Removing Segmentation Inconsistencies with Semi-Supervised Non-Adjacency Constraint

Pierre-Antoine Ganaye, Michaël Sdika, Bill Triggs, Hugues Benoit-Cattin

Removing Segmentation Inconsistencies with Semi-Supervised Non-Adjacency Constraint

Pierre-Antoine Ganaye[a], Michaël Sdika[a,], Bill Triggs[b], Hugues Benoit-Cattin[a]

[a] *Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69100, Lyon, France*
[b] *Laboratoire Jean Kuntzmann, Université Grenoble-Alpes, CNRS, Grenoble INP, INRIA, B.P. 53, 38041 Grenoble Cedex 9, France*

ARTICLE INFO

Article history:

Keywords:
Brain-region segmentation
Anatomical adjacency constraints
Semi-supervised training
Magnetic resonance imaging

ABSTRACT

The advent of deep learning has pushed medical image analysis to new levels, rapidly replacing more traditional machine learning and computer vision pipelines. However segmenting and labelling anatomical regions remains challenging owing to appearance variations, imaging artifacts, the paucity and variability of annotated data, and the difficulty of fully exploiting domain constraints such as anatomical knowledge about inter-region relationships. We address the last point, improving the network's region-labeling consistency by introducing NonAdjLoss, an adjacency-graph based auxiliary training loss that penalizes outputs containing regions with anatomically-incorrect adjacency relationships. NonAdjLoss supports both fully-supervised training and a semi-supervised extension in which it is applied to unlabeled supplementary training data. The approach substantially reduces segmentation anomalies on the MICCAI-2012, IBSRv2 brain MRI datasets and the Anatomy3 whole body CT dataset, especially when semi-supervised training is included.

1. Introduction

Medical image segmentation is a critical technology for localizing, classifying and quantifying anatomical structures. Clinical examination of segmented cerebral regions is one of the key applications of neuroimaging, for example to investigate structural brain connectivity (Frau-Pascual et al., 2019). Diffeomorphic atlas-based methods have long been a robust choice (Vercauteren et al., 2009; Ashburner, 2007; Sdika, 2008, 2013), providing theoretical guarantees and yielding consistent segmentation maps that preserve both the topologies and interrelationships of structures. Machine learning approaches have been used to complement the traditional multi-atlas fusion step (Wang and Yushkevich, 2013; Coupé et al., 2011; Sdika, 2010), correcting for potential voting errors. However multi-atlas methods are computationally burdensome owing to need to register the input to each atlas.

In comparison, Convolutional Neural Network (CNN) approaches are proving to be both efficient and accurate. (Moeskops et al., 2016) introduced a CNN-based multi-scale patch-level classifier for segmentation. Although patch-based methods can exploit contextual information around each pixel and side information such as the image position of the patch (Ganaye et al., 2018; de Brebisson and Montana, 2015; Ghafoorian et al., 2017), they are unable to exploit global constraints such as volumetric and anatomical consistency and this limits their overall performance. CNN's based on multiscale encoder-decoder architectures are able to take account of the entire input image, for example using "fully convolutional" approaches (Long et al., 2015; Badrinarayanan et al., 2017; Ronneberger et al., 2015). This has proved to be more effective than the patch based approach and it allows the inclusion of loss terms that exploit richer forms of 2D and 3D domain information. For example for brain MRI, (Roy et al., 2017) proposed an encoder-decoder model pre-trained on a dataset annotated automatically with Freesurfer then fine-tuned with a specific loss focused on mining hard negatives. For 2D/3D segmentation (Sudre et al., 2017) formulated a generalized Dice loss that is more robust to highly imbalanced problems.

In part the success of CNNs has been driven by the advent of larger annotated training sets, formerly a rare commodity for various reasons (the complexity of manual segmentation, storage costs, ethical requirements, ...). CNNs are typically trained to minimize low-level image-based differences between the inferred segmentation and the annotated ground truth, as measured by cost functions such as cross-entropy and Dice. So their precision is limited by both their myopic view of correctness and the quantity and quality of the available training data (ideally this should capture the full range of inter-subject variability

e-mail: michael.sdika@creatis.insa-lyon.fr (Michaël Sdika)

and the annotations should embody a broad consensus among experts). For these reasons early CNN methods were not always a clear improvement on traditional segmentation pipelines and there have been various attempts to harness properties such as anatomical invariance (Oktay et al., 2018; Kervadec et al., 2018; Ravishankar et al., 2017) and semantic knowledge (Xu et al., 2018) within the CNN framework.

Dense (pixel level) conditional random fields (CRF) (Krähenbühl and Koltun, 2011) have been used in many computer vision problems as a post-processing method to correct labelling inconsistencies, despite their computational cost. In the broader deep learning community such priors (contextual information, spatial position) can be used either as additional inputs or soft constraints. For brain image segmentation, (de Brebisson and Montana, 2015; Ganaye et al., 2018) integrated a spatial localization feature as a representation prior, requiring the segment positions to be correlated with their anatomical structures. In constrained image segmentation, (Oktay et al., 2018) used an auto-encoder to learn a prior on the label space, extracting features directly on the label maps and using them to penalize the output segmentations during training. (Ravishankar et al., 2017) formulate a method that learns and integrates a shape prior, while (Kervadec et al., 2018) imposes restrictions on the volumes of the segmented structures during training. To include containment and exclusion priors between the structures in multi-label segmentation, (BenTaieb and Hamarneh, 2016) defined a loss inspired from the conditional random field (CRF) approach: the last layer of the network is a sigmoid function which forces the model to segment correctly while penalizing impossible configurations between labels. Compared to the latter, our loss directly penalizes impossible adjacencies on the joint probability space.

The contributions of this paper can be summarized as follows:

- We propose a new methodology that reduces the number of segmentation abnormalities by penalizing violations of the known adjacency relationships between anatomical regions. We use a 2D encoder-decoder model inspired by (Roy et al., 2017), but during fine-tuning we augment its label-based segmentation loss (Dice or cross-entropy) with an original, fully differentiable, structure adjacency loss named NonAdjLoss.

- We show that the non-adjacency penalty can also be used in a semi-supervised fashion, supplementing the annotated training data with additional unlabeled images to improve generalization without compromising accuracy.

- We explore a change of architecture that expands the scope of the original 2D segmentation method to 3D, at the same time reformulating the NonAdjLoss to consider spatial arrangements between regions.

- We show that our methods provide a remarkable reduction in segmentation outliers on two neuroimaging datasets, MICCAI 2012 (Landman) and IBSR V2 (Worth), and a multi-organ dataset, Anatomy3 (Jimenez-del-Toro et al., 2016). We attribute this improvement to the NonAdjLoss and semi-supervised training.

Our implementation of NonAdjLoss training is available at https://github.com/trypag/NonAdjLoss.

2. Methods

Sect. 2.1 introduces our 2D segmentation architecture and Sect. 2.2 shows how we extract adjacency rules from ground truth label maps. Sect. 2.3 formulates a differentiable non-adjacency loss that operates directly on output segmentation maps, and shows how it can be enforced on unannotated images. In Sect. 2.5 the 2D architecture is extended to 2.5D and the NonAdjLoss is reformulated to account for the relative spatial displacements between structures. Finally, Sect. 2.4 details our practical algorithm for optimizing this loss.

2.1. Encoder-Decoder architecture

Our first CNN architecture is an encoder-decoder inspired by (Roy et al., 2017). This network (Fig. 1) takes 7 consecutive 2D slices as input and uses these to segment the middle one. The additional slices bring contextual information about the central one, improving the overall robustness. The network is a U-net composed of four 2× downsampling layers (the encoding path), followed by four upsampling steps based on max-unpooling (the decoding path). Each decoding layer also has direct connections from the corresponding encoding one.

2.2. Anatomical adjacency matrix

We make the assumption that all subjects will have the same anatomical adjacencies and thus inter-region connectivities, even though their region geometries may vary. In the image, the adjacency relationships between each pair of regions i and j can be represented by an adjacency matrix $\mathbf{A}$, where $\mathbf{A}_{ij}$ is the total number of voxels on the boundary between the annotated segments labelled i and j. Formally,

$$\mathbf{A}_{ij} = \sum_{x} \sum_{v \in V} \delta_{i,s(x)} \delta_{j,s(x-v)}, \tag{1}$$

where x are the voxels, $s(x)$ is the label at x, δ is the Kronecker delta function and V defines a local neighborhood. $\mathbf{A}$ encodes the surface area of the contours shared between pairs of structures in the 3D volume. The volumes of anatomical structures may vary considerably between subjects owing to inter-person variability and neuropathologies. For this reason we choose to binarize $\mathbf{A}$ to $\tilde{\mathbf{A}} = (\mathbf{A} > 0)$ as this is invariant to homeomorphic image deformations (see Fig. 2 as an example). We define the set of impossible transitions between structures to be $F = \{(i, j) \mid \tilde{\mathbf{A}}_{ij} = 0\}$, for $\tilde{\mathbf{A}}$ in the training set. This defines the set of anatomical adjacencies that we want to forbid during the training of the model.

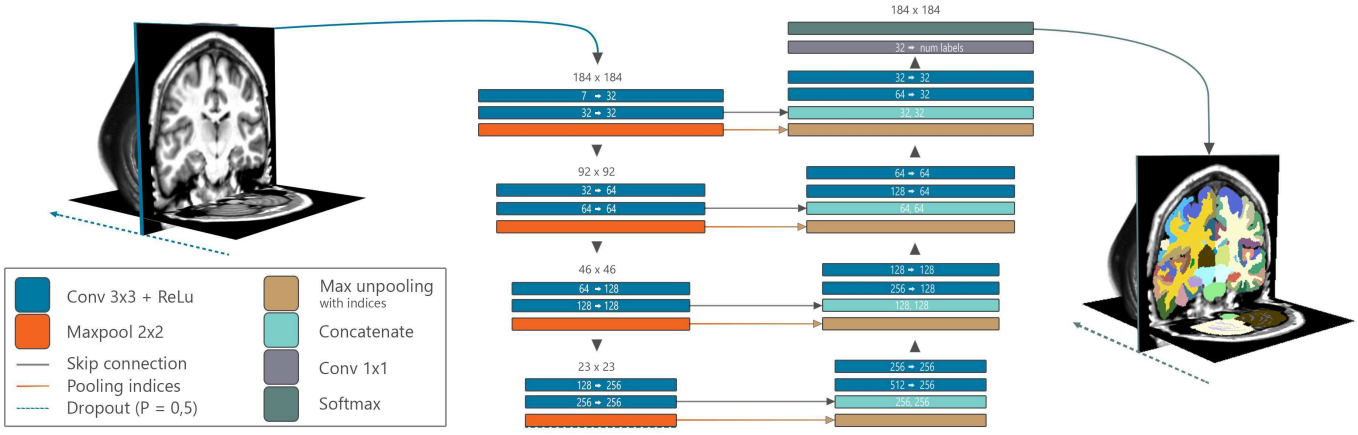

Fig. 1: Our pipeline for 2D image segmentation. Seven adjacent slices are given as input to the neural network, which outputs the label map of the central slice only. A fully convolutional U-net encoder-decoder architecture is used to obtain a fast slice-by-slice volume segmentations. The network has about 3 million parameters.

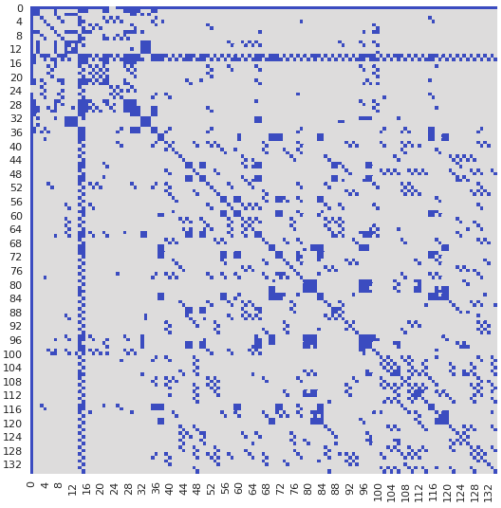

Fig. 2: Binary adjacency matrix $\tilde{\mathbf{A}}_{ij}$ extracted on the training dataset of the MICCAI 2012 multi-atlas segmentation challenge. Blue denote adjacencies between structures in a 3×3×3 neighborhood.

2.3. Training a segmentation network with adjacency constraints

Constrained training. The main objective of this work is to train segmentation networks to produce outputs that satisfy the anatomical constraints encoded within F. To this end we define a constraint function $G(\mathbf{w})$ of the network parameters $\mathbf{w}$, which is zero when all constraints are satisfied for all images of the dataset and increases with the number of inconsistencies. The network is trained by solving the following optimisation problem:

$$\min_{\mathbf{w} \mid G(\mathbf{w})=0} \frac{1}{|D_S|} \sum_{(\mathbf{I},\mathbf{S})\in D_S} L(\phi(\mathbf{I},\mathbf{w}),\mathbf{S})), \tag{2}$$

where

$$G(\mathbf{w}) = \sum_{\mathbf{I}\in D_G} \sum_{(i,j)\in F} a_{ij}(\phi(\mathbf{I},\mathbf{w})). \tag{3}$$

Here, $\mathbf{I}$ is a greylevel image and $\mathbf{S}$ is its annotated label map. D_S and D_G are the training datasets used respectively for the segmentation loss and the NonAdjLoss (D_G will typically include D_S plus some unannotated supplementary images). ϕ is the function defined by the neural network. For an image $\mathbf{I}$ and given the network's weights $\mathbf{w}$, $\phi(\mathbf{I},\mathbf{w})$ is the network's output : a multi-channel image providing for every pixel the probability of belonging to each class. The a_{ij} function is defined below.

Adjacency functions. The function a_{ij} measures the soft adjacency between output labels i and j based on the network's output probability maps. Its form is inspired by Eq. 1 but the function $\delta_{\cdot,s(x)}$ needs to be softened for use in the context of soft output labelings and gradient descent training. Let $\phi_i(x)$ be the probability map for label i in image $\mathbf{I}$, as given by the neural network output. When two regions i and j should not be adjacent, $(i,j)\in F$, the probability of respectively belonging to i and j should be simultaneously null for a pixel and all of its neighbors. A simple means of enforcing this is to require $\phi_i(x)\phi_j(x-v)$ to be low for x and its neighbors $x-v$. To apply this penalty over the image we define a_{ij} as :

$$a_{ij}(\phi) = \sum_x \sum_{v\in V} \phi_i(x)\phi_j(x-v), \tag{4}$$

where ϕ is the label probability vector map. If we define $\tilde{\phi} = \phi * \mathbb{1}_V$ as the convolution of ϕ with the indicatrix of the neighborhood element V, this expression can be simplified for faster computation:

$$a_{ij}(\phi) = \sum_x \phi_i(x) \sum_{v\in V} \phi_j(x-v) \tag{5}$$

$$= \sum_x \phi_i(x)\tilde{\phi}_j(x) \tag{6}$$

If the network gives crisp outputs, $\phi_i(x)$ becomes $\delta_{i,\mathrm{argmax}_k\, p_k(x)}$ and $a_{ij}(\phi)$ reduces to $\mathbf{A}_{ij}(\phi)$.

The constraint $G(w)$ is the sum of these functions for all forbidden adjacencies and all images. As with most objective functions used for deep neural network training, $G(w)$ is not convex with respect to the network weights and care is needed with the numerical resolution of 2 (see section 2.4). However G is at least positive and quadratic (but non-convex) with respect to the network's output probabilities $\phi_i(x)$, which ensures that the objective in is bounded below.

Extension to semi-supervised learning. Once the adjacency matrix **A** has been extracted from the annotated label maps of D_S, it is considered to be the ground truth forbidden adjacency rule for every input image. Using the network with a_{ij} as a differentiable adjacency metric, we can then estimate the forbidden connectivity of any image, whether or not it is in the original annotated dataset D_S. This allows us to include unannotated images in the adjacency-constraint training dataset D_G, giving a form of semi-supervised training in which the network is simultaneously optimized to segment structures based on full annotations when available, and to enforce the NonAdjLoss on all images whether or not they are annotated. As we will show in the experiments, this gives us a great deal of scope to improve the anatomical reliability of the output labelings by including multi-centric datasets during training. (see fig. 3).

2.4. Constrained optimization algorithm

Optimization. In practice, we solve the constrained optimization problem using a penalty method similar to (Nocedal and Wright, 2006): the network is trained by continuation using the constraint G as a penalty term with gradually increasing weight λ:

$$\min_{\mathbf{w}} \frac{1}{|D_S|} \sum_{(\mathbf{I},y)\in D_S} L(\phi(\mathbf{I},\mathbf{w}),y)) + \lambda\, G(\mathbf{w}). \tag{7}$$

It turns out to be important to pre-train the network using the standard segmentation loss before activating the NonAdjLoss constraints. The overall procedure is detailed in Algo. 1, where train(λ) denotes the result of the optimization problem Eq. 7 under the given soft constraint.

Algorithm 1 Constrained learning algorithm

Initialization
$L_0, G_0 = train(0)$
$\lambda = \lambda_{ratio} \times \frac{L_0}{G_0}$
for $i = 0$ to $i = n_{epochs}$ **do**
  $L_i = train(\lambda)$
  if $i \mod n_{update}$ **then**
    if $L_0 - L_i < \epsilon$ **then**
      $\lambda = \lambda * \lambda_{increase}$
    else
      $\lambda_{increase} = \lambda_{increase} * \lambda_{reduction\_factor}$
      $\lambda = \lambda * \lambda_{reduction}$
    end if
  end if
end for

Here and below, L_i and G_i denote respectively the average Dice or cross-entropy and the average NonAdjLoss on the training set at the end of the i-th epoch. Initially λ is set to make the non-adjacency loss contribution a fraction λ_{ratio} of the segmentation loss – in practice $\lambda_{ratio} = 0.3$. High λ_{ratio} settings (0.8 for example) tend to lead to overly-local flipping of pixel labels, nominally removing adjacency errors but adversely affecting the Dice metric and the segmentation topology. Low λ_{ratio} settings slow the method's convergence to optimality. While training, if the validation-set Dice is steady or improving, λ is increased by $\lambda_{increase}$ every n_{update} epochs. Conversely, if the Dice falls more than ϵ below that of the initial unconstrained iteration, λ is rolled back to a lower value and the step size $\lambda_{increase}$ is also reduced. $\lambda_{reduction\_factor}$ is the constant reduction factor applied to $\lambda_{increase}$ when the Dice drops: low $\lambda_{increase}$ values slow the convergence while high ones create training instabilities.

Multi-objective model selection. To choose the best model, one should usually look for the epoch at which the validation metric reaches its best level. However in this paper we are interested in both segmentation and adjacency metrics so we propose a simple multi-objective selection rule. To select the final set of model parameters, we take the epochs with the five best validation-set Dice scores and choose the model with the lowest validation-set non-adjacency loss among these. This strategy plays an important role in finding optimal models with regard to the validation set and it helps to reduce overfitting on the training set.

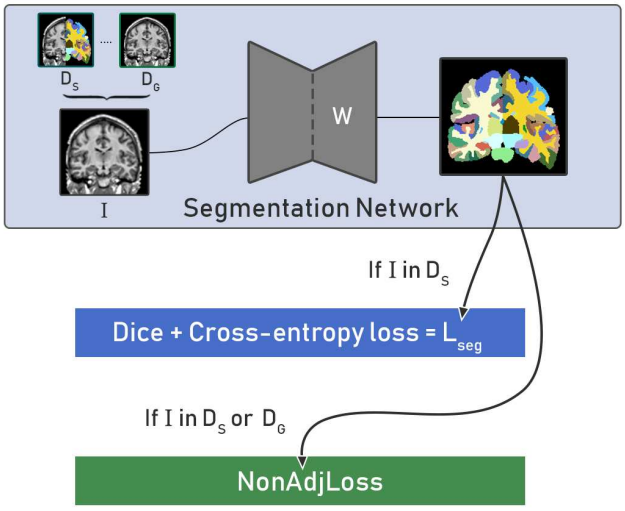

Fig. 3: Global overview of the semi-supervised scheme, where the network parameters **w** are optimized using L_{seg} and NonAdjLoss on annotated images and L_{graph} alone on unannotated ones.

2.5. Extension to 3D

2.5D architecture. To benefit from the 3D nature of brain images we would like to extend the adjacency constraints in the depth dimension. This is problematic within a lightweight 2D architecture. Using full 3D convolutions would solve the problem, but at the expense of a larger network requiring significantly more memory than our GPU's had available. Instead we altered the architecture of Fig. 1 to segment the three central slices out of the seven input, instead of just the central one. Specifically, we replaced the final 1×1 convolution with three parallel 1×1 convolutions, estimating segmentation probability maps for both the central slice and the ones immediately above and below it. Each branch is optimized based on its own ground truth so the segmentation loss becomes a sum ofthese 3 terms. We will refer to this 3-slice architecture as the 2.5D one. It allows the use of a 3D neighborhood to compute the adjacency

constraints, but it requires a simple map-fusion strategy at inference: the final probability map of a given slice is the average of the three corresponding maps generated from the current block and the ones immediately above and below it. (see fig 4).

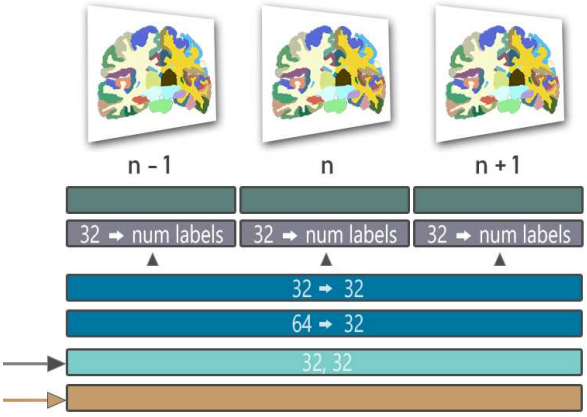

Fig. 4: The configuration of the last convolution block for the 2.5D architecture. The final convolution is converted into 3 parallel convolutions, generating 3 distinct maps.

Oriented adjacencies. The adjacency matrix $\mathbf{A}$ was found by evaluating, for each pixel, the labels of all of its neighbours in a symmetric 3D neighbourhood. This overlooks anatomical sidedness constraints. For instance in neuroanatomy one knows that the right putamen is not merely adjacent to the right pallidum but also to the right of it. We can strengthen the anatomical constraint by replacing $\mathbf{A}$ with six separate matrices, one for each of the six available orientations $o \in O = \{\text{front, back, top, bottom, left, right}\}$. These are built in the same way as $\mathbf{A}$ but using oriented neighbourhoods that encode adjacency in each direction separately. The network optimizer needs to enforce all six constraints $G_o(\mathbf{w})$:

$$\min_{\mathbf{w} \mid G_o(\mathbf{w})=0, \forall o \in O} \frac{1}{|D_S|} \sum_{(\mathbf{I},y)\in D_S} L(\phi(\mathbf{I},\mathbf{w}),y)) \tag{8}$$

The numerical procedure used to solve this constraint problem is the same as in section 2.4. The only difference is that the penalty function is the sum of all six G_o functions.

3. Experiments

We tested the NonAdjLoss and our semi-supervised training method on two neuroimaging datasets and a whole-body dataset (Section 3.1). Post-processing with a dense conditional random field was evaluated for comparison. The hyperparameters used for training are detailed in Section 3.2. To quantify our methods' ability to reduce the incidence of adjacency errors in segmented images, we propose new quality metrics that count unique connections and volumes (Section 3.3).

3.1. Data

Neurological Imaging. The method was evaluated on brain-region segmentations from T1-weighted MR images using the MICCAI 2012 multi-atlas challenge (Landman) and IBSRv2 datasets (Worth). Each brain imaging dataset was split into training/validation/test subsets as presented in table 1. We followed the official train/test experimental protocol for the MICCAI challenge, however no official data split is provided for the IBSRv2 dataset. The OASIS dataset (Marcus et al., 2010) was used as the source of unlabelled training data for the semi-supervised experiments, excluding the subjects who also appear in MICCAI 2012. In IBSRv2, 6 of the 39 labels were removed from the segmentation problem (ones such as Lesion, Blood vessel, and Unknown).

	subjects	labels	train	validation	test
MICCAI12	35	135	10	5	20
IBSRv2	18	33	10	3	5
OASIS	406	0	284	122	0

Table 1: Characteristics of the three brain MRI datasets used in the experiments: numbers of patients, labels, training images, validation images, and test images. The OASIS dataset is entirely composed of unannotated images.

All of the images were affine-registered to a reference atlas in the MNI space with FSL FLIRT, then resampled to 1mm voxel spacing. Bias field correction was applied with N4ITK (Tustison et al., 2010). The mean and standard deviation were estimated for each dataset and the corresponding images were centered and reduced. Skull stripping was not used as a preprocessing step because we found that our CNN's were able to label the skull as background with high precision. During training, the images from the annotated (D_S) dataset were artificially augmented with elastic deformations (Simard et al., 2003) in order to simulate the inherent natural variability of anatomy.

Whole Body Imaging. Anatomy3 is a multi-organ dataset composed of CT scans and MRIs with and without contrast agent, where 20 anatomical regions were annotated by trained experts. It was created for the purpose of the Visceral (Jimenez-del-Toro et al., 2016) segmentation challenge, which is not running anymore. We did not have access the test set used for the challenge, however a "Silver Corpus" set was released publicly, with annotations crowd-sourced by merging the segmentations obtained from the participants' models. Our train/validation/test split was as follows :

- 10 images for training, 10 for validation, with all data from the official training set.
- 25 images for testing, 30 images for semi-supervision (excluding annotations), with all data from the "Silver Corpus".

During pre-processing, all images were clamped between $[-1000; 2000]$ Hounsfield units and intensity-normalized to set their mean to 0 and standard deviation to 1. For computational speed and to conserve GPU memory, we sub-sampled the xy axis (acquisition plan) from resolution 512×512 to 256×256, while the z axis was preserved at its original resolution. Cubic interpolation was used for image re-sampling, and nearest neighbor interpolation for the label maps.

3.2. Implementation Details

Network optimization was performed using stochastic gradient descent with momentum of 0.9. The batch size was set to 8. The learning rate was initialized to 0.01 and updated using the polynomial rate policy (Chen et al., 2016). During training only the cross-entropy or Dice loss was used for the first 300 epochs, then the NonAdjLoss was progressively enforced, adjusting its weight λ using Algorithm 1. The parameters of Algorithm 1 were: $\lambda_{ratio} = 0.3$, $\lambda_{increase} = 1.3$, $\lambda_{reduction\_factor} = 0.98$, $\lambda_{reduction} = 0.90$, $n_{update} = 5$ and $\epsilon = 0.02$. The non-adjacency penalization was applied for 170 epochs, using a decreased learning rate of 0.001.

When optimizing the cross-entropy losses, we found that class imbalance caused issues owing to the large number of label classes and the considerable volume variations between structures. Following (Roy et al., 2017), median frequency weighting was applied with success. The Dice loss did not need such an imbalance correction.

When used, the dense CRF inference (Krähenbühl and Koltun, 2011) was run for 15 steps, with a unary term based on the network's probability maps and two pairwise terms: position dependant and image dependant. The pairwise terms take advantage of $\tilde{A}$, the binarized adjacency matrix, as the label compatibility term. Increasing the number of steps to 50 decreased the performance and led to a processing time of 1 hour.

3.3. Evaluation

For all of the experiments we report Dice, Hausdorff Distance and Mean Surface Distance. These metrics do not directly measure topological defects such as adjacency errors, so to quantify such anatomical inconsistencies we introduce two new metrics:

$$CA^{unique}(A^I) = \frac{|O^I \cap H|}{|H|} \quad (9)$$

$$CA^{volume}(A^I) = \frac{\sum_{(i,j)\in(O^I\cap H)} A^I_{ij}}{vol_{contour}}, \quad (10)$$

where A^I is an output segmentation map's adjacency graph for a given image I, $\tilde{A}$ the binarized ground truth adjacency matrix, $O^I = \{(i,j) \mid A^I_{ij} > 0\}$, $H = \{(i,j) \mid \tilde{A}_{ij} = 0\}$ and $vol_{contour}$ is total number of contour voxels in the inferred segmentation. CA^{unique} is the percentage of all the forbidden inter-class adjacencies that appear somewhere in the image, while CA^{volume} is the percentage of the region-boundary voxels in the image that have forbidden adjacencies. The former measures the fraction of all the incorrect region adjacencies that appear, while the latter gives the volumetric ratio of pixels with adjacency errors.

4. Results

Section 4.1 reports results showing the impact of including the NonAdjLoss and semi-supervised training. To account for adjacencies at different scales, we test a sum of NonAdjLosses of varying neighborhood sizes in Section 4.2. Finally, the original 2D architecture is extended to 2.5D in Section 2.5, and combined with the oriented non-adjacency penalization in Section 4.4.

4.1. NonAdjLoss and semi-supervision

The results of our method are presented in tables 2 and 3, where *Baseline* is for the model with unconstrained training (Dice and cross-entropy losses) and $NonAdjLoss(n)$ for the model with constrained training and n unannotated images from the OASIS dataset. Note that all three datasets (MICCAI12, IBSRv2, Anatomy3) see significant improvements in average Hausdorff Distance across models (with a confidence level of 95%). Likewise, the non-adjacency metrics CA^{unique} and CA^{volume} evaluated on 30 images from OASIS indicate that forbidden connectivities are being reduced quite effectively. As a comparison with methods using spatial pairwise priors, the *Baseline* was post-processed using the dense CRF approach proposed in (Krähenbühl and Koltun, 2011) with a unary potential based on the negative log likelihoods of the network output maps. Table 2 shows that including the CRF inference does lead to a slight improvement in distance and connectivity metrics, but at the cost of 13 minutes of post-processing time, compared to less than a second for the underlying *Baseline* and $NonAdjLoss$ models.

The results for the Visceral Anatomy3 organ segmentation tests (table 3) show the same tendencies as for MICCAI12 and IBSRv2, with a sharp decrease of 30mm in the mean Hausdorff Distance as well as a constant reduction of abnormal connectivity. However the Dice scores are slightly decreased (by 0.01): this can be attributed the hyperparameter settings and our multi-objective selection criteria.

Adjacency graphs obtained by merging all output-label transitions seen anywhere in the MICCAI12 and IBSRv2 test sets are presented in Fig. 6, for which we set the neighborhood size to $3 \times 3 \times 3$. No pre-processing was required on the adjacency matrix. We did try varying the thresholding levels in an attempt to eliminate erroneous transitions introduced by annotation errors, but no improvement was observed. In both datasets the baseline models produce a large number of forbidden inter-class transitions (red dots), while the same CNNs trained with the NonAdjLoss make significantly fewer errors, as confirmed at 95% confidence level by a paired t-testin which each models' mean is compared to the *Baseline* mean. Labeling errors that are spatially distant from their ground truth regions are corrected as expected, while the allowed transitions (blue dots) are preserved. For both the 2D and 2.5D architectures, the best results for distance and connectivity metrics are obtained when semi-supervised training is used to further reinforce the NonAdjLoss.

Semi-supervision proves to be a powerful addition to the framework if one needs to enforce stronger structural adjacencies. This is especially true in cases where little annotated imagery is supplied for training, but a great deal of raw imagery is available. For example Fig. 7 (left) shows (on a log scale) the total number of incorrect adjacencies across all OASIS subjects for each anatomical structure, comparing the baseline, $NonAdjLoss(0)$ and $NonAdjLoss(50)$ methods. The semi-supervised model (blue) is meaningfully more reliable than either the baseline or the raw $NonAdjLoss$ one, providing anomaly-free segmentation (according to our non-adjacency criteria) for a large number of anatomical regions on

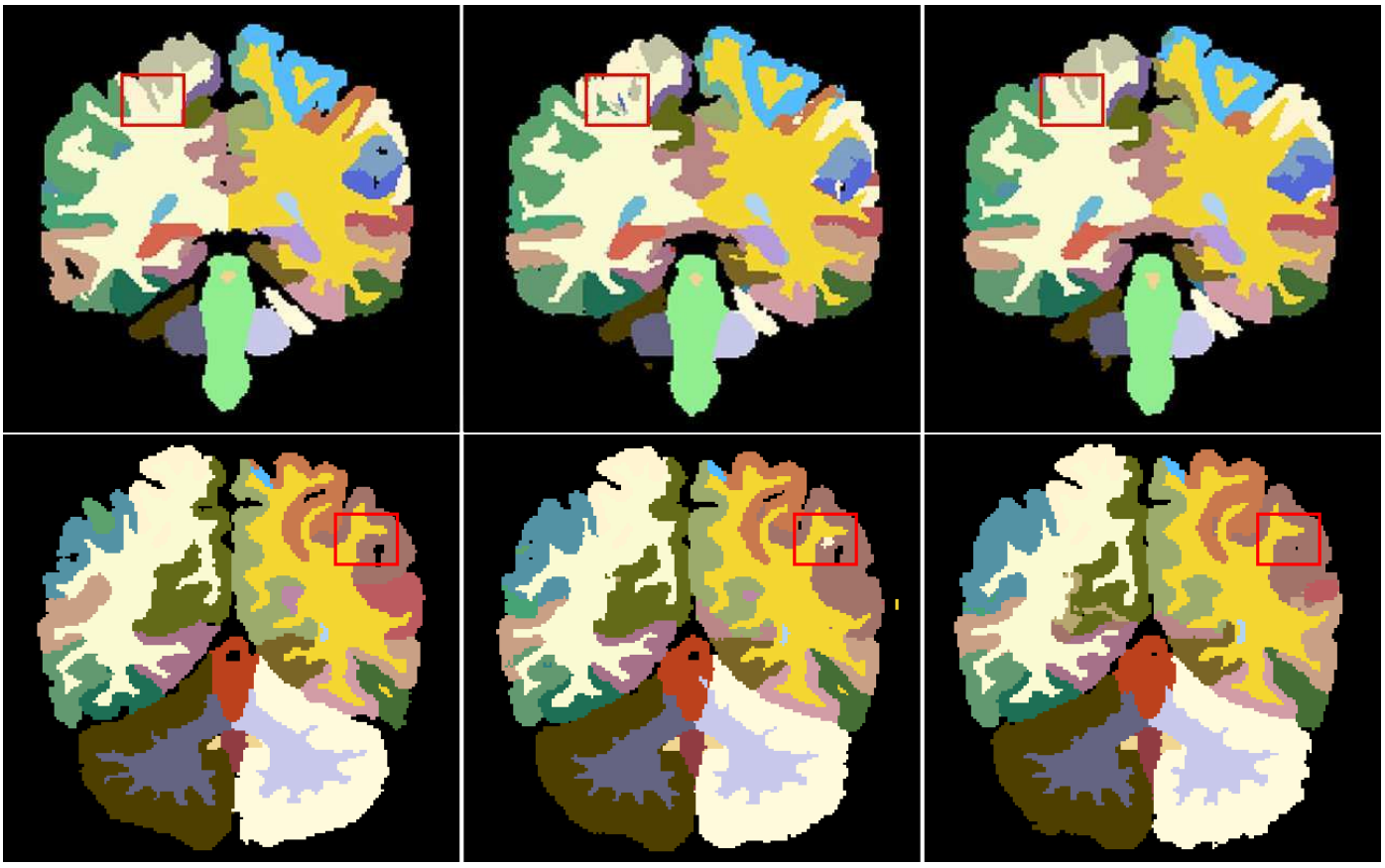

Fig. 5: Segmentation maps for two subjects from the MICCAI12 dataset, from left to right: ground truth, model based only on L_{seg} loss, model including NonAdjLoss with semi-supervision. The red boxes highlight areas where anatomical inconsistencies were corrected.

MICCAI12	Dice	HD (mm)	MSD (mm)	CA^{unique}	CA^{volume}
Baseline	0.740 ± 0.11	20.93 ± 9.50	1.18 ± 0.40	5.1e-2 ± 6.8e-2	1.8e-2 ± 4.9e-2
Baseline + CRF	0.739 ± 0.11*	18.86 ± 8.03*	1.17 ± 0.40	4.4e-2 ± 6.8e-2*	1.5e-2 ± 4.5e-2*
NonAdjLoss(0)	0.734 ± 0.10*	12.37 ± 4.62*	1.10 ± 0.34	2.7e-3 ± 6.6e-3*	2.6e-4 ± 9.4e-4*
NonAdjLoss(20)	0.739 ± 0.10*	11.19 ± 4.40*	1.06 ± 0.34	5.8e-4 ± 1.4-e3*	2.8e-5 ± 9.6e-5*
NonAdjLoss(50)	0.741 ± 0.10	**10.97** ± **4.37***	1.04 ± 0.33	**3.9e-4** ± **9.9e-4***	**1.4e-5** ± **4.8e-5***
NonAdjLoss(100)	0.743 ± 0.10	11.31 ± 4.69*	1.04 ± 0.33	4.7e-4 ± 1.5e-3*	1.9e-5 ± 6.8e-5*
Multi-scale$\{3^3, 7^3, 11^3\}$	0.737 ± 0.10*	12.46 ± 5.23*	1.09 ± 0.36	4.4e-3 ± 1.0e-2*	5.4e-4 ± 1.5e-3*
Multi-scale$\{3^3, 11^3, 15^3\}$	0.734 ± 0.10*	12.01 ± 4.69*	1.09 ± 0.34	2.6e-3 ± 5.4e-3*	3.0e-4 ± 8.4e-4*
IBSRv2	Dice	HD (mm)	MSD (mm)	CA^{unique}	CA^{volume}
Baseline	0.833 ± 0.11	15.99 ± 15.27	0.78 ± 0.37	1.0e-1 ± 8.8e-2	1.5e-3 ± 3.0e-3
NonAdjLoss(0)	0.835 ± 0.10*	14.04 ± 15.45	**0.76** ± **0.34***	7.0e-3 ± 2.1e-2*	3.1e-5 ± 1.5e-4*
NonAdjLoss(20)	0.834 ± 0.10	12.75 ± 13.26	0.77 ± 0.34*	1.2e-3 ± 2.2e-3*	**3.4e-7** ± **8.7e-7***
NonAdjLoss(50)	0.832 ± 0.10	**11.92** ± **12.65***	0.77 ± 0.37	1.6e-3 ± 4.6e-3*	1.8e-6 ± 8.1e-6*

Table 2: Distance, similarity and connectivity metrics for each model. HD denotes Hausdorff Distance and MSD denotes Mean Surface Distance, both in millimeters. Dice, HD and MSD were averaged over the test set while CA^{unique} and CA^{volume} were averaged over 30 unlabeled images from the OASIS test set. NonAdjLoss(n) denotes the same network architecture as the baseline model, but trained using NonAdjLoss penalization and n images of semi-supervised data. * indicates that the metric's mean is significantly different from the *Baseline* with a confidence level of 95%. We report average score ± standard deviation.

this dataset.

In tables 2 and 3, although clear improvements are seen for the connectivity metrics (CA^{unique}, CA^{volume}) and the surface metrics (HD, MSD), no such improvements are seen for the Dice score. The $NonAdjLoss(n)$ models have similar Dice scores to the *Baseline* ones, sometimes slightly better and sometimes slightly worse. This follows from the nature of the constraints imposed by $NonAdjLoss$: for the most part, clearing inconsistencies removes small erroneous regions that are far from their true anatomical locations. As these are small their impact on the Dice is limited (especially after averaging over the classes), whereas if their distances to their true locations are large their impact on the Hausdorff Distance is more important. During the final revision of this paper we became aware of the

Visceral Anatomy3	Dice	HD (mm)	MSD (mm)	CA^{unique}	CA^{volume}
Baseline	0.682 ± 0.26	88.76 ± 52.30	3.88 ± 2.31	9.2e-2 ± 3.9e-2	3.9e-4 ± 6.4e-4
NonAdjLoss(0)	0.679 ± 0.26	58.44 ± 39.46*	3.38 ± 2.01	1.1e-2 ± 1.3e-2*	3.5e-5 ± 8.3e-5*
NonAdjLoss(30)	0.674 ± 0.26*	57.53 ± 41.45*	3.17 ± 1.87*	5.4-e3 ± 5.9e-3*	7.8e-6 ± 2.0e-5*

Table 3: Distance, similarity and connectivity metrics for each model. HD denotes Hausdorff Distance and MSD denotes Mean Surface Distance, both in millimeters. Dice, HD and MSD were averaged over the test set. NonAdjLoss(n) denotes the same network architecture as the baseline model, but trained using NonAdjLoss penalization and n images of semi-supervised data. * indicates that the metric's mean is significantly different from the *Baseline* with a confidence level of 95%. We report average score ± standard deviation.

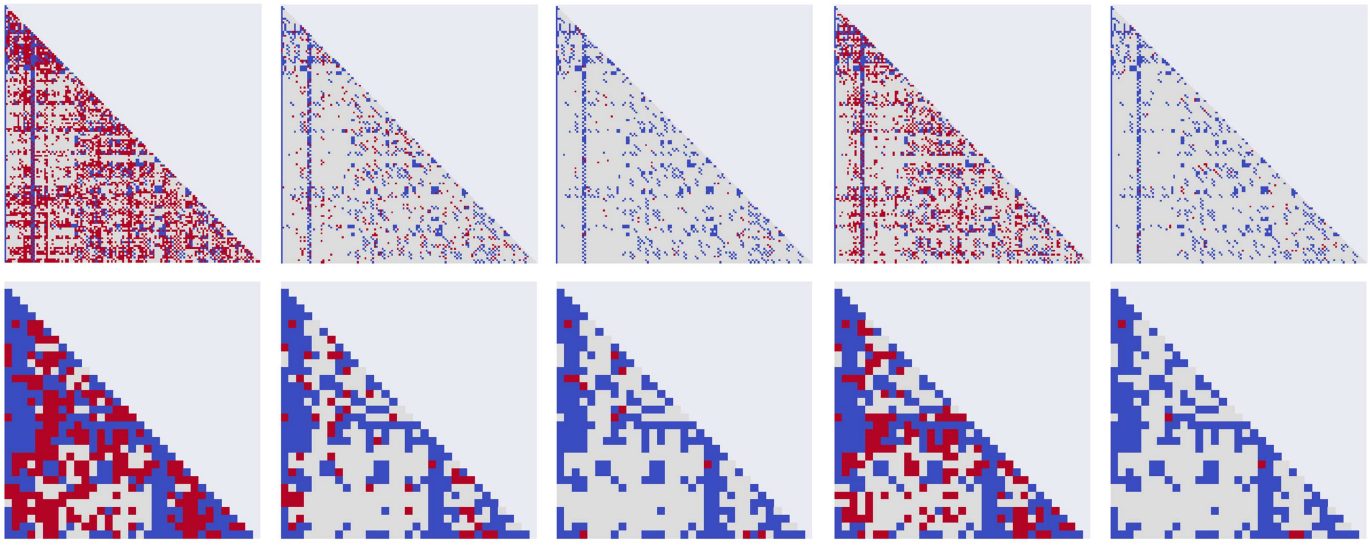

Fig. 6: Binary output-class adjacency matrices summarizing adjacencies seen anywhere on the MICCAI 2012 (top row) and IBSRv2 (bottom row) datasets. Blue denotes correct adjacencies, red forbidden ones. From left to right the methods are: (i) 2D without NonAdjLoss; (ii) 2D with NonAdjLoss; (ii) 2D with NonAdjLoss and semi-supervision; (iv) 2.5D with fusion; (v) 2.5D with fusion and semi-supervision. We report average score ± standard deviation.

work of (Painchaud et al., 2019) which shares the objective of removing inconsistencies from segmentation maps. It is interesting to observe that even though their method is completely different (post-processing with a variational auto-encoder to remove a set of 16 inconsistencies in cardiac segmentation), they observe systematic decreases in average Dice scores when applying their post-processing algorithm to a set of ten different segmentation methods.

Two visual examples of corrected adjacency anomalies are shown in Fig. 5, where small incorrect regions produced in the 2D baseline are ultimately suppressed. For the 2D Baseline the training time was 11.1 hours (300 epochs), fine-tuning with the NonAdjLoss required 8.5 hours (170 epochs), while semi-supervised training (100 images) required 46 hours (170 epochs). These execution times were measured on a server equipped with a single K80 GPU.

4.2. Multi-scale NonAdjLoss

In an attempt to make better use of the multi-scale nature of inter-structure distances, we briefly tested multi-scale adjacency penalties. For these experiments, the global constraint term is the sum of non-adjacency losses using several different neighborhoods (V in Eq. 4) with increasing sizes. We evaluated this on MICCAI12 for two different three-neighbourhood configurations with sizes given by $n_0 = \{3^3, 7^3, 11^3\}$ and $n_1 = \{3^3, 11^3, 15^3\}$. Table 2 shows that the segmentation and connectivity results were broadly similar to those for the corresponding single-scale loss (3×3×3 neighborhood), so we did not pursue this further.

MICCAI12	Dice	HD (mm)
Baseline 2.5D	0.733 ± 0.11	19.77 ± 9.52
Baseline 2.5D + fusion	0.738 ± 0.11	16.06 ± 7.11
NonAdjLoss(0) + fusion	0.736 ± 0.10	12.10 ± 4.75
NonAdjLoss(50) + fusion	0.744 ± 0.10	**10.19** ± **3.73**
IBSRv2	Dice	HD (mm)
Baseline 2.5D	0.832 ± 0.11	14.48 ± 16.00
Baseline 2.5D + fusion	0.834 ± 0.11	12.60 ± 14.60
NonAdjLoss(0) + fusion	0.837 ± 0.10	**9.71** ± **10.38**
NonAdjLoss(50) + fusion	0.835 ± 0.10	10.94 ± 13.96

Table 4: Distance and similarity metrics for each model. HD denotes Hausdorff distance. Each metric is averaged over the test dataset and we report average score ± standard deviation.

4.3. 2.5D Architecture

To see whether more of the 3D connectivity information from the brain images could be exploited at a reasonable cost, we trained the baseline CNN with the proposed 2.5D output augmentation. During inference we applied a simple fusion based post-processing strategy, that consists in summing the overlapping maps while sliding the window over the entire vol-

MICCAI12	Dice	HD (mm)	MSD (mm)	CA^{unique}	CA^{volume}
Baseline 2D	0.740 ± 0.11	20.93 ± 9.50	1.18 ± 0.40	5.1e-2 ± 6.8e-2	1.8e-2 ± 4.9e-2
Baseline 2.5D + fusion	0.738 ± 0.11	16.06 ± 7.11	1.20 ± 0.39	1.9e-2 ± 3.1e-2	8.0e-3 ± 2.6e-2
2.5D + NonAdjLoss(0) + Fus.	0.736 ± 0.10	12.10 ± 4.74	1.15 ± 0.38	4.0e-3 ± 1.1e-2	1.3e-3 ± 5.9e-3
2.5D + NonAdjLoss(50) + Fus.	0.744 ± 0.10	**10.19 ± 3.73**	**1.05 ± 0.34**	3.2e-4 ± 1.4e-3	4.4e-5 ± 2.2e-4
2.5D + NonAdjLoss(50) + Fus. + M-O	0.734 ± 0.10	10.27 ± 4.02	1.10 ± 0.34	**1.7e-4 ± 5.4e-4**	**1.2e-5 ± 4.4e-5**
IBSRv2	Dice	HD (mm)	MSD (mm)	CA^{unique}	CA^{volume}
Baseline 2D	0.833 ± 0.11	15.99 ± 15.27	0.78 ± 0.37	1.0e-1 ± 8.8e-2	1.5e-3 ± 3.0e-3
Baseline 2.5D + fusion	0.834 ± 0.11	12.60 ± 14.60	0.78 ± 0.34	5.6e-2 ± 5.5e-2	1.5e-3 ± 2.5e-2
2.5D + NonAdjLoss(0) + Fus.	0.837 ± 0.10	**9.71 ± 10.39**	**0.75 ± 0.33**	5.3e-3 ± 1.2e-2	6.2e-5 ± 2.1e-4
2.5D + NonAdjLoss(50) + Fus.	0.835 ± 0.10	10.94 ± 13.96	0.76 ± 0.32	5.3e-4 ± 2.8e-3	1.3e-6 ± 7.0e-6
2.5D + NonAdjLoss(50) + Fus. + M-O	0.836 ± 0.10	9.99 ± 11.45	0.76 ± 0.35	**4.2e-4 ± 2.3e-3**	**1.3e-6 ± 7.0e-6**

Table 5: Distance, similarity and connectivity metrics for each model. HD denotes Hausdorff Distance and MSD denotes Mean Surface Distance, both in millimeters. For each model we report average score ± standard deviation. The Dice, HD and MSD scores were averaged over the stated test set while the CA^{unique} and CA^{volume} ones were averaged over 30 test images from OASIS. The same CNN architectures were used for the baseline models and the NonAdjLoss(n) ones, with the baseline models trained using L_{seg} loss alone and the NonAdjLoss(n) ones including NonAdjLoss penalization and n additional images of semi-supervision.

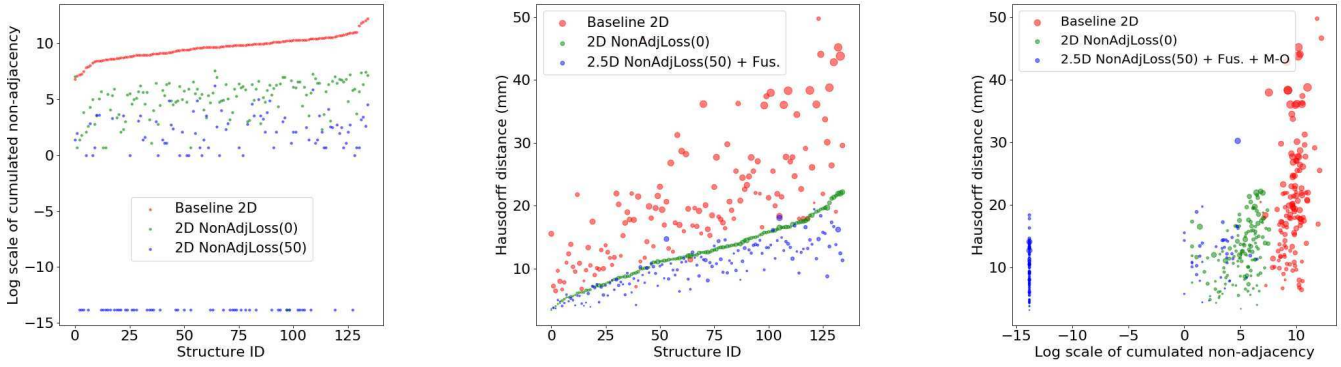

Fig. 7: Some illustrative non-adjacency statistics for each anatomical region on 30 test images from OASIS, using various models trained on MICCAI12. The log total adjacency error counts of regions without errors are set to −14. For Hausdorff Distances the point diameters are proportional to their standard deviations. The regions are ordered by their (left) error frequency on Baseline and (center) Hausdorff Distance on NonAdjLoss(0).

ume. Table 4 shows that the fusion post-processing reduces outliers, leading to decreased Hausdorff Distances. Combining the 2.5D model, NonAdjLoss with semi-supervision, and post-processing gives the best results seen in our experiments. Fig. 7 (center) shows the test-set average Hausdorff Distance for each anatomical structure, comparing the 2D baseline against 2D $NonAdjLoss(0)$ and 2.5D $NonAdjLoss(50)$ with fusion: the latter has lower errors than the others for almost all anatomical regions.

4.4. Oriented Non-adjacency Loss

The oriented NonAdjLoss provides finer-grained anatomical constraints but training a model using all eight 3D orientations requires full 3D output maps to be available. In practice we used our proposed 2.5D networks with a 6-orientation NonAdjLoss restricted to their 3 adjacent output slices, with the averaging fusion strategy. Table 5 shows that although the oriented loss improves the *CA* non-adjacency scores on both datasets, it slightly degrades the Dice, Hausdorff Distance and MSD scores. However for the semi-supervised training of the multi-oriented loss the setup was not identical to that for $NonAdjLoss(50)$: in order to to evaluate 6 losses instead of one, we were forced to reduce the batch size due to GPU memory constraints. The small degradations seen might also be due to sub-optimal hyperparameter settings, such as the learning rate or network's weights initialization. In the longer term we think that the oriented loss will prove to be a useful technique for anatomical applications. Fig. 7 (right) shows a scatter plot of average Hausdorff Distance versus log total non-adjacency counts for each anatomical region. This suggests that for most regions, reducing their connectivity errors also decreases their Hausdorff Distances significantly. Indeed the 2.5D oriented $NonAdjLoss(50)$ with fusion brings many of the regions down to zero adjacency errors. However even for these regions some of the usual Hausdorff Distance (inter-region boundary location) errors persist.

5. Conclusion

We have introduced NonAdjLoss, a loss constraint that suppresses known-forbidden region adjacencies in anatomical segmentations. Only the network training procedure is changed: the underlying network architecture remains unchanged and there is no additional cost during inference. Although the method had little effect on the Dice segmentation quality scores,

it clearly improved the Hausdorff Distance, Mean Surface Distance and connectivity metrics. It should be especially valuable for complex anatomical segmentation problems such as cortical region labelling because increasing the number of anatomical regions also increases the number of active constraints. The method's ability to handle partly unannotated data during training is another major advantage, allowing models to be trained on larger datasets.

Acknowledgments.

This work was funded by the CNRS PEPS "APOCS". It was performed within the framework of the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon under the program "Investissements d'Avenir"(ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). We gratefully acknowledge the support of NVIDIA Corporation, who donated the Titan X Pascal GPU used for this research. We would also like to thank the IN2P3 computing center for sharing their resources.

References

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38, 95–113.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence .

BenTaieb, A., Hamarneh, G., 2016. Topology aware fully convolutional networks for histology gland segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 460–468.

de Brebisson, A., Montana, G., 2015. Deep Neural Networks for Anatomical Brain Segmentation. arXiv:1502.02445 [cs, stat] URL: http://arxiv.org/abs/1502.02445. arXiv: 1502.02445.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. CoRR abs/1606.00915. URL: http://arxiv.org/abs/1606.00915.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54, 940 – 954. URL: http://www.sciencedirect.com/science/article/pii/S1053811910011997, doi:https://doi.org/10.1016/j.neuroimage.2010.09.018.

Frau-Pascual, A., Fogarty, M., Fischl, B., Yendiki, A., Aganj, I., 2019. Quantification of structural brain connectivity via a conductance model. NeuroImage URL: http://www.sciencedirect.com/science/article/pii/S1053811919300333, doi:https://doi.org/10.1016/j.neuroimage.2019.01.033.

Ganaye, P.A., Sdika, M., Benoit-Cattin, H., 2018. Towards integrating spatial localization in convolutional neural networks for brain image segmentation, in: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, IEEE. pp. 621–625.

Ghafoorian, M., Karssemeijer, N., Heskes, T., Bergkamp, M., Wissink, J., Obels, J., Keizer, K., de Leeuw, F.E., Ginneken, B., Marchiori, E., Platel, B., 2017. Deep multi-scale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin. NeuroImage: Clinical 14, 391 – 399. URL: http://www.sciencedirect.com/science/article/pii/S2213158217300311, doi:https://doi.org/10.1016/j.nicl.2017.01.033.

Jimenez-del-Toro, O., Müller, H., Krenn, M., Gruenberg, K., Taha, A.A., Winterstein, M., Eggel, I., Foncubierta-Rodríguez, A., Goksel, O., Jakab, A., Kontokotsios, G., Langs, G., Menze, B.H., Salas Fernandez, T., Schaer, R., Walleyo, A., Weber, M., Dicente Cid, Y., Gass, T., Heinrich, M., Jia, F., Kahl, F., Kechichian, R., Mai, D., Spanier, A.B., Vincent, G., Wang, C., Wyeth, D., Hanbury, A., 2016. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. IEEE Transactions on Medical Imaging 35, 2459–2475. doi:10.1109/TMI.2016.2578680.

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.B., 2018. Constrained-cnn losses forweakly supervised segmentation. arXiv preprint arXiv:1805.04628 .

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials, in: Advances in neural information processing systems, pp. 109–117.

Landman, B., . Miccai 2012 workshop on multi-atlas labeling. URL: https://my.vanderbilt.edu/masi/workshops/.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2010. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. Journal of cognitive neuroscience 22, 2677–2684.

Moeskops, P., Viergever, M.A., Mendrik, A.M., Vries, L.S.d., Benders, M.J.N.L., Išgum, I., 2016. Automatic Segmentation of MR Brain Images With a Convolutional Neural Network. IEEE TMI 35, 1252–1261. doi:10.1109/TMI.2016.2548501.

Nocedal, J., Wright, S., 2006. Numerical optimization, chapter 17. Springer Science & Business Media.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T., O'Regan, D.P., Kainz, B., Glocker, B., Rueckert, D., 2018. Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation. IEEE TMI 37, 384–395. doi:10.1109/TMI.2017.2743464.

Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A., Jodoin, P.M., 2019. Cardiac MRI Segmentation with Strong Anatomical Guarantees, in: MICCAI, Springer International Publishing.

Ravishankar, H., Venkataramani, R., Thiruvenkadam, S., Sudhakar, P., Vaidya, V., 2017. Learning and incorporating shape models for semantic segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 203–211.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Roy, A.G., Conjeti, S., Sheet, D., Katouzian, A., Navab, N., Wachinger, C., 2017. Error corrective boosting for learning fully convolutional networks with limited data, in: MICCAI, Springer. pp. 231–239.

Sdika, M., 2008. A fast nonrigid image registration with constraints on the Jacobian using large scale constrained optimization. IEEE Transactions on Medical Imaging 27, 271–81. URL: https://hal.archives-ouvertes.fr/hal-01902507, doi:10.1109/TMI.2007.905820.

Sdika, M., 2010. Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote. Med Image Anal 14, 219–26. URL: https://hal.archives-ouvertes.fr/hal-00617791. 1361-8423 (Electronic) 1361-8415 (Linking) Journal Article Research Support, Non-U.S. Gov't.

Sdika, M., 2013. A Sharp Sufficient Condition for B-Spline Vector Field Invertibility. Application to Diffeomorphic Registration and Interslice Interpolation. SIAM Journal on Imaging Sciences 6, 2236–2257. URL: https://hal.archives-ouvertes.fr/hal-01902498, doi:10.1137/120879920.

Simard, P.Y., Steinkraus, D., Platt, J.C., 2003. Best practices for convolutional neural networks applied to visual document analysis, in: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., pp. 958–963. doi:10.1109/ICDAR.2003.1227801.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Cardoso, M.J., Arbel, T., Carneiro, G., Syeda-Mahmood, T., Tavares, J.M.R., Moradi, M., Bradley, A., Greenspan, H., Papa, J.P., Madabhushi, A., Nascimento, J.C., Cardoso, J.S., Belagiannis, V., Lu, Z. (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham. pp. 240–248.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4itk: improved n3 bias correction. IEEE TMI 29, 1310–1320.

Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2009. Diffeomorphic

demons: Efficient non-parametric image registration. NeuroImage 45, S61–S72.

Wang, H., Yushkevich, P., 2013. Multi-atlas segmentation with joint label fusion and corrective learning—an open source implementation. Frontiers in neuroinformatics 7, 27.

Worth, A., . Internet Brain Segmentation Repository. URL: `https://www.nitrc.org/projects/ibsr/`.

Xu, J., Zhang, Z., Friedman, T., Liang, Y., den Broeck, G.V., 2018. A semantic loss function for deep learning with symbolic knowledge. URL: `https://openreview.net/forum?id=HkepKG-Rb`.