Btrfly Net: Vertebrae Labelling with Energy-based Adversarial Learning of Local Spine Prior

Anjany Sekuboyina, Markus Rempfler, Jan Kukačka, Giles Tetteh, Alexander Valentinitsch, Jan S. Kirschke*, and Bjoern H. Menze*

Technische Universität München, Munich, Germany

Abstract. Robust localisation and identification of vertebrae is an essential part of automated spine analysis. The contribution of this work to the task is two-fold: (1) Inspired by the human expert, we hypothesise that a sagittal and coronal reformation of the spine contain sufficient information for labelling the vertebrae. Thereby, we propose a butterfly-shaped network architecture (termed Btrfly Net) that efficiently combines the information across the reformations. (2) Underpinning the Btrfly net, we present an energy-based adversarial training regime that encodes the local spine structure as an anatomical prior into the network, thereby enabling it to achieve state-of-art performance in all standard metrics on a benchmark dataset of 302 scans without any post-processing during inference.

1 Introduction

The localisation and identification of anatomical structures is a significant part of any medical image analysis routine. In spine's context, labelling of vertebrae has immediate diagnostic and modelling significance, e.g.: localised vertebrae are used as markers for detecting kyphosis or scoliosis, vertebral fractures, in surgical planning, or for follow-up analysis tasks such as vertebral segmentation or their bio-mechanical modelling for load analysis.

Vertebrae labelling. Like several analysis approaches off-late, vertebrae labelling has seen successful utilisation of machine learning. One of the incipient and notable works by Glocker et al. [2], followed by [3] used context-based features with regression forests and Markov models for labelling. In spite of their intuitive motivation, these approaches suffer a setback due to limited FOVs or presence of metal insertions. On a similar footing, [7] proposed a deep multi-layer perceptron using long-range context features. With the emergence of convolutional neural networks (CNN), Chen et al. [1] proposed a joint-CNN as a combination of a random forest for initial candidate selection followed by a CNN trained to identify the vertebra based on its appearance and a conditional dependency on its neighbours. Without hand-crafting features this approach performed remarkably well. However, since the CNN works on a limited region around the

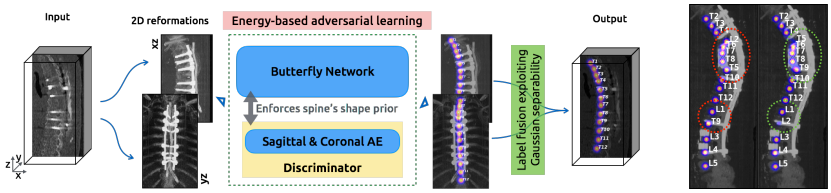
* Joint supervising authors.

vertebra, it results in a high variability of the localisation distance. Recently, Yang et al., with [8] and [9], proposed a deep, volumetric, fully-convolutional 3D network (FCN) called DI2IN with deep-supervision. The output of DI2IN is improved in subsequent stages that employ either message-passing across channels or a convolutional LSTM followed by further tuning with a shape dictionary.

Owing to the equivariance of the convolutional operator and limited receptive field, an FCN doesn't always learn the anatomy of the region-of-interest. This is a severe limitation as human-equivalent learning includes utilisation of anatomical details aided with prior knowledge. An immediate remedy is to increase the receptive field by going deeper, however this either comes at the cost of higher model complexity or is just unfeasible due to memory constraints when working with volumetric data. This opens an interesting research problem of encoding anatomical knowledge into the network during training.

Prior & adversarial learning in CNNs. Recent work in [5] and [4] propose encoding (anatomical) segmentation priors into an FCN by learning the shape representation using an auto encoder (AE). The segmentation is expressed in terms of a pre-learnt latent space for evaluating a prior-oriented loss, which is then used to guide the FCN into predicting an anatomically sound segmentation. Our approach shares similarities with this approach with certain fundamental differences: (1) Our approach is aimed at localisation, which requires a redefinition of the notion of anatomical *shape*. (2) We employ an AE for shape regularisation, but do not 'pre-train' it to learn the latent space. We train the AE adversarially in tandem with the FCN. Parallels can be drawn between end-to-end learning of priors and learning the distribution of priors using generative adversarial networks (GANs). Both have two networks, a predictor (generator) and an auxiliary network which works on the 'goodness' of the prediction. In medical image analysis where the scan size is large and the data samples are few, inspired from an energy-based adversarial generation framework (Zhao et al., [11]), it is preferable to employ an adversary providing an anatomically-inspired supervision instead of the usual adversaries that supervise with a binary value (vanilla GAN) or Wasserstein distance.

Our contribution. In this work, we propose an end-to-end solution for vertebrae labelling by training an FCN in an adversarial fashion thereby encoding the local spine structure into it. More precisely, relying on the sufficiency of information in certain 2D projections of 3D data, we propose: (1) A butterfly-shaped network that operates on 2D sagittal and coronal reformations, combining information across these views at a large receptive field, (2) Encoding the spine's structure into the Btrfly net using an energy-based, fully-convolutional, adversarial auto encoder acting as a discriminator. Our approach attains identification rates above 85% without any post-processing stages, achieving state-of-art performance.



(a) (b)

Fig. 1: (a) Overview of our approach. (b) Label correcting capability of the AE when trained as a denoising convolutional auto-encoder (red: corrupted, green: corrected). This motivates the discriminator in our adversarial framework.

2 Methodology

We present our approach in two stages. First, we describe the Btrfly network tasked with the labelling of the vertebrae. Then we present the adversarial learning of the local spine shape with an energy-based auto-encoder acting as the discriminator. Fig. 1a gives an overview of the proposed approach and the motivation for prior-encoding is illustrated in Fig. 1b.

2.1 Btrfly Network

Working with 3D volumetric data is computationally restrictive, more so for localisation and identification that rely on a large context so as to capture spatially distant landmarks. Consequently, there is a trade-off between working with low-resolution data or resorting to shallow networks. Therefore, we propose working in 2D with *sufficiently–representative* projections of the volumetric data. The choice of projection is application dependant. Since we are working with bone, we work on sagittal and coronal maximum intensity projections (MIP). The former captures the spine's curve and the latter captures the rib-vertebrae joints, both of which are crucial markers for labelling. Note that a naive MIP might not always be the optimal choice of projection. Such cases are handled with a pre-processing stage detailed in the Supplement.

Annotations. We formulate the problem of learning the vertebrae labels as a multi-variate regression. The ground-truth annotation $\mathbf{Y} \in \mathbb{R}^{(h \times w \times d \times 25)}$ is a 25-channeled, 3D volume with each channel corresponding to each of the 24 vertebrae (C1 to L5), and one for the background. Each channel i is constructed as a Gaussian heat map of the form $\mathbf{y}_i = e^{-||x-\mu_i||^2/2\sigma^2}$, $x \in \mathbb{R}^3$ where μ_i is the location of the i^{th} vertebra and σ controls the spread. The background channel is constructed as, $\mathbf{y}_0 = 1 - \max_i(\mathbf{y}_i)$. The sagittal and coronal MIPs of $\mathbf{Y}$ are denoted by $\mathbf{Y}_{\text{sag}} \in \mathbb{R}^{(h \times w \times 25)}$ and $\mathbf{Y}_{\text{cor}} \in \mathbb{R}^{(h \times d \times 25)}$, respectively.

Architecture. We employ an FCN to perform the task of labelling. Since essential information is contained in both the sagittal and coronal reformations, and since the spine is approximately spatially centred in both, exchange of this information across views leads to an improved identification. We propose a

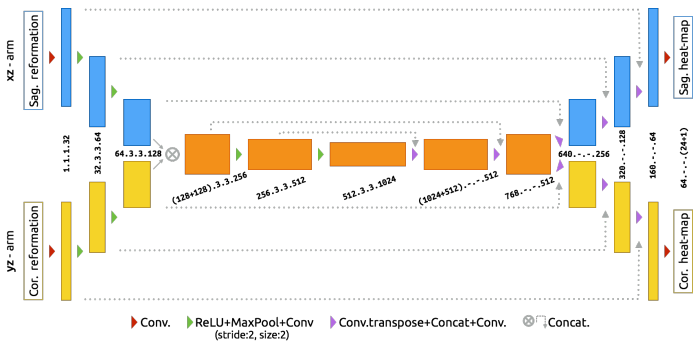


Fig. 2: The Btrfly architecture. The xz- (blue) and the yz-arms (yellow) correspond to the sagittal and coronal views. The kernel's shape resulting in each of the blocks is indicated as: {input channels} · {kern. height} · {kern. width} · {output channels}

butterfly-like network (cf. Fig 2) with two arms (xz- and yz-arms) each concerned with one of the views. The feature maps of both the views are combined after a certain depth in order to learn their inter-dependency. The x and y dimensions could be made equal by zero-padding the smaller dimension.

Loss. We choose an ℓ_2 distance as the primary loss supported by a cross-entropy loss over the softmax excitation of the ground truth and the prediction. The total loss is expressed as:

$$\mathcal{L}_{\text{b,sag}} = ||\mathbf{Y}_{\text{sag}} - \tilde{\mathbf{Y}}_{\text{sag}}||^2 + \omega H(\mathbf{Y}^{\sigma}_{\text{sag}}, \tilde{\mathbf{Y}}^{\sigma}_{\text{sag}}), \tag{1}$$

where $\tilde{\mathbf{Y}}_{\text{sag}}$ is the prediction of the net's xz-arm, H is the cross-entropy function, and $\mathbf{Y}^{\sigma}_{\text{sag}} = \sigma(\mathbf{Y}_{\text{sag}})$, the softmax excitation. ω is the median frequency weighing map (described in [6]), boosting the learning of less frequent classes. The loss for the yz-arm is constructed in a similar fashion and the total loss of the Btrfly net is given by $\mathcal{L}_{\text{b}} = \mathcal{L}_{\text{b,sag}} + \mathcal{L}_{\text{b,cor}}$.

2.2 Energy-based adversary for encoding prior

Since the Btrfly net is fully-convolutional, its predictions across voxels are independent of each other owing to the spatial invariance of convolutions. Whatever information it encodes is solely due to its receptive field, which may not be anatomically consistent across the image. We propose to impose the anatomical prior of the spine's shape onto the Btrfly net with *adversarial* learning.

Denoting the projected annotation as $\mathbf{Y}_{\text{view}}$, where view$\in$\{sag,cor\}, a sample annotation consists of a 2D Gaussian at the vertebral location in each channel (except $\mathbf{y}_0$). Looking at $\mathbf{Y}_{\text{view}}$ as a 3D volume enables us in learning the spread of Gaussians across channels and consequently the vertebral labels. However, owing to the extreme variability of FOVs and scan sizes, it is preferable to learn the spread of the vertebrae in parts. Therefore, we employ a fully-convolutional, 3D auto encoder (AE) with a receptive field covering a part of the spine at a

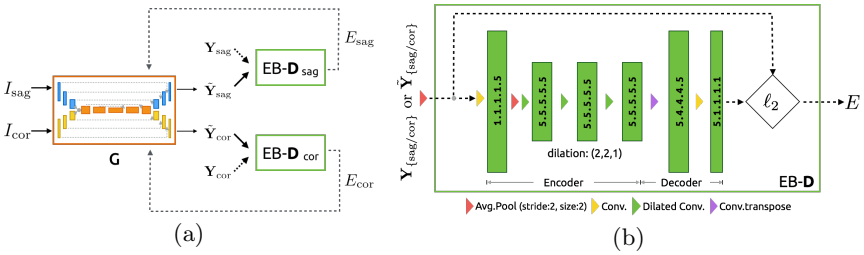


Fig. 3: (a) A overview of adversarial training showing the input to and the energy-based supervision signal from the discriminators. (b) The composition of the energy-based discriminator (EB-D). It gives the ℓ_2 reconstruction error as output.

time. The absence of fully-connected layers in the AE also removes the necessity to resize the data, making it end-to-end trainable with the Btrfly net. Fig. 3a shows the arrangement of the AEs as adversaries w.r.t the Btrfly net. In an adversarial framework, the Btrfly net acts as the generator (G), and the local manifolds learnt from $\mathbf{Y}_{\text{view}}$ influence $\tilde{\mathbf{Y}}_{\text{view}}$ and vice versa.

Discriminator. We devise the 3D adversary (D, cf. Fig. 3b) consisting of the AE as a functional predicting the ℓ_2 distance between the input $\mathbf{Y}_{\text{view}}$ and its reconstruction by the AE, $rec(\mathbf{Y}_{\text{view}})$: $D(\mathbf{Y}_{\text{view}}) = E = ||\mathbf{Y}_{\text{view}} - rec(\mathbf{Y}_{\text{view}})||$. This energy, E is fed back into G for adversarial supervision, as in [11]. As it is an energy-based functional, we interchangeably refer to the discriminator as EB-D. Since $\mathbf{Y}_{\text{view}}$ consists of Gaussians, it is less informative than an image. Therefore, we avoid using max-pooling by resorting to average pooling. In order to have a receptive field covering multiple vertebrae without using pooling operations, we employ spatially dilated convolution kernels [10] of size $(5 \times 5 \times 5)$ with a dilation rate of 2 (only in image plane), resulting in a receptive field of 76×76 pixels. At 1 mm isotropic resolution, this covers 2 to 3 vertebrae in the lumbar region and more elsewhere.

Losses. As in any adversarial setup, EB-D is shown real ($\mathbf{Y}_x (\equiv \mathbf{Y}_{\text{view}})$) and generated annotations ($\mathbf{Y}_g (\equiv \tilde{\mathbf{Y}}_{\text{view}})$), and it learns to discriminate between both by predicting a low E for real annotations, while G learns to generate annotations that would trick D. For a given positive, scalar margin m, the following generator and discriminator losses are optimised:

$$\mathcal{L}_D = D(\mathbf{Y}_x) + \max(0, m - D(\mathbf{Y}_g)), \text{ and} \tag{2}$$

$$\mathcal{L}_G = D(\mathbf{Y}_g) + \mathcal{L}_{\text{b,view}}. \tag{3}$$

The joint optimisation of (2) and (3) for both the EB-Ds results in a G that performs vertebrae labelling while respecting the spatial distribution of the vertebrae across channels. We refer to this prior-encoded G as the 'Btrfly$_{\text{pe}}$' net.

2.3 Inference

Once trained, an inference for a given input scan of size $(h \times w \times d)$ proceeds as: the desired sagittal and coronal MIP reformations are obtained and given

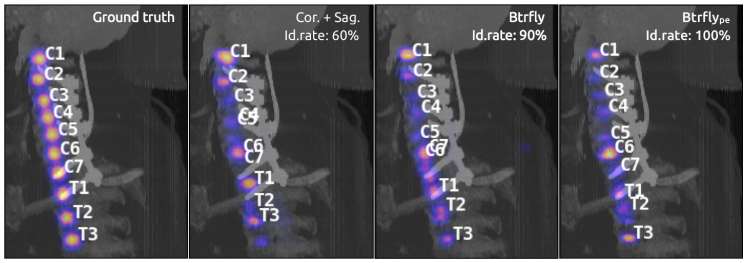


Fig. 4: Effect of prior encoding: the prior-encoded Btrfly$_{\text{pe}}$ net successfully performs its task of prevent overlapping labels (C6 & C7), consequently reordering all the vertebral labels. The reported id. rates are per volume. Additional results in Supplement.

as input to the xz- and yz-arms of the Btrfly net, resulting in a $(h \times w \times 25)$ sagittal heatmap and $(h \times d \times 25)$ coronal heatmap. The values below a threshold (T) are ignored in order to remove noisy predictions. As the Gaussian kernel is separable, an outer product of the predictions results in the final heat map as $\tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}_{\text{sag}} \otimes \tilde{\mathbf{Y}}_{\text{cor}}$, where $\otimes$ denotes the outer product. The 3D location of the vertebral centroids are obtained as the maxima in their corresponding channels. Note that the EB-D is no longer required during inference as its role in encoding the prior ends with the convergence of the Btrfly$_{\text{pe}}$ net.

3 Experiments

The evaluation is performed using a dataset introduced in [3] with a total of 302 CT scans (242 for training and 60 for testing) including various challenges such as scoliotic spines, metal insertions, and highly restrictive FOVs. However, these are cropped to a region around the spine which excludes the ribcage. Thus, a naive sagittal and coronal MIP suffices to obtain the input images for our approach. In order to enhance the net's robustness, 10 MIPs are obtained from one 3D scan, each time randomly choosing half the slices of interest. This leads to a total of 2420 reformations per view for training (incl. a validation split of 100). We present the experiments with the Btrfly net trained as stand-alone as well as with the prior-encoding discriminator EB-D. Batch-normalisation is used after every convolution layer, along with 20% dropout in the fused layers of Btrfly. Additionally, so as to validate the necessity of the combination of views, we compare the Btrfly net's performance with that of two networks working individually on the views (denoted by Cor.+Sag. nets). The architecture of each of these networks is similar to one arm of the Btrfly net. The optimiser's setup in all the three cases is similar: an Adam optimiser is employed with an initial learning rate of $\lambda = 1 \times 10^{-3}$, working on data resampled to a 1 mm isotropic resolution. λ is decayed by a factor of $3/4th$ every 10k iterations to 0.2×10^{-3}. The training of Cor.+Sag. nets and Btrfly net converges at 60k iterations. Due to an increased model complexity of joint networks, the Btrfly$_{\text{pe}}$ converges later

Measures	[3]	[1]	DI2IN[8]	DI2IN*[8]	Cor.+Sag.	Btrfly	**Btrfly$_{pe}$**
Id.rate	74.0	84.2	76.0	85.0	78.1	81.8	**86.1**
d_{mean}	13.2	8.8	13.6	8.6	9.3	7.5	**7.4**
d_{std}	17.8	13.0	37.5	7.8	8.0	**5.4**	9.3

Table 1: Performance evaluation and comparison of our approach (without thresholding) with Glocker et al. [3], Chen et al. [1], & Yang et al. [8]. DI2IN refers to stand-alone FCN, while DI2IN* includes use of message passing and shape dictionary. We do not compare with experiments in [8] that use additional undisclosed data.

(≈80k iterations). As suggested in [11], we initialise m with 20 and gradually reduce it to zero where the Nash equilibrium is attained.

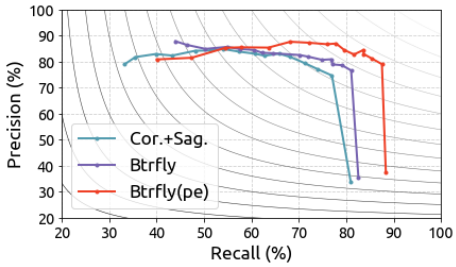


Fig. 5: A precision-recall curve with F1 isolines, illustrating the effect of the T during inference. For any T, Btrfly$_{pe}$ offers a better trade-off between P and R.

Approach	P	R	F1
Cor.+Sag.$_{(T=0.05)}$	74.7	77.0	75.8
Btrfly$_{(T=0.1)}$	78.7	79.1	78.9
Btrfly$_{pe}$ $_{(T=0.2)}$	**84.6**	**83.7**	**84.1**

Table 2: The optimal P and R values based on F1 score, along with the optimal T. R at optimal-F1 of Btrfly$_{pe}$ is comparable to state-of-art.

Evaluation & discussion. For evaluating and comparing the performance of our network with prior work, we use two metrics defined in [2] namely, the *identification rates* (id. rate, in %) and *localisation distances* (d_{mean} & d_{std}, in mm). We report the measures in Table 1. It lists the performance of three variants of our network and compares them with several recent approaches. We address three main questions through our experiments: *(1) Why the butterfly shape?* Compared to Cor.+Sag. nets, performance improves with the Btrfly net. This is because the predictions across views in the Btrfly net are spatially consistent with each other. In case of separate nets, the lack of this consistency weakens the Gaussians in the final stage of heat map fusion. We observe a 6% improvement in the id.rate over a naive 3D FCN (DI2IN) with a significantly better d_{mean}. *(2) Why the adversarial prior-encoding?* In addition to the advantages of the Btrfly net, the Btrfly$_{pe}$ net possesses adversarially encoded spatial distribution of the vertebrae. This results in about a 4% increase in the id. rate. Compared to the prior work, Btrfly$_{pe}$ net achieves state-of-art measures in both the metrics, and it does so by being a single network trained end-to-end. (cf. Fig. 4) *(3)*

Relation to latent-space learning? Learning latent representations and projecting the generated annotations onto this latent space has been addressed in [5,4]. Conversely, EB-D is more flexible as it learns from scratch and converges to a latent manifold best representing the true as well as generated data. The reconstruction capability of the AE for a generated sample is of interest. Using the output of the AE instead of $\text{Btrfly}_{\text{pe}}$, we achieve an id.rate of 75% with a d_{mean} of 19 mm. This is in spite of the AE not receiving any explicit supervision on the 'true-ness' of the annotation it is reconstructing, indicating the AEs' capability of transferring the learning to contrastive samples.

Precision & Recall. The localisation distance and the id.rate capture the ability of the network in accurately labelling a vertebra. However, both the measures are agnostic to false positive predictions. Accounting for spurious predictions becomes important especially when dealing with FCNs, as the predictions depend on a locally constrained receptive field. In our case, the false positives are controlled by the threshold T as described in Section 2.3. Accounting for these, we define two measures, *precision* (P) and *recall* (R) as: $P = \frac{\#\text{hits}}{\#\text{predicted}}$ and $R = \frac{\#\text{hits}}{\#\text{actual}}$, where #hits is the number of vertebrae satisfying the condition of identification as defined for id.rate, #predicted is the vertebrae in the prediction, and #actual is the vertebrae actually present in the image. Observe that R is synonymous to id.rate. We calculate these metrics for each scan and report the average over the entire test set. Fig. 5 shows a precision-recall curve generated by varying T between 0 to 0.8 in steps of 0.05, while Table 2 shows the performance at the F1-optimal threshold. In spite of not choosing an recall-optimistic threshold, our networks perform comparably well. Notice the over-arcing nature of Btrfly over Cor.+Sag. nets and that of $\text{Btrfly}_{\text{pe}}$ over others.

4 Conclusions

We validate the sufficiency of 2D orthogonal projections of the spine for localising and identifying the vertebrae by combining information across the projections using a butterfly-like architecture. In addition to looking at a local receptive field like any FCN, our approach considers the local structure of the spine thanks to an adversarial energy-based prior encoding, thereby outperforming the state-of-art approaches as a stand-alone network without any post-processing stages.

References

1. Chen, H., et al.: Automatic localization and identification of vertebrae in spine ct via a joint learning model with deep neural networks. In: MICCAI. pp. 515–522. Springer (2015)
2. Glocker, B., et al.: Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. In: MICCAI. pp. 590–598. Springer (2012)
3. Glocker, B., et al.: Vertebrae localization in pathological spine ct via dense classification from sparse annotations. In: MICCAI. pp. 262–270. Springer (2013)

4. Oktay, O., et al.: Anatomically constrained neural networks (ACNN): application to cardiac image enhancement and segmentation. CoRR abs/1705.08302 (2017)
5. Ravishankar, H., et al.: Learning and incorporating shape models for semantic segmentation. In: MICCAI. pp. 203–211. Springer (2017)
6. Roy, A.G., et al.: Error corrective boosting for learning fully convolutional networks with limited data. In: MICCAI. pp. 231–239. Springer (2017)
7. Suzani, A., et al.: Fast automatic vertebrae detection and localization in pathological ct scans - a deep learning approach. In: MICCAI. pp. 678–686 (2015)
8. Yang, D., et al.: Automatic vertebra labeling in large-scale 3d ct using deep image-to-image network with message passing and sparsity regularization. In: IPMI. pp. 633–644. Springer (2017)
9. Yang, D., et al.: Deep image-to-image recurrent network with shape basis learning for automatic vertebra labeling in large-scale 3d ct volumes. In: MICCAI. pp. 498–506. Springer (2017)
10. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2016)
11. Zhao, J.J., et al.: Energy-based generative adversarial network. CoRR abs/1609.03126 (2016)

5 Supplementary Material

5.1 Case study on a non-spine-centred scan

The benchmark dataset used in Section 3 of our work is mostly spine-centred, and the naive maximum intensity projections contain no occlusions. However, in certain full-body scans, the spine is obstructed by the ribcage in a MIP of the entire scan, or the spine is not spatially centred in both the views, thus not taking full advantage of Btrfly net's view fusion (cf. Fig. 6a). Such cases can be handled by a introducing a pre-processing step before the $\text{Btrfly}_{\text{pe}}$ net in the form of an 'object-detection' network.

For such scenario, we construct the MIPs in two stages. The first MIP is constructed on the entire scan. On this, we use a *single-shot object detection* (SSD) inspired architecture [1] trained to identify occluded spines (cf. Fig. 6a). Once the spine is located, we construct the second pair of MIPs based on the *spine*-slices, which are then used as inputs to the $\text{Btrfly}_{\text{pe}}$ net (cf. Fig. 6b,c). The ground truth for the SSD net can be constructed from the ground truth annotation of the vertebral centroids. We use a generic 16-layer residual CNN with an SSD extension. This use-case is illustrated on a scan from the training set of the xVertSeg [2] dataset. Note that we used the xVertSeg data only for inference and not for re-training the network. The centroids of the vertebrae are obtained from the maximum point of the distance transform of the segmentation map (xVertSeg has voxel-level annotations from L1 to L5).

References

1. Liu W. et al.: SSD: Single Shot MultiBox Detector. CoRR abs/1512.02325 (2015), http://arxiv.org/abs/1512.02325.
2. The xVertSeg Challenge, http://lit.fe.uni-lj.si/xVertSeg/

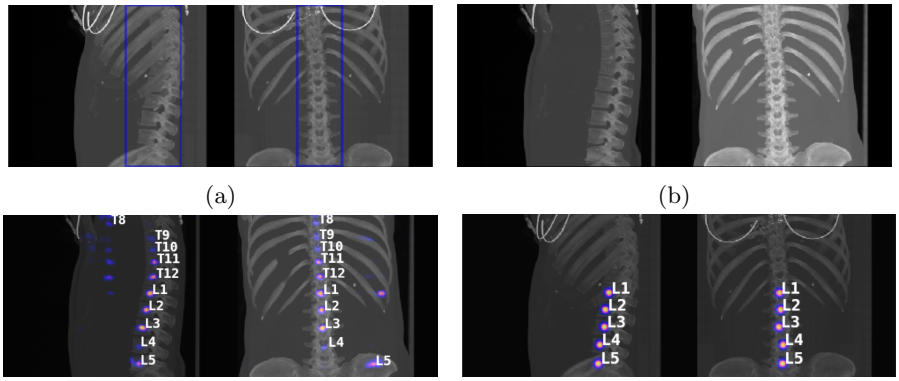


Fig. 6: An illustration of the extension to $\text{Brtfly}_{\text{pe}}$ net. **(a)** Naive sagittal and coronal MIPs on the entire scan with the bounding box predictions (in blue) of our SSD net. Observe the ribcage obstructing the spine. **(b)** Improved MIPs constructed from the slices containing the spine based on the localisation in (a). **(c)** Output of the $\text{Btrfly}_{\text{pe}}$ net, resulting in an 80 % id.rate. Also observe the incorrect localisation of T8 and L5, along with prediction noise in sagittal view owing to the non-aligned spine in both views. We believe that aligning the spine using its detection could further improve the prediction. **(d)** The ground truth centroids constructed from the voxel-level annotation map of scan. Since xVertSeg data only has lumbar annotations, we visualise the lumbar centroids.

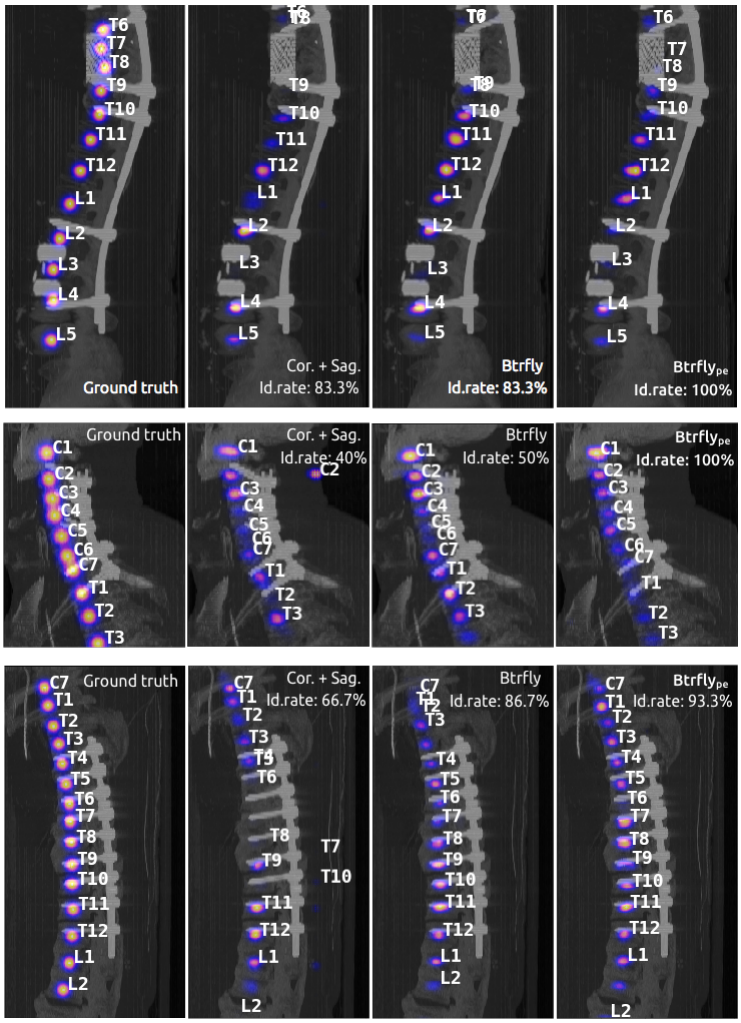


Fig. 7: **Additional quantitative results.** MIP images with predictions of the three variants of our approach at T=0 for all cases. The spine's local structure is conserved in the predictions of Btrfly$_{pe}$. Also observe that, as a consequence of prior encoding, in some cases labels are predicted in spite of no useful spatial information, albeit the strength of these predictions is less.