# Latent Shape Constraint for Anatomical Landmark Detection on Spine Radiographs

Florian Kordon[1,2,3], Andreas Maier[1,2,4], Holger Kunze[1,3]

[1]Pattern Recognition Lab, Universität Erlangen-Nürnberg (FAU), Erlangen
[2]Erlangen Graduate School in Advanced Optical Technologies (SAOT), Universität Erlangen-Nürnberg (FAU), Erlangen
[3]Siemens Healthcare GmbH, Forchheim
[4]Machine Intelligence, Universität Erlangen-Nürnberg (FAU), Erlangen
florian.kordon@fau.de

**Abstract.** Vertebral corner points are frequently used landmarks for a vast variety of orthopedic and trauma surgical applications. Algorithmic approaches that are designed to automatically detect them on 2D radiographs have to cope with varying image contrast, high noise levels, and superimposed soft tissue. To enforce an anatomically correct landmark configuration in presence of these limitations, this study investigates a shape constraint technique based on data-driven encodings of the spine geometry. A contractive PointNet autoencoder is used to map numerical landmark coordinate representations onto a low-dimensional shape manifold. A distance norm between prediction and ground truth encodings then serves as an additional loss term during optimization. The method is compared and evaluated on the SpineWeb16 dataset. Small improvements can be observed, recommending further analysis of the encoding design and composite cost function.

## 1 Introduction

Anatomical landmark localization is an important prerequisite in medical image processing. It mostly serves as a semantic prior for subsequent tasks which operate on single-point information and their spatial relationships [1,2]. Traditionally, strategies to assess such a localization task either involve independent detection of a single landmark or incorporate information about relative positioning, spatial constraints, a priori knowledge, and characteristics of the local feature vicinity [2,3]. In contrast to natural images, this additional information often presents itself as a natural choice to alleviate image-quality based ambiguities. This is due to anatomical landmarks following a rather rigid configuration constrained by the bio-mechanical range of motion of the human body.

A typical example of such a configuration can be observed for the localization of vertebral corner points on antero-posterior spine radiographs. On the micro level, the vertebrae within one spinal region are very similar to each other

and describe a convex quadrilateral by their corner vertices. On the macro level, the vertical progression of all vertebrae on the X-ray image can be described by a plane curve. In the case of healthy patients, this curve approximates a straight line, and in cases of more severe scoliosis, it is more bent. If we want to translate these anatomical constraints for automatic localization methods using Deep Learning, we can distinguish between two basic approaches: (1) *Explicit Constraints* using domain knowledge, i.e. enforcing adjacency constraints or geometric rules [4], and (2) *Implicit Constraints* using a data-driven extraction of statistics about plausible and aberrant configurations [3,5].

While using an explicit constraints scheme is an attractive approach, it is limited in its ability to generalize well to unseen data and to data that does not match the underlying geometric model. In contrast, a data-driven approach in theory can capture these variations, but presumes a sufficiently large training data corpus to do so. However, depending on the specific methodology such implicit constraints might require additional topological modifications to the learning model's architecture. To circumvent this additional computational effort during inference, [5] proposed a lightweight shape-aware method which they evaluated for segmentation and super-resolution tasks. The latent representation of an autoencoder is used to map both the ground truth as well as the neural network predictions onto a low-dimensional manifold. During optimization, the distance between both encodings serves as additional constraint to the cost function to pull back aberrant predictions onto anatomically accurate solutions.

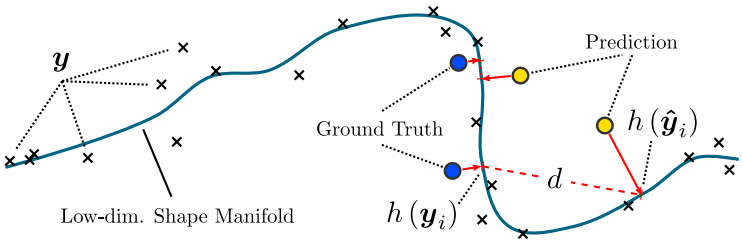
Based on this idea, this study investigates an extension of this method for a joint coordinate representation of vertebral corner points. After estimating the spatial position for each of the 68 corner points using 2D heatmaps, normalized numerical coordinates are extracted using a differentiable *spatial-to-numerical* transform (DSNT) [6]. A shape representation of this coordinate set is then extracted as the latent encoding of a PointNet-Autoencoder and compared to the encoded shape of the corresponding ground truth. The influence of the proposed constraint is evaluated on the SpineWeb16 dataset [3].

## 2 Materials and methods

### 2.1 Geometric constraint via latent shape encoding

A common way to represent data as a set of discriminative and abstract features is the use of autoencoders. An autoencoder can learn to map the data features to a low-dimensional manifold in an unsupervised fashion by encoding and decoding a latent (vector) representation $h(\cdot)$ and measuring the quality of the input reconstruction. If we consider a set of representative landmark configurations of the vertebral corner points $\boldsymbol{y}$, such an autoencoder optimizes an abstraction of possible and anatomically correct spine shapes from which it can reconstruct all individual corner points. Consequently, if the distance between two latent vectors of a ground truth sample $h\left(\boldsymbol{y}_i\right)$ and a test sample $h\left(\tilde{\boldsymbol{y}}_i\right)$ is large, it can be assumed that the test sample describes a shape that is not suitable for decoding and anatomically aberrant (Fig. 1).

**Fig. 1.** Illustration of a low-dimensional shape manifold based on ground truth $\boldsymbol{y}$ and the learned mapping function $h(\cdot)$.

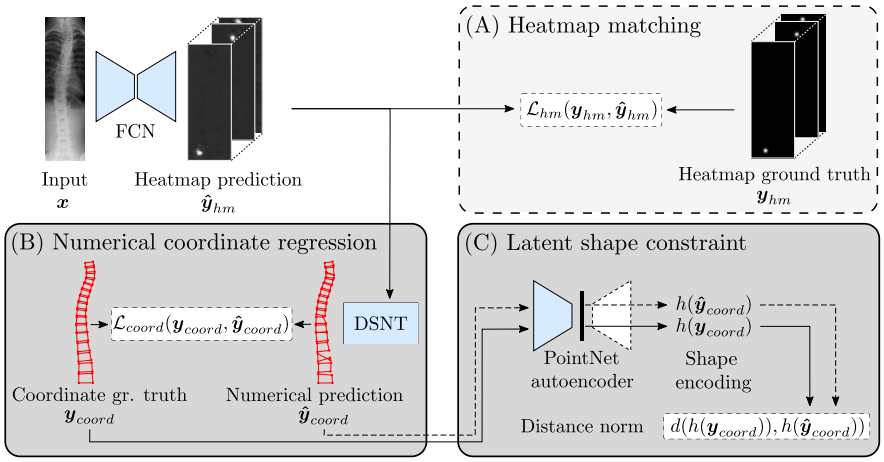



This idea of a latent distance measure naturally translates to an additional loss term within the cost function of an optimization problem [5]. Given the cost function for some optimization task $\mathcal{L}_{opt}$, a distance norm $d$ (here L2-norm) on the latent vectors, and a weighting term $\lambda$, this can be described as $\mathcal{L}_{total} = \mathcal{L}_{opt} + \lambda \cdot d\left(h\left(\boldsymbol{y}\right)\right), h\left(\tilde{\boldsymbol{y}}\right))$.

### 2.2 Architecture design variants

A popular way to estimate the position of 2D landmarks is to encode their spatial likelihood as heatmaps. This is achieved by placing a 2D Gaussian distribution with standard deviation $\sigma$ and compact support $\pm 3\sigma$ on the landmark coordinates. The heatmaps are then estimated with a Fully-Convolutional Neural Network (FCN) and explicitly compared to the ground truth via image-to-image matching using a mean-squared error cost function (Fig. 2-A) [1,7]. If we consider a computationally demanding task such as the prediction of a large set of corner points, the integration of the introduced shape constraint necessitates a comparably large autoencoder to encode all predicted heatmaps. To reduce this computational footprint, we employ a *spatial-to-numerical* transform [6] to map the heatmaps to normalized numerical coordinates. In contrast to a spatial argmax operation which is used to obtain the maximum response in case of heatmap matching, this transform is completely differentiable. This means that a cost functions can directly operate on the numerical coordinates while implicitly optimizing the heatmaps. Besides, using a numerical coordinate representation as input and target for the autoencoder allows for a lightweight autoencoder architecture (Fig. 2-B,C). To be able to find meaningful shape representations also in case of noisy label data, we propose to use a contractive autoencoder topology inspired by PointNet [8]. Given a landmark configuration in the form of a $(N, 2)$ tensor with $N$ marking the number of landmarks, the input is first transformed to $(N, 1)$ using a number of convolutional blocks before a symmetry function is applied. Besides performing the final encoding step, this symmetry function (here max-pooling) makes the autoencoder invariant to the order of input landmarks. We assume this characteristic to aid in learning a global shape context and implicit geometric adjacency relations.

**Fig. 2.** Visualization of the model variants that are compared for the task of vertebral corner point detection. (A) Explicit heatmap prediction with subsequent spatial argmax. (B) Implicit heatmap prediction with *spatial-to-numerical* transform DSNT [6]. (C) Additional shape constraint penalizing the latent vector distance.




### 2.3 Data and experiment protocol

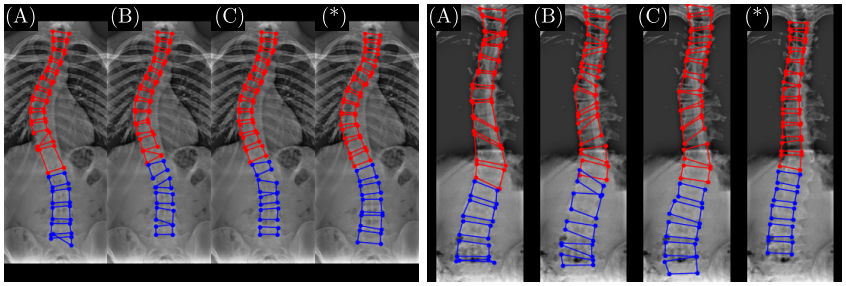
The comparison of the architecture variants is performed on the SpineWeb16 dataset which consists of 609 spinal antero-posterior radiographs [3] with labels for 12 thoracic and 5 lumbar vertebrae, totaling 68 vertebral corner points. An initial semi-automatic screening of the annotation data revealed point and vertebrae permutations, shifting of upper and lower plates, as well as clinically implausible corners in areas of low contrast. A subset of 520 suitable images was selected after automatic and manual corrections of the annotation material[1]. For all variants, we devised a 5(+1)-fold cross-validation scheme. The extra sixth fold was selected for an initial optimization of the shape constraint weight $\lambda \in [0.001, 0.003, 0.006, 0.010, ..., 1.000]$ and was used as additional training data in the subsequent cross-validation. As a main model for heatmap prediction we trained a single Hourglass module [7] with a feature root of 256 and additional instance normalization layers [1]. For the shape-constrained variant (C), a PointNet autoencoder was pre-trained for the corresponding fold configuration and included via an additional cost term with optimized weighting factor $\lambda = 0.003$ (Subsec. 2.1). For every architecture variant, the images were down-sampled and zero-padded to a common spatial resolution of [h:512 × w:192] px. Every image was processed with a homomorphic filtering operation to increase the image contrast in low-intensity areas and min-max normalized to the interval of $[0, 1]$. During training, an online augmentations scheme with randomized Gaussian blurring, contrast scaling, rotation, translation, scaling, and slight axis-aligned shrinking was applied. Training was performed for 250 epochs and the best model was selected based on the performance on the validation set.

[1] Curated annotations are available at doi:10.5281/zenodo.4413665.

**Table 1.** Cross-validation results for the three architecture variants ((A),(B),(C)) and autoencoder reconstruction (*). The average and max scores are reported across the five folds and are based on the individual mean/max of the Euclidean distance (ED) over all samples within the respective test fold. The reported scores are scaled w.r.t. the original image size and normalized to a reference spatial resolution of [h:1000 × w:1000] px.

| Architecture variant | Average ED (px) | Max ED (px) |
|---|---|---|
| (A) Heatmap matching | $23.23 \pm 1.47$ | 92.73 |
| (B) Num. regression | $21.67 \pm 1.48$ | 104.03 |
| (C) Num. regression + Shape constraint | $21.25 \pm 1.24$ | 84.55 |
| (*) Autoencoder reconstruction | $45.57 \pm 4.84$ | 121.20 |

**Fig. 3.** Example predictions of a successful shape constraint (a) and a failure case (b). (*) marks the reconstruction obtained by the PointNet autoencoder. Red and blue coloring denotes the thoracic and lumbar spinal section respectively.




(a) (b)

## 3 Results

As presented in Tab. 1, small performance gains can be observed when using a numerical coordinate representation (B) or shape constraint variant (C). The magnitude of the improvement however does suggest the benefits to not be significant. When analyzing the qualitative results, the shape constraint yields improvements for a subset of images with overall good image contrast (Fig. 3). However, in case of more severe contrast differences and obscured image parts due to superimposed soft tissue, neither the DSNT variant nor the shape constraint approach benefit the landmark predictions. For such image characteristics in general, no model variant yields anatomically accurate landmark configurations. Interestingly, the autoencoder reconstruction frequently estimates much smoother landmark configurations at the cost of spatial precision.

## 4 Discussion

The analysis shows that while the proposed shape constraint on average benefits the spinal shape, no consistent improvements w.r.t. the positional quality of the vertebral corner points can be achieved, especially in the case of low-quality images. Based on the findings of the qualitative analysis, several potential problem

factors can be identified. Although the quality of the autoencoder reconstruction indicates that a mapping function onto a meaningful shape manifold can be learned, the abstraction towards a probabilistic mean shape could be too strong to actually enforce geometrically meaningful landmarks. This assumption is supported by a small optimal weighting factor $\lambda$ and a general pull back to a rectified spine shape (Fig. 3). Such behavior also relates to the observations by [3] who showcase that standard convolutional features do not warrant sufficiently accurate landmark positions. Also, a linear composition of the cost function might cause a suboptimal optimization landscape due to conflicting gradients, which could be solved by either gradient manipulation [9] or an adaptive weighting scheme [1]. And lastly, adaptive pre-processing based on region-specific image statistics could help to alleviate image quality based ambiguities which often occur in the thoracic region. With these limitations in mind, we seek to extend our evaluation and combine a data-driven approach with explicit shape constraints.

**Acknowledgement.** The authors gratefully acknowledge funding of the Erlangen Graduate School in Advanced Optical Technologies (SAOT) by the Bavarian State Ministry for Science and Art.

**Disclaimer.** The methods and information presented here are based on research and are not commercially available.

## References

1. Kordon F, Fischer P, Privalov M, et al. Multi-task localization and segmentation for X-Ray guided planning in knee surgery. In: Shen D, Liu T, Peters TM, et al., editors. Proc MICCAI. Springer; 2019. p. 622–630.
2. Urschler M, Ebner T, Štern D. Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. Med Image Anal. 2018;43:23–36.
3. Wu H, Bailey C, Rasoulinejad P, et al. Automatic landmark estimation for adolescent idiopathic scoliosis assessment using BoostNet. In: Descoteaux M, Maier-Hein L, Franz A, et al., editors. Proc MICCAI. Springer; 2017. p. 127–135.
4. Imran AAZ, Huang C, Tang H, et al.. Bipartite distance for shape-aware landmark detection in spinal X-Ray images; 2020. arXiv:2005.14330v1 [eess.IV].
5. Oktay O, Ferrante E, Kamnitsas K, et al. Anatomically constrained neural networks (ACNN): Application to cardiac image enhancement and segmentation. IEEE Trans Med Imaging. 2018;37(2):384–395.
6. Nibali A, He Z, Morgan S, et al. Numerical coordinate regression with convolutional neural networks. CoRR. 2018;abs/1801.07372.
7. Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. In: Leibe B, Matas J, Sebe N, et al., editors. Proc ECCV. Springer; 2016. p. 483–499.
8. Charles RQ, Su H, Kaichun M, et al. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit; 2017. p. 77–85.
9. Yu T, Kumar S, Gupta A, et al.. Gradient surgery for multi-task learning; 2020. arXiv:2001.06782v3 [cs.LG].