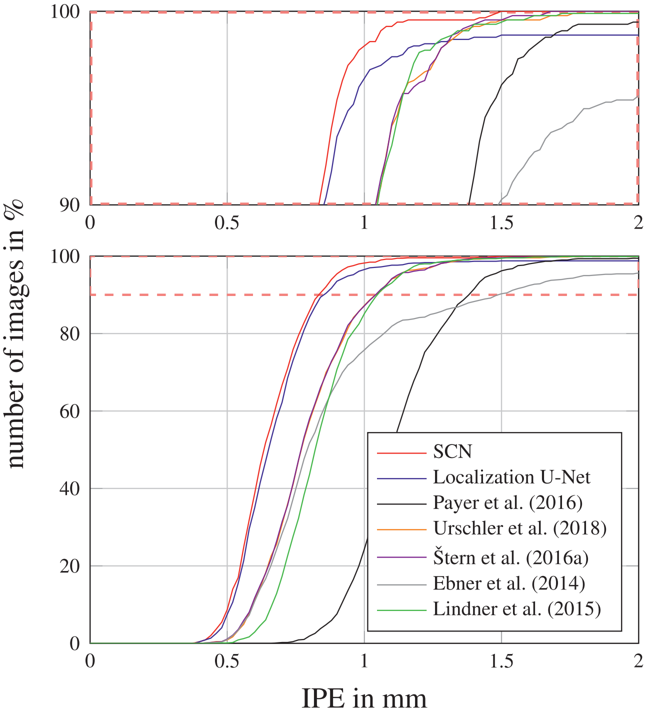


Fig. 5. Cumulative distribution of the image-specific point-to-point error on 2DHandFull dataset. The zoomed-in region of the dashed red box is shown on top. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

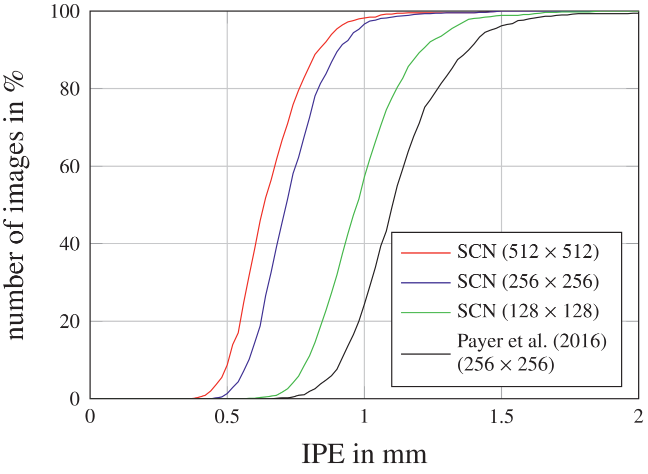


Fig. 6. Cumulative distribution of IPE on 2DHandFull dataset for varying input/heatmap target sizes of our SCN architecture.

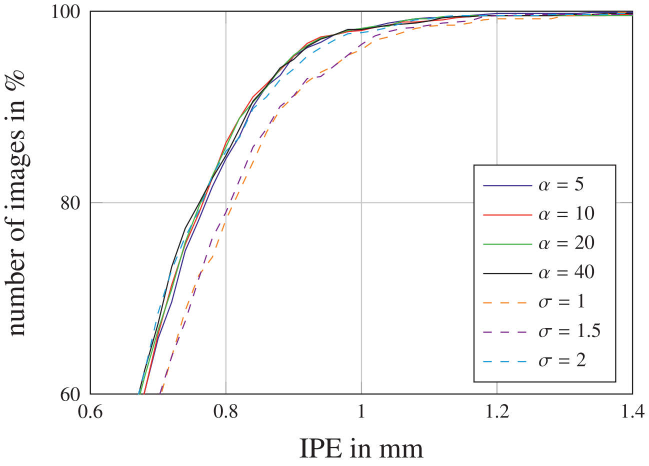


Fig. 7. Cumulative distribution of IPE on 2DHandFull dataset for fixed σ and different weighting factors α when learning $\boldsymbol{\sigma}$ in our SCN.

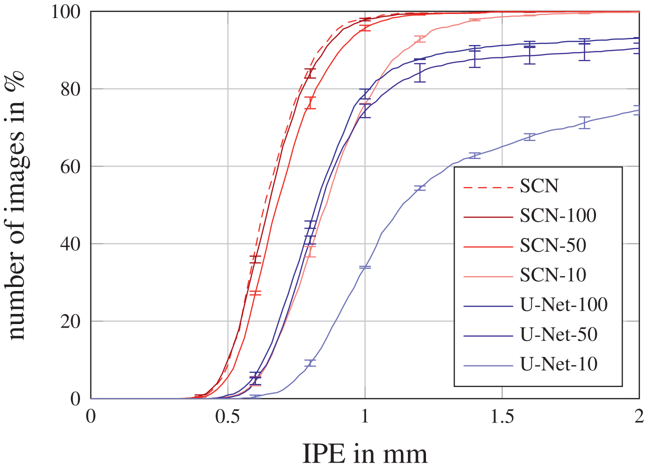


Fig. 8. Cumulative distribution of IPE on 2DHandReduced dataset for different numbers of training images.

Secondly, improvements in localization performance also come from our novel objective function involving learnable heatmap peak widths $\boldsymbol{\sigma}$ for each landmark. To investigate the behavior of this extension, we perform an experiment comparing different fixed values for all σ_i, as proposed in Payer et al. (2016), with the variant where $\boldsymbol{\sigma}$ is learned independently per landmark, and weighting factor α is varied. Fig. 7 shows the results of this experiment, which confirms that varying $\boldsymbol{\sigma}$ independently outperforms the variants with the same fixed σ_i for all landmarks L_i, while the difference in localization performance for varied values of α is small, indicating that this is an uncritical parameter. With independently learned $\boldsymbol{\sigma}$ values, the heatmap peak widths adapt to the data by encoding the uncertainty of network predictions, see Fig. 2. Landmarks that are easier to predict unambiguously develop a smaller σ_i than those that show more variation or ambiguity in the training dataset. Thus, in contrast to fixed σ_i values, no predetermined tradeoff between large σ_i generating oversmoothed, potentially inaccurate predictions, and small σ_i leading to potentially highly accurate responses but with multiple peaks in close proximity has to be made.

When comparing our proposed SCN to other state-of-the-art algorithms, Fig. 5 and Table 1 show that it significantly outperforms all our previous methods, i.e., methods that either solely rely on random regression forests (Ebner et al., 2014; Urschler et al., 2018), or that additionally incorporate explicit graphical models (Štern et al., 2016a), as well as our previous CNN-based method (Payer et al., 2016). Moreover, our SCN also outperforms the state-of-the-art landmark localization method of Lindner et al. (2016), who applied their algorithm to our preprocessed 2DHandFull dataset with the same cross-validation split. Overall, our SCN achieves an unprecedented low number of three outliers at 10 mm for this dataset. Interestingly, in terms of #O$_r$ our optimized Localization U-Netimplementation performs nearly as good as the SCN on the 2DHandFull dataset (see Table 1). This presumably is a consequence of the comparatively large amount of training data that were available for such a deep architecture derived from the U-Net to be trained. However, from the cumulative IPE distribution in Fig. 5, we see that already at an error radius of 1.25 mm, Localization U-Netshows a drop in performance compared to most of the other state-of-the-art methods. Thus, from results of #O$_r$ and IPE it can be concluded that Localization U-Netleads to a few outliers in many images, while methods incorporating graphical models like Lindner et al. (2015) or Urschler et al. (2018) lead to less images with outliers, but more of them per respective image.

4.1.2. Reduced set of hand radiographs (2DHandReduced)

In this experiment, we use the 2DHandFull dataset with its total of 895 images to systematically evaluate how our proposed SCN performs when decreasing the number of training images, while still using the same three-fold cross validation setup as before. Thus, to generate 2DHandReduced, instead of using all 600 training images per fold, we solely train on a random subset of 10, 50, or 100 images, while we test on the same 300 images per fold, such that still every image is tested exactly once. Other training hyperparameters remain the same as in Section 4.1.1. For each cross validation round of 10, 50, and 100 training images, we report our evaluation metrics as averages from three experiments to compensate for randomly selected subsets of images.

Results of this experiment in Fig. 8 and Table 2 show that for smaller training datasets, our SCN performs much better than Localization U-Net, since our SCN-10 has 30 outliers at $r = 10$ mm when trained from solely 10 images, while Localization U-Net-10 trained from 10 images has 633 outliers for the same error radius. Even trained on 100 images, Localization U-Net-100 still has 151 outliers. Interestingly, already when training with 50 images, landmark localization performance of our SCN-50 is almost the same as the original SCN trained on the 2DHandFull dataset. This experiment shows that increasing the number of training images leads to better localization performance of CNN based methods in general, while it also confirms our hypothesis that by incorporating *spatial configuration* inside our SCN, a smaller amount of data is sufficient for training.

4.1.3. Local appearance ⇔ spatial configuration

As shown in the experiments on 2DHandFull and 2DHandReduced, in our work it is not necessary to provide a large amount of training data to achieve a low landmark localization error. This is due to splitting the localization problem into two simpler tasks by multiplying the predictions from the *local appearance* and *spatial configuration* components, as shown in Fig. 3. Again, we used the 2DHandFull dataset to show this simplification in an experiment, where we separately extract the maxima of the heatmaps of both components $\mathbb{H}^{\text{LA}}$ and $\mathbb{H}^{\text{SC}}$ of the normally trained SCN. Results of this experiment are given in Fig. 9 as cumulative distributions of IPE for LA and SC, respectively.

To show that the increased performance of the SCN is not solely caused by the improved LA component, we performed an experiment, where a network consisting only of the LA component (LA-Net) was trained. Due to the reduced receptive field of the LA-Net and the missing spatial configuration component, it performs

Table 2
Cross-validation results on 895 images from 2DHandReduced dataset, averaged from three random selections of training images.

# Training	Method	PE_{all} (in mm)		$\#O_r$ (in %)		
Images		Median	Mean ± SD	$r = 2$ mm	$r = 4$ mm	$r = 10$ mm
10	SCN:	**0.55**	**0.91** ± **1.13**	**3564 (10.76%)**	**707 (2.14%)**	**30 (0.09%)**
	Localization U-Net:	0.71	2.00 ± 9.38	4855 (14.66%)	1627 (4.92%)	633 (1.91%)
50	SCN:	**0.45**	**0.72** ± **0.84**	**2204 (6.66%)**	**379 (1.15%)**	**8 (0.02%)**
	Localization U-Net:	0.59	1.19 ± 5.44	2977 (8.99%)	656 (1.98%)	199 (0.60%)
100	SCN:	**0.44**	**0.69** ± **0.81**	**1895 (5.72%)**	**293 (0.89%)**	**8 (0.02%)**
	Localization U-Net:	0.59	1.09 ± 4.68	2766 (8.35%)	539 (1.63%)	151 (0.46%)

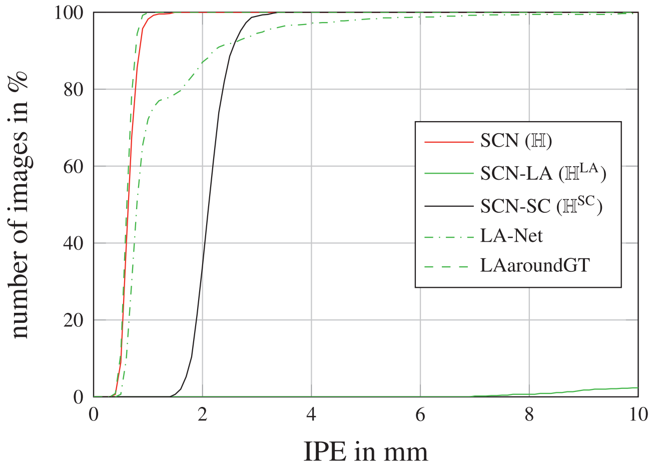


Fig. 9. Cumulative distribution of IPE on 2DHandFull dataset for local appearance (LA) and spatial configuration (SC) components of our SCN (see Fig. 3). LAaroundGT shows LA predictions restricted to a radius within 10 mm of the artificially provided ground truth landmark locations during testing.

worse not only compared to the SCN (see Fig. 9), but also to the Localization U-Net (compare to Fig. 5).

Furthermore, we artificially generate results for the *local appearance* component of our SCN by making available the ground truth landmark locations during testing. Thus, in Fig. 9, the dashed line LAaroundGT shows that there is a locally accurate prediction around the ground truth locations within a radius of 10 mm. However, due to ambiguous structures these predictions are neither the only ones nor the strongest ones that generate a response in the *local appearance* heatmap $\mathbb{H}^{LA}$, as visualized in Fig. 3 and also shown by the low values of the cumulative error distribution for LA in Fig. 9. Still, when combining the locally accurate but ambiguous candidate predictions from the *local appearance* heatmap $\mathbb{H}^{LA}$ with the robust but inaccurate prediction of the *spatial configuration* component $\mathbb{H}^{SC}$, the ambiguities from the *local appearance* component are eliminated, while the local accuracy of our proposed SCN remains the same as for LAaroundGT.

4.2. Hand MRIs (3DHand)

To show the applicability of our SCN to volumetric MR data and to compare to our previously published results, we use an in-house dataset of 60 T1-weighted 3D gradient echo hand MR scans with 28 annotated landmarks, which is intended for automatic forensic age estimation (Štern et al., 2016b). The average volume size is 294 × 512 × 72 with a voxel resolution of 0.45 × 0.45 × 0.9 mm^3. Input and heatmap target size of the volumes is 96 × 128 × 32 voxels. The SCN uses 3 × 3 × 3 kernels in the local appearance block; for the spatial configuration block, we use a downsampling factor of $\frac{1}{4}$, and 7 × 7 × 7 kernels. For the U-Net implementation we use 3 × 3 × 3 kernels and four depth levels. The number of filter outputs for each intermediate convolution layer is set to 64 for both SCN and U-Net. We use the same five-fold cross validation setup as Ebner et al. (2014); Payer et al. (2016); Urschler et al. (2018), with each round consisting of 43 training and 17 testing images. We train the networks for 20,000 iterations. As we know the physical voxel resolution of each image, $s^{(j)}$ is set to 1.

In Table 3, we compare the results to our previous works Ebner et al. (2014); Payer et al. (2016); Urschler et al. (2018). Our results demonstrate that our 3D SCN gives best localization performance of all compared methods also for 3D MR images. We clearly outperform our preliminary work from Payer et al. (2016) in terms of landmark localization error (PE_{all}) using the same voxel resolution in both approaches. We are also significantly better in terms of localization error compared to the approaches based on random regression forests (Ebner et al., 2014; Urschler et al., 2018), although these methods use double the voxel resolution in each dimension. Interestingly, besides our proposed method, the only other method that achieves a perfect outlier rate at $r = 10$ mm on this dataset also makes extensive use of information on spatial configuration of landmarks (Urschler et al., 2018). Additionally to our proposed 3D SCN that shows the best localization performance, also Localization U-Netoutperforms all our previous methods on this small dataset comprised of only 60 volumes. We presume that this good performance comes from our 3D data augmentation scheme involving elastic and intensity transformations, which is highly beneficial for training such deep CNNs.

4.3. Lateral cephalograms (2DSkull)

To compare our proposed SCN in the context of a landmark localization challenge, we apply it on the publicly available dataset that was used for the ISBI 2015 Cephalometric X-ray Image Analysis Challenge (Wang et al., 2016). It consists of 400 lateral cephalograms from 400 different subjects, with 19

Table 3
Cross-validation localization results on 60 images from 3DHand with 28 annotated landmarks.

Method	Input Size	PE_{all} (in mm)		$\#O_r$ (in %)		
	(in vox)	Median	Mean ± SD	$r = 2$ mm	$r = 4$ mm	$r = 10$ mm
SCN:	96 × 128 × 32	**0.90**	**0.84** ± **0.62**	**96 (4.03%)**	**5 (0.21%)**	**0 (0.00%)**
Localization U-Net:	96 × 128 × 32	**0.90**	0.90 ± 1.16	123 (5.17%)	10 (0.42%)	2 (0.08%)
Payer et al. (2016):	96 × 128 × 32	1.01	1.20 ± 1.48	215 (9.03%)	15 (0.63%)	3 (0.13%)
Urschler et al. (2018):	172 × 300 × 72	1.10	1.31 ± 0.72	293 (12.15%)	23 (0.97%)	**0 (0.00%)**
Ebner et al. (2014):	172 × 300 × 72	1.27	1.44 ± 1.51	416 (17.48%)	29 (1.22%)	6 (0.25%)

Table 4
Localization results on 250 test images from 2DSkull dataset with 19 landmarks. We report the mean results of both test datasets of the ISBI 2015 Grand Challenge according to the evaluation protocol in Wang et al. (2016).

Method	#O$_r$ (in %)			
	$r = 2$ mm	$r = 2.5$ mm	$r = 3$ mm	$r = 4$ mm
SCN:	**26.67%**	**21.24%**	**16.76%**	10.25%
Localization U-Net:	27.85%	22.17%	17.96%	11.20%
Urschler et al. (2018):	29.79%	23.05%	17.92%	10.99%
Lindner et al. (2015):	29.35%	23.07%	17.83%	**10.15%**
Ibragimov et al. (2014):	31.87%	25.37%	20.23%	13.13%

Table 5
Localization results on the 3DSpine dataset with at most 26 landmarks per volume. We report results for the two-fold cross-validation with 224 images (*Set1*) and the MICCAI CSI 2014 Challenge test set with 60 images (*Set2*). For *Set2*, we report results on the unmodified test set with mislabeled vertebrae in the groundtruth (*Set2*-mislabeled), and on the test set, where we corrected these mislabeled vertebrae (*Set2*-corrected).[5]

Method	*Set1*		*Set2*-mislabeled	
	PE$_{all}$ (mm)	ID$_{rate}$	PE$_{all}$ (mm)	ID$_{rate}$
SCN:	**6.2 ± 9.9**	**86.1%**	6.0 ± 16.1	90.9%
Localization U-Net:	7.3 ± 12.9	82.7%	9.1 ± 22.1	83.0%
Liao et al. (2018):	–	–	6.5 ± 8.6	88.3%
Sekuboyina et al. (2018)	–	–	7.4 ± 9.3	86.1%
Yang et al. (2017):	9.1 ± 7.2	80.0%	8.6 ± 7.8	85.0%
Chen et al. (2015):	–	–	8.8 ± 13.0	84.2%
Glocker et al. (2013):	12.4 ± 11.2	70.0%	13.2 ± 17.8	74.0%
Method			*Set2*-corrected	
			PE$_{all}$ (mm)	ID$_{rate}$
SCN:			**2.9 ± 4.4**	**96.0%**
Localization U-Net:			4.9 ± 7.7	87.5%

annotated landmarks. The 2D images have a size of 1935 × 2400 with a physical resolution of 0.1×0.1 mm^2 per pixel. We set up the networks with 512 × 512 pixels input/heatmap target size. We use the train/test split (150 training, 250 testing images) and evaluation protocol described in Wang et al. (2016). We train the networks for 30,000 iterations. As the results presented in Wang et al. (2016) lack cumulative distribution graphs, we only compare the reported values of the best performing methods from Ibragimov et al. (2014) and Lindner et al. (2015).

As shown in Table 4, our proposed SCN outperforms previous winners of this challenge. With Lindner et al. (2015) resembling the current state-of-the-art on this dataset, our results demonstrate that our method is more accurate locally (#O$_r$ with $r = 2$ mm), while robustness towards landmark misidentification is the same (#O$_r$ with $r = 4$ mm). We found that most problems of our method on this particularly challenging dataset are at anatomically ill-defined landmarks (e.g., the anterior landmark at the chin in Fig. 4c), whose position are hard to estimate without strong geometric constraints from handcrafted graphical models. This can also be seen by comparing the performance of other methods in terms of outliers with $r = 4$ mm in Table 4. While the method of Lindner et al. (2015) based on a variant of the statistical shape model shows an increase in performance with increasing error radius, Localization U-Netthat learns geometric constraints solely from data shows the largest drop in performance.

4.4. Spine CTs (3DSpine)

To compare our proposed SCN with other recent methods, we also evaluate on a publicly available volumetric dataset of pathological spine CT scans (Glocker et al., 2013) used for the MICCAI CSI 2014 Vertebrae Localization and Identification Challenge. This dataset includes various challenges such as scoliotic spines, vertebrae fractures, metal insertions causing severe image artifacts, and highly restrictive field of views. Thus, due to a significant variation in local appearance, as well as repetitive structures in the spine CT scans, the spatial configuration of landmarks has to be learned to distinguish different vertebrae. In line with previously reported results, we split the evaluation on this dataset into two sets: In *Set1*, the 224 CT scans of the challenge training set are evaluated with two-fold cross-validation. In *Set2*, the networks are trained on all 224 CT scans plus additional 18 CT scans and evaluated on the challenge test set, which consists of 60 CT scans. The average volume size is 512 × 512 × 160 with a voxel resolution of $0.34 \times 0.34 \times 2.06$ mm^3. Input and heatmap target size of the volumes are 96 × 96 × 192 voxels; we resample the input images to have an isotropic spacing of 2 mm per dimension. With these input sizes, the network can process volumes with a physical extent up to $192 \times 192 \times 384$ mm^3. As some volumes have a larger extent in the z-axis (i.e., the axis perpendicular to the axial plane) that would not fit into the network, we process such volumes as follows: During training, we crop a subvolume at a random position at the z-axis. During testing, we split the volumes at the z-axis into multiple subvolumes that overlap for 96 pixels, and process them one after another. Then, we merge the network predictions of the overlapping subvolumes by taking the maximum response over all predictions. The network architectures are the same as for the 3DHand dataset. We train the networks for 40,000 iterations. As we know the physical voxel resolution of each image, we set the normalization factor $s^{(j)} = 1$.

We compare our network predictions with the latest reported results on this highly difficult dataset, see Table 5. Our approach outperforms all state-of-the-art methods evaluated on this challenge dataset, on both cross validation set (*Set1*) and challenge test set (*Set2*[5]). Although the Localization U-Netshows overall good results on this dataset, neither landmark localization error (PE$_{all}$) in general, nor robustness towards landmark misidentification (ID$_{rate}$) specifically are as good as for our SCN. Due to the lack of explicitly modeling the spatial configuration of landmarks, the Localization U-Netshows a significant decrease of 8.5% in ID$_{rate}$ for *Set2*. As compared to random forests (Glocker et al., 2013) and other CNNs (Chen et al., 2015; Yang et al., 2017; Liao et al., 2018), our SCN performing heatmap regression shows the best localization performance. In the very recent work of Sekuboyina et al. (2018), the authors propose to project the 3D information of spine anatomy into 2D sagittal and coronal views, and solely use these views as input for their CNN. Although their method is tailored towards spine datasets, where the landmarks are situated along the axis orthogonal to the axial plane, our generic SCN that directly processes 3D information outperforms their method. Overall, we show that our SCN outperforms all state-of-the-art methods, without the need for sophisticated training procedures (e.g., Liao et al. (2018)) or pre-/postprocessing (e.g., Yang et al., 2017; Sekuboyina et al., 2018).

5. Discussion and conclusion

In this work we have shown that by incorporating *spatial configuration* into a CNN based deep learning architecture that

[5] When evaluating our proposed algorithm on the testing dataset, we identified three volumes where all groundtruth vertebrae labels were shifted by up to five vertebrae, resulting in wrong landmark positions with more than 100 mm distance to the actually correct groundtruth position. For comparison with previously reported results, we show results on the original dataset (*Set2*-mislabeled). In agreement with the challenge organizers, we also report results on the corrected groundtruth (*Set2*-corrected), while the organizers have updated the challenge dataset accordingly.

regresses heatmaps for landmark localization, we achieve high localization performance even in the presence of limited training datasets. We have evaluated localization performance of our proposed SCN method on four size-limited datasets that contain 2D radiographs and 3D MRIs and CTs of different anatomical structures, demonstrating the generic applicability of our proposed method. Our results show that when training on 43 hand MRI volumes (Section 4.2), 112 spine CT images (Section 4.4), 150 skull radiographs (Section 4.3), or 600 hand radiographs (Section 4.1), our method outperforms the state-of-the-art approaches. In Section 4.1.2, where we purposely limited the 2DHand training images up to an extremely low number of 10 images, our proposed network still achieves results that are comparable to the state-of-the-art, when evaluated on the same images as in Section 4.1.1 (i.e., 2DHandFull).

Inspired by the use of prior knowledge to constrain landmark configurations to anatomically feasible ones, in our proposed SCN constraints on the relative position of landmarks are automatically learned from training data and integrated inside its *spatial configuration* component. Previous state-of-the-art approaches in medical image analysis (Donner et al., 2013; Lindner et al., 2015) introduce appearance features and handcrafted models resembling anatomical constraints in separate components. Instead of using a handcrafted model, in (Urschler et al., 2018) we have shown a possibility of learning the automatic integration of appearance information and geometric configuration into a single random forest framework for localization. Nevertheless, differently to all these previous methods that require the two separate components to be trained sequentially without any interaction between them, in our end-to-end trained SCN, the *local appearance* and *spatial configuration* components are simultaneously optimized, together providing both high robustness towards landmark misidentification as well as high accuracy locally at each identified landmark.

By simultaneously optimizing the two components, the problem of landmark localization is separated into two simpler subproblems that can be modeled from a low amount of training data, as shown in our experiments. Such a simplification is only possible when the regression objective, which is calculated by multiplying the output of the two network components, is optimized in a single process and in an end-to-end manner. This is different to the methods of Tompson et al. (2014) and Pfister et al. (2015), where each of their network components has to be locally accurate and robust towards misidentification simultaneously. It is important to notice that when the multiplication given in (4) is replaced by additional convolutional layers, as Pfister et al. (2015) used for human pose estimation in videos, this does not lead to the simplification of the localization problem, thus the need for large amounts of training data remains. Furthermore, our proposed network architecture learns to dedicate the *local appearance* component to locally accurate candidate predictions, without the need to distinguish locally similar structures, while the *spatial configuration* component solely focuses on eliminating ambiguities to improve robustness towards landmark misidentification, without the need for being locally accurate (see Fig. 9). Although there is no theoretical guarantee that the optimization process will lead to such a separation, we have observed this behavior in all our experiments, and we have shown it in more detail for the 2DHandFull experiment in Section 4.1.3.

Intrigued by the recent success of the fully convolutional U-Net (Ronneberger et al., 2015) in segmentation problems, we adapted this architecture for landmark localization using heatmap regression. After extensively tuning it for this task, our Localization U-Netachieved competitive localization performance in the experiments, often being outperformed only by our proposed SCN. However, when evaluated only regarding robustness towards landmark misidentification, Localization U-Netis not showing state-of-the-art results. This can be seen especially in the 3DSpine experiment with its large anatomical and pathological variation, where our SCN is 8.5% better in the $\mathrm{ID}_{\mathrm{rate}}$ evaluation. We think that this drop in performance is due to the multi-scale Localization U-Netarchitecture not using the prior knowledge that landmarks are not uniformly distributed in image space but are constrained by other anatomical landmarks. Thus, without prior knowledge of the existence of such constraints on *spatial configuration*, Localization U-Netrequires a large amount of training data to learn these constraints on multiple scales solely from the data. This can best be seen from the 3DSpine and 2DHandReduced experiments, where our proposed SCN benefits from this prior knowledge when learning the same anatomical variation from a limited amount of training data.

When comparing our SCN to other state-of-the-art approaches, on our two in-house datasets 2DHandFull and 3DHand we outperform all our previously reported random forest based results (Ebner et al., 2014; Štern et al., 2016a; Urschler et al., 2018), and also our preliminary results based on CNNs (Payer et al., 2016). Moreover, on the 2DHandFull dataset, we show better results than the current state-of-the-art localization method based on random forests of Lindner et al. (2015), who applied their method on the same cross-validation split, using our landmark annotation. On the 2DSkull dataset, which was used for two public challenges in previous years, we also outperform both challenge winners, Lindner et al. (2015) as well as Ibragimov et al. (2014). For 3DSpine, Chen et al. (2015) use both random forest and CNN predictions to significantly improve results compared to the pure random forest based method of Glocker et al. (2013). Similar results were obtained by the complex method of Yang et al. (2017) using only CNNs. Recently, they both were outperformed by CNN methods that are highly tailored to this dataset (Sekuboyina et al., 2018; Liao et al., 2018). However, our SCN outperforms all these methods in terms of landmark localization error on this challenging dataset, without the need for complex or tailored implementations.

In conclusion, we have shown how to combine information of *local appearance* and *spatial configuration* into a single end-to-end trained network for anatomical landmark localization. Our generic architecture does not require any postprocessing step for achieving state-of-the-art results in terms of landmark localization error on different 2D and 3D datasets, even when limited amounts of training images are available. In future work, we intend to compare our SCN with the recently presented Attention U-Net (Oktay et al., 2018) that shows some similarities in suppressing irrelevant features for segmentation tasks. Furthermore, we are currently looking into extending our SCN regarding occluded structures and multi-object localization, and into adapting our SCN for semantic segmentation problems (see Payer et al., 2018 for preliminary results), where structural constraints may be used in a similar manner.

Conflict of interest

None.

Acknowledgments

We thank all four anonymous reviewers for their thorough work and excellent comments that helped us to improve our manuscript. Moreover, we thank Claudia Lindner for applying their algorithm on our 2DHand dataset and Thomas Ebner for aiding in annotations of our datasets. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used in this research. This work was supported by the Austrian Science Fund (FWF): P28078-N33.

References

Beichel, R., Bischof, H., Leberl, F., Sonka, M., 2005. Robust active appearance models and their application to medical image analysis. IEEE Trans. Med. Imaging 24 (9), 1151–1169. doi:10.1109/TMI.2005.853237.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. doi:10.1023/A:1010933404324.

Chen, H., Shen, C., Qin, J., Ni, D., Shi, L., Cheng, J.C., Heng, P.A., 2015. Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks. In: Proc. Med. Image Comput. Comput. Interv., pp. 515–522. doi:10.1007/978-3-319-24553-9_63.

Ciresan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: Proc. Med. Image Comput. Comput. Interv., pp. 411–418. doi:10.1007/978-3-642-40763-5_51.

Cootes, T., Taylor, C., Cooper, D., Graham, J., 1995. Active shape models - Their training and application. Comput. Vis. Image Underst. 61 (1), 38–59. doi:10.1006/cviu.1995.1004.

Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., Siddiqui, K., 2013. Regression forests for efficient anatomy detection and localization in computed tomography scans. Med. Image Anal. 17 (8), 1293–1303. doi:10.1016/j.media.2013.01.001.

Donner, R., Menze, B.H., Bischof, H., Langs, G., 2013. Global localization of 3D anatomical structures by pre-filtered hough forests and discrete optimization. Med. Image Anal. 17 (8), 1304–1314. doi:10.1016/j.media.2013.02.004.

Ebner, T., Štern, D., Donner, R., Bischof, H., Urschler, M., 2014. Towards automatic bone age estimation from MRI: localization of 3D anatomical landmarks. In: Proc. Med. Image Comput. Comput. Interv. Springer, pp. 421–428. doi:10.1007/978-3-319-10470-6_53.

Felzenszwalb, P.F., Huttenlocher, D.P., 2005. Pictorial structures for object recognition. Int. J. Comput. Vis. 61 (1), 55–79. doi:10.1023/B:VISI.0000042934.15159.49.

Ghesu, F.C., Georgescu, B., Grbic, S., Maier, A., Hornegger, J., Comaniciu, D., 2018. Towards intelligent robust detection of anatomical structures in incomplete volumetric data. Med. Image Anal. 48, 203–213. doi:10.1016/j.media.2018.06.007.

Ghesu, F.C., 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 Trans. Pattern Anal. Mach. Intell. 41, 176–189. doi:10.1109/TPAMI.2017.2782687.

Glocker, B., Feulner, J., Criminisi, A., Haynor, D.R., Konukoglu, E., 2012. Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans.. In: Proc. Med. Image Comput. Comput. Interv., 15, pp. 590–598. doi:10.1007/978-3-642-33454-2_73.

Glocker, B., Zikic, D., Konukoglu, E., Haynor, D.R., Criminisi, A., 2013. Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In: Proc. Med. Image Comput. Comput. Interv., pp. 262–270. doi:10.1007/978-3-642-40763-5_33.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proc. Int. Conf. Comput. Vis.. IEEE, pp. 1026–1034. doi:10.1109/ICCV.2015.123.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. Comput. Vis. Pattern Recognit.. IEEE, pp. 770–778. doi:10.1109/CVPR.2016.90.

Heimann, T., Meinzer, H.P., 2009. Statistical shape models for 3D medical image segmentation: a review. Med. Image Anal. 13 (4), 543–563. doi:10.1016/j.media.2009.05.004.

Ibragimov, B., Likar, B., Pernuš, F., Vrtovec, T., 2012. A game-theoretic framework for landmark-based image segmentation. IEEE Trans. Med. Imaging 31 (9), 1761–1776. doi:10.1109/TMI.2012.2202915.

Ibragimov, B., Likar, B., Pernuš, F., Vrtovec, T., 2014. Shape representation for efficient landmark-based segmentation in 3-D. IEEE Trans. Med. Imaging 33 (4), 861–874. doi:10.1109/TMI.2013.2296976.

Johnson, H.J., Christensen, G.E., 2002. Consistent landmark and intensity-based image registration. IEEE Trans. Med. Imaging 21 (5), 450–461. doi:10.1109/TMI.2002.1009381.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Adv. Neural Inf. Process. Syst., pp. 1097–1105.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. doi:10.1038/nature14539.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2323. doi:10.1109/5.726791.

Liao, H., Mesfin, A., Luo, J., 2018. Joint vertebrae identification and localization in spinal CT images by combining short- and long-range contextual information. IEEE Trans. Med. Imaging 37 (5), 1266–1275. doi:10.1109/TMI.2018.2798293.

Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.F., 2015. Robust and accurate shape model matching using random forest regression-voting. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1862–1874. doi:10.1109/TPAMI.2014.2382106.

Lindner, C., Wang, C.-W., Huang, C.-T., Li, C.-H., Chang, S.-W., Cootes, T.F., 2016. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. Sci. Rep. 6, 33581. doi:10.1038/srep33581.

Liu, D., Zhou, K.S., Bernhardt, D., Comaniciu, D., 2010. Search strategies for multiple landmark detection by submodular maximization. In: Proc. Comput. Vis. Pattern Recognit., pp. 2831–2838. doi:10.1109/CVPR.2010.5540016.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In: Proc. Int. Conf. Mach. Learn., 28, pp. 1–6.

Nesterov, Y., 1983. A method of solving a convex programming problem with convergence rate O(1/k^2). In: Sov. Math. Dokl., 27, pp. 372–376.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: learning where to look for the pancreas. In: Proc. 1st Conf. Med. Imaging with Deep Learn. arXiv abs/1804.03999.

Payer, C., Štern, D., Bischof, H., Urschler, M., 2016. Regressing heatmaps for multiple landmark localization using CNNs. In: Proc. Med. Image Comput. Comput. Interv.. Springer, pp. 230–238. doi:10.1007/978-3-319-46723-8_27.

Payer, C., Štern, D., Bischof, H., Urschler, M., 2018. Multi-label whole heart segmentation using CNNs and anatomical label configurations. In: Stat. Atlases Comput. Model. Hear. ACDC MMWHS Challenges. STACOM 2017. Springer, pp. 190–198. doi:10.1007/978-3-319-75541-0_20.

Pfister, T., Charles, J., Zisserman, A., 2015. Flowing ConvNets for human pose estimation in videos. In: Proc. Int. Conf. Comput. Vis., pp. 1913–1921. doi:10.1109/ICCV.2015.222.

Potesil, V., Kadir, T., Platsch, G., Brady, M., 2015. Personalized graphical models for anatomical landmark localization in whole-body medical images. Int. J. Comput. Vis. 111 (1), 29–49. doi:10.1007/s11263-014-0731-7.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: Proc. Med. Image Comput. Comput. Interv. Springer, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252. doi:10.1007/s11263-015-0816-y.

Sekuboyina, A., Rempfler, M., Kukačka, J., Tetteh, G., Valentinitsch, A., Kirschke, J.S., Menze, B.H., 2018. Btrfly Net: vertebrae labelling with energy-based adversarial learning of local spine prior. In: Proc. Med. Image Comput. Comput. Interv. Springer, pp. 649–657. doi:10.1007/978-3-030-00937-3_74.

Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (4), 640–651. doi:10.1109/TPAMI.2016.2572683.

Sirinukunwattana, K., Pluim, J.P.W., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B.B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D.R.J., Rajpoot, N.M., 2017. Gland segmentation in colon histology images: the glas challenge contest. Med. Image Anal. 35, 489–502. doi:10.1016/j.media.2016.08.008.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Štern, D., Ebner, T., Urschler, M., 2016. From local to global random regression forests: exploring anatomical landmark localization. In: Proc. Med. Image Comput. Comput. Interv.. Springer, pp. 221–229. doi:10.1007/978-3-319-46723-8_26.

Štern, D., Likar, B., Pernuš, F., Vrtovec, T., 2011. Parametric modelling and segmentation of vertebral bodies in 3D CT and MR spine images. Phys. Med. Biol. 56 (23), 7505–7522. doi:10.1088/0031-9155/56/23/011.

Štern, D., Payer, C., Lepetit, V., Urschler, M., 2016. Automated age estimation from hand MRI volumes using deep learning. In: Proc. Med. Image Comput. Comput. Interv.. Springer, pp. 194–202. doi:10.1007/978-3-319-46723-8_23.

Toews, M., Arbel, T., 2007. A statistical parts-based model of anatomical variability. IEEE Trans. Med. Imaging 26 (4), 497–508. doi:10.1109/TMI.2007.892510.

Tompson, J., Jain, A., LeCun, Y., Bregler, C., 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In: Adv. Neural Inf. Process. Syst., pp. 1799–1807.

Toshev, A., Szegedy, C., 2014. DeepPose: human pose estimation via deep neural networks. In: Proc. Comput. Vis. Pattern Recognit., pp. 1653–1660. doi:10.1109/CVPR.2014.214.

Urschler, M., Ebner, T., Štern, D., 2018. Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. Med. Image Anal. 43, 23–36. doi:10.1016/j.media.2017.09.003.

Urschler, M., Zach, C., Ditt, H., Bischof, H., 2006. Automatic point landmark matching for regularizing nonlinear intensity registration: application to thoracic CT images. In: Proc. Med. Image Comput. Comput. Interv., pp. 710–717. doi:10.1007/11866763_87.

Wang, C.W., Huang, C.T., Lee, J.H., Li, C.H., Chang, S.W., Siao, M.J., Lai, T.M., Ibragimov, B., Vrtovec, T., Ronneberger, O., Fischer, P., Cootes, T.F., Lindner, C., 2016. A benchmark for comparison of dental radiography analysis algorithms. Med. Image Anal. 31, 63–76. doi:10.1016/j.media.2016.02.004.

Yang, D., Xiong, T., Xu, D., Huang, Q., Liu, D., Zhou, S.K., Xu, Z., Park, J.H., Chen, M., Tran, T.D., Chin, S.P., Metaxas, D., Comaniciu, D., 2017. Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization. In: Proc. Inf. Process. Med. Imaging, pp. 633–644. doi:10.1007/978-3-319-59050-9_50.

Zhang, J., Liu, M., Shen, D., 2017. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. IEEE Trans. Image Process. 26 (10), 4753–4764. doi:10.1109/TIP.2017.2721106.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S., 2015. Conditional random fields as recurrent neural networks. Proc. Int. Conf. Comput. Vis. 1529–1537. doi:10.1109/ICCV.2015.179.