Medical Image Analysis (2025)

Contents lists available at ScienceDirect

Medical Image Analysis

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

An End-to-end Approach to Segmentation in Medical Images with CNN and Posterior-CRF

Shuai Chen[a,*], Zahra Sedghi Gamechi[a], Florian Dubost[a], Gijs van Tulder[a], Marleen de Bruijne[a,b]

[a]*Biomedical Imaging Group Rotterdam, Department of Radiology & Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands.*
[b]*Machine Learning Section, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark.*

ARTICLE INFO

Article history:
Received – – 2021
Received in final form – – 2021
Accepted – – 2021
Available online – – 2021

Keywords: Segmentation, CNN, CRF, Graph model, Medical images

ABSTRACT

Conditional Random Fields (CRFs) are often used to improve the output of an initial segmentation model, such as a convolutional neural network (CNN). Conventional CRF approaches in medical imaging use manually defined features, such as intensity to improve appearance similarity or location to improve spatial coherence. These features work well for some tasks, but can fail for others. For example, in medical image segmentation applications where different anatomical structures can have similar intensity values, an intensity-based CRF may produce incorrect results. As an alternative, we propose *Posterior-CRF*, an end-to-end segmentation method that uses CNN-learned features in a CRF and optimizes the CRF and CNN parameters concurrently. We validate our method on three medical image segmentation tasks: aorta and pulmonary artery segmentation in non-contrast CT, white matter hyperintensities segmentation in multi-modal MRI, and ischemic stroke lesion segmentation in multi-modal MRI. We compare this with the state-of-the-art CNN-CRF methods. In all applications, our proposed method outperforms the existing methods in terms of Dice coefficient, average volume difference, and lesion-wise F1 score.

1. Introduction

After the breakthrough of deep learning in computer vision (Krizhevsky et al., 2012; He et al., 2016; Long et al., 2015), deep convolutional neural networks (CNNs) and their variants (Ronneberger et al., 2015; Çiçek et al., 2016; Kamnitsas et al., 2017) quickly started to dominate medical image segmentation, outperforming traditional machine learning methods in many applications (Yu et al., 2016; Bakas et al., 2018; Kuijf et al., 2019; Maier et al., 2015). To refine the prediction from the CNN, it is common to combine CNN with a conditional random field (CRF) (Krähenbühl and Koltun, 2011). By modeling pairwise relationships and interactions between voxel-wise variables over the whole image, the CRF can improve the coherence of the segmentation. In previous work, CRFs based on predefined features such as intensity similarity and spatial coherence have been used as an efficient post-processing technique or trained in an end-to-end manner in a recurrent neural network to refine the CNN outputs (Chen et al., 2017; Dou et al., 2017; Kamnitsas et al., 2017; Zheng et al., 2015).

Most often, a CRF uses a combination of voxel intensity and voxel location as pairwise potentials. Although this works well in several computer vision applications (Zheng et al., 2015; Schwing and Urtasun, 2015), there can be challenges in other applications. The approach assumes that voxels that have similar intensity and are close to each other in the image are likely to belong to the same class. There are many applications among others in medical image analysis in which this assumption does not hold. For example, the intensity-based features of the CRF

*Corresponding author. Email addresses: s.chen.2@erasmusmc.nl (S. Chen); marleendebruijne@erasmusmc.nl (M.d. Bruijne);

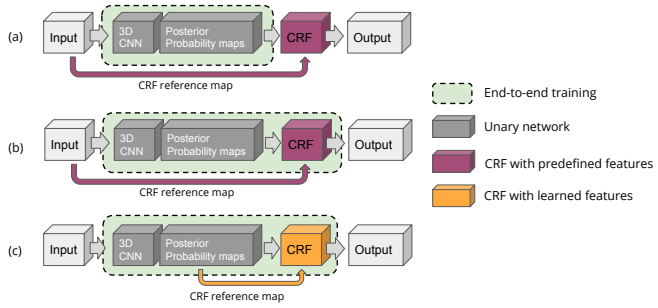

Fig. 1. Different CRF-based approaches For each graph: (a) Post-processing CRF (Chen et al., 2017; Kamnitsas et al., 2017); (b) End-to-end training CRF with predefined features (Zheng et al., 2015); (c) Proposed Posterior-CRF, which uses CNN feature maps as CRF reference maps. Best viewed in color with zoom.

are not sufficient for problems where the intensity is not informative enough to identify object boundaries, such as the artery segmentation problem in Figure 2a. The spatial component of the CRF, on the other hand, requires extra careful tuning when the CRF is applied to data with isolated small objects, such as the white matter hyperintensities in Figure 2b, which may be erroneously removed by excessive smoothing. In stroke lesion segmentation, a large appearance difference between lesion objects of the same class also goes against the CRF assumption that the same class objects should have similar intensity (see Figure 2c).

In this paper, we propose *Posterior-CRF*, a new learning-based CRF approach for image segmentation that allows the CRF to use features learned by a CNN, optimizing the CRF and CNN parameters concurrently. The learning-based CRF makes the CNN features update to work best with CRF in an end-to-end manner. During training, the CRF inference works in the CNN feature space, which is more likely to contain useful high-level features for segmentation compared to the original intensity values.

We demonstrate our method in three medical image analysis applications. Our first application is the segmentation of the aorta and pulmonary artery in non-contrast, non-ECG-gated chest CT scans. In these images, the aorta and the pulmonary artery share similar intensity values, which goes against the CRF assumption that similar classes should share similar intensity (Sedghi Gamechi et al., 2018; Xie et al., 2014). The boundaries between the objects are not recognizable by intensity alone, making a standard CRF less effective (Figure 2a). Our second application is the segmentation of white-matter hyperintensities in brain MRI. These small objects are sparsely distributed in the brain (see Figure 2b) and may be removed by the CRF, which optimizes for the spatial coherence of segmentation. Our third application is the segmentation of ischemic stroke lesions in brain MRI, which have very heterogeneous intensities and shapes within the same lesion class (Figure 2c).

Contributions

1. We present a new end-to-end trainable algorithm for image segmentation called *Posterior-CRF* using learnable features in CRF pairwise potentials. We explore how the proposed method affects CNN learning during training.

2. We compare the performance of a fully-connected CRF in several settings: post-processing, end-to-end training with predefined features, and end-to-end training with learned features. Ablation experiments are conducted to investigate the influence of CRF parameters and which level of the CNN feature maps are more likely to benefit the CRF inference. We found that the features in the last CNN feature maps provide a more consistent improvement than features in early CNN layers and predefined intensity features.

3. We evaluate our methods in three applications: aorta and pulmonary artery segmentation in non-contrast CT, which can be used to compute important biomarkers such as the pulmonary artery to aorta diameter ratio (Sedghi Gamechi et al., 2018); white matter hyperintensities segmentation in multi-sequence MRI, which is of key importance in many neurological research studies (Kuijf et al., 2019); and ischemic stroke lesion segmentation in multi-sequence MRI, which can provide biomarkers for stroke diagnosis (Maier et al., 2015). In the experiments, the proposed Posterior-CRF outperforms CNN without CRF, post-processing CRF, end-to-end intensity-based CRF, and end-to-end spatial-based CRF.

A preliminary version of this work, focused on a single application and with less validation, appeared as an extended abstract in (Chen and de Bruijne, 2018).

2. Related Work

2.1. End-to-end Training of CRF and CNN

CRF is widely used as an efficient post-processing method to refine the output of CNN segmentation models (for example, (Chen et al., 2017; Dou et al., 2017; Kamnitsas et al., 2017)). However, applying a CRF as post-processing means that the CNN is not able to adapt its output to the CRF. Zheng et al. (Zheng et al., 2015) proposed to optimize CNN and CRF jointly by reformulating the CRF inference as a recurrent neural network (RNN) operation, such that the CRF weights can be learned together with the CNN. This approach makes the unary potentials and the kernel weights in pairwise potentials trainable, which saves the computational cost of grid search for other approaches to tune these weights, although the CRF still works in the predefined fixed feature space. In this paper, we focus on a new CRF approach where the CRF inference works in a learning-based CNN feature space.

2.2. Locally-connected CRFs with Learned Potentials

While conventional CRFs use predefined Gaussian edge potentials, the potentials can also be learned through a neural network. Vemulapalli et al. (Vemulapalli et al., 2016) learn the pairwise potentials of a Gaussian CRF in a bipartite graph structure. This approach uses a simpler continuous CRF model which provides better convergence of mean-field inference than the conventional discrete CRF models. In this paper, we focus on the most widely used discrete CRF model which is a natural fit for the dense segmentation problem. Lin et al. (Lin et al., 2016), Li et al. (Li and Ping, 2018) and Wang et al. (Wang et al., 2018a) learn pairwise CRF potentials to model patch-wise (or

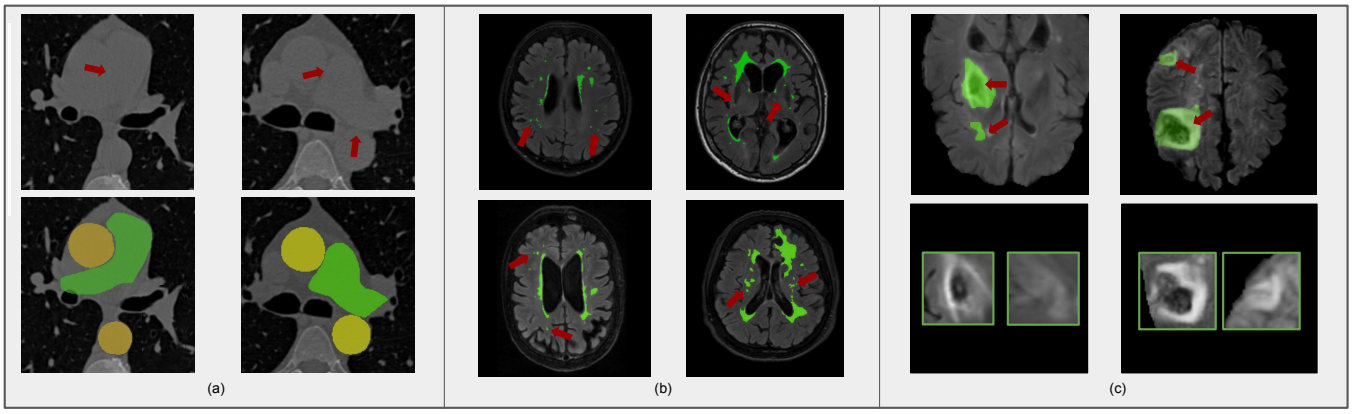

Fig. 2. Difficult cases for conventional CRF inference in medical image segmentation. (a) Segmentation of arteries in CT: first row shows two axial slices of the CT scan with red arrows indicating indistinguishable boundaries; second row shows the corresponding ground truth of the aorta (yellow) and pulmonary artery (green); (b) White matter hyperintensities segmentation in MRI: four examples are shown with the ground truth of the lesions (green), red arrows indicate small isolated lesions that can be easily removed by CRF; (c) Ischemic stroke lesions segmentation in MRI: first row shows the ground truth of the lesions (green) where large appearance difference between lesions can be observed (red arrows); second row shows a close-up view of the lesions. Best viewed in color with zoom.

local) relationships using free form functions learned by neural network rather than a combination of predefined Gaussians to calculate the pairwise potentials. The patch-wise potentials provide a better ability to model the semantic compatibility between image regions and have different effects compared to our approach, where we do not consider patch-wise relationships. Our method uses traditional Gaussian edge potentials (Krähenbühl and Koltun, 2011) similar to Zheng et al. (Zheng et al., 2015) which are easier to compute in a fully-connected manner. Unlike Zheng et al., we derive the potentials from the feature space learned by a CNN. This allows us to model global interactions between voxel-wise variables using learning-based features.

2.3. Other Methods Related to CRF

Next to CRF, there are several other approaches that aim to model interactive relationships or add global information to neural networks. Graph neural networks (GNN) (Scarselli et al., 2008; Selvan et al., 2018) model interactions between variables by applying graph convolution filters, which allow them to learn global relationships between voxels. We further address GNN in the Discussion. The recently proposed non-local CNN (Wang et al., 2020) uses layer-wise self-attention (Vaswani et al., 2017; Wang et al., 2018; Yuan et al., 2019) to make each layer in the network focus on the areas that encoded the most non-local information in the preceding layer. While this allows non-local CNNs to model long-range dependencies, they are unable to model the interactions that can be learned by a CRF or GNN. In this paper, we focus on the fully-connected CRF model which is an efficient approach of modeling both interactive relationships and global information.

3. Methodology

Our method consists of two parts that are optimized jointly: 3D CNN and 3D CRF. In Section 3.1, we describe the CNN model, which provides unary potentials for the CRF inference as well as features for the pairwise potentials for the proposed Posterior-CRF. Then we introduce the CRF in Section 3.2. We show two previously proposed ways to perform CRF inference using predefined features: post-processing (Section 3.3.1) and end-to-end training with predefined features (Section 3.3.2). Our proposed end-to-end training with learned features is presented in Section 3.4, followed by Section 3.4.1 about the back-propagation of the proposed learning-based CRF. The mean-field inference algorithm used in the proposed method is explained in Appendix Section 8.

3.1. CNN Model

Our CNN model is based on UNet (Ronneberger et al., 2015), the most widely used network architecture for medical image segmentation. It has a multi-scale design with skip-connections that connect the encoding and decoding parts of the network, which allow the decoding path to use the early, high resolution feature maps without losing information through pooling. We use 3D UNet as the basic CNN architecture to provide the unary potentials for CRF inference as well as features for the pairwise potentials for the proposed Posterior-CRF. Details of the network layout used in our experiments are given in Figure 3.

3.2. Conditional Random Fields

In this section, we describe the CRF as proposed in (Krähenbühl and Koltun, 2011). In image segmentation, a CRF models voxel-wise variable x_i taking values in $\{1, ..., C\}$ as a set of random variables $\mathcal{X} = \{x_1, ..., x_N\}$, where C is the number of classes and N is the number of voxels in the image. During training, x_i is converted into a soft classification vector of length C, indicating for each class the probability that the ith voxel belongs to that class, with the L_1 norm $|x| = 1$. x_i obey a Markov property conditioned on a global observation, the image $\mathbf{I}$ consisting of variables $\mathcal{I} = \{I_1, ..., I_N\}$. In this paper, $\mathbf{I}$

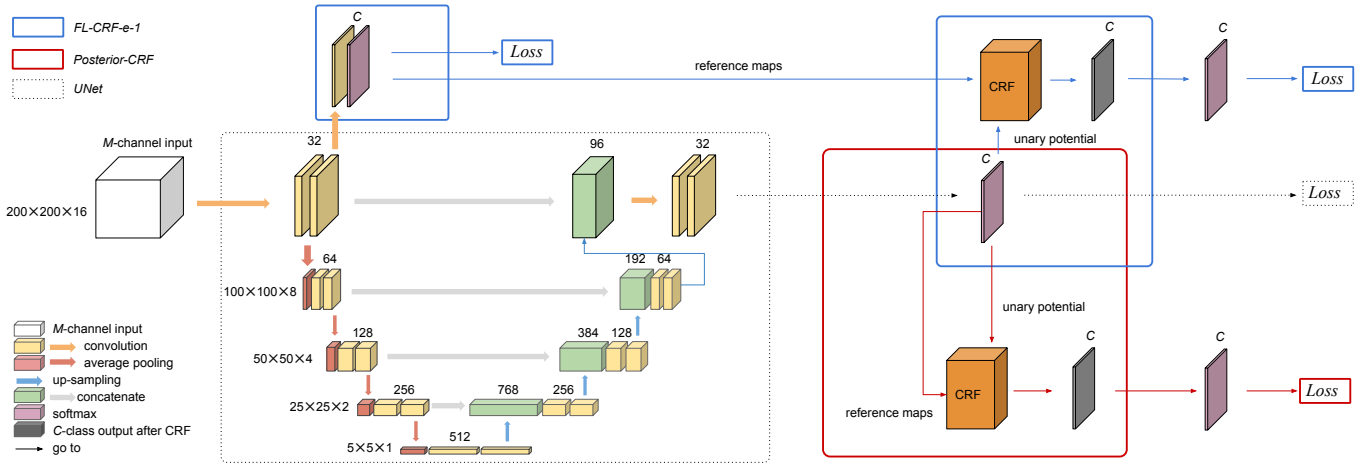

Fig. 3. Proposed feature-learning-based CRF using early/later CNN feature maps. The backbone architecture is based on 3D UNet. The skip-connections concatenate the feature maps from the encoder path with the upsampled ones from the decoder path. The CRF module is placed on top of the CNN and infers the most likely posterior class probability conditioned on the CRF features. M is the number of input imaging modalities. C is the number of output classes. Two proposed CRF variants are shown in this figure: 1. *Posterior-CRF* (red rectangle and arrows), which uses the last CNN layer as CRF reference maps; 2. *FL-CRF-e-1* (blue rectangles and arrows), which uses the first level CNN layer as CRF reference maps. Best viewed in color with zoom.

is the observed 3D CT/MRI scans, with its length given by the number of imaging modality channels M times the number of voxels per channel N.

Consider a fully-connected pairwise CRF model $(\mathbf{X},\mathbf{I})$ characterized by a prior Gibbs distribution:

$$P(\mathbf{X}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})}\exp(-\sum_{c\in C_\zeta}\phi_c(\mathbf{X}_c|\mathbf{I})) \tag{1}$$

where $\zeta = (\mathcal{V},\mathcal{E})$ is an undirected graph describing the random field $\mathbf{X}$. Each clique c in a complete set of unary and pairwise cliques C_ζ in ζ, and ϕ is the potential for each clique. We seek a maximum a posteriori probability (MAP) estimation $\mathbf{x}$ that minimizes the corresponding Gibbs energy $E(\mathbf{X}=\mathbf{x}|\mathbf{I})$:

$$E(\mathbf{X}=\mathbf{x}|\mathbf{I}) = \sum_i \varphi_u(x_i|\mathbf{I}) + \sum_{i<j}\varphi_p(x_i,x_j|\mathbf{I}) \tag{2}$$

$$MAP(P(\mathbf{X}|\mathbf{I})) : \mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} E(\mathbf{X}=\mathbf{x}|\mathbf{I}) \tag{3}$$

where i and j range from 1 to N. The first term $\varphi_u(x_i)$ in Equation 2 is the unary potential, which in our case is the current C length vector of voxel i representing the class probabilities in the CNN posterior probability maps. The second term $\varphi_p(x_i,x_j)$ is the pairwise potential:

$$\varphi_p(x_i,x_j) = \mu(x_i,x_j)\sum_{m=1}^{K}\omega_m k_m \tag{4}$$

where $\mu(x_i,x_j)$ is the label compatibility function that describes the interactive influences between different pairs of classes, ω_m is the linear combination weight of different pre-defined kernels k_m and K is the total number of kernels. Each k_m is a modified Gaussian kernel with specific feature vector $\mathbf{f}$:

$$k(\mathbf{f}_i,\mathbf{f}_j) = \prod_{s=1}^{S}\mathbf{exp}(-\frac{1}{2}(f_i^s - f_j^s)^{\mathrm{T}}\mathbf{\Lambda}^s(f_i^s - f_j^s)) \tag{5}$$

The feature vector $\mathbf{f}$ is defined from S arbitrary feature spaces. $\mathbf{\Lambda}$ is a symmetric positive-definite precision matrix that defines the shape of each kernel. In semantic segmentation, typically a combination of intensity (I) and position features (p) has been used (Krähenbühl and Koltun, 2011; Zheng et al., 2015; Kamnitsas et al., 2017):

$$\begin{aligned}\varphi_p(x_i,x_j) =& \mu(x_i,x_j)[\omega_1\mathbf{exp}(-\frac{|p_i-p_j|^2}{2\theta_\alpha^2} - \frac{|I_i-I_j|^2}{2\theta_\beta^2}) \\ &+ \omega_2\mathbf{exp}(-\frac{|p_i-p_j|^2}{2\theta_\gamma^2})]\end{aligned} \tag{6}$$

where the first kernel controlled by ω_1 is called *appearance kernel* and the second kernel controlled by ω_2 is called *smoothness kernel*. The parameters θ_α, θ_β and θ_γ control the influence of the corresponding feature spaces. The appearance kernel is inspired by the observation that nearby voxels with similar intensity are likely to be in the same class, while voxels that are either further away or have larger intensity difference are less likely to be in the same class. The smoothness kernel can remove isolated regions and produce smooth segmentation results (Krähenbühl and Koltun, 2011; Kamnitsas et al., 2017). Note that the position feature appears in both appearance kernel and smoothness kernel, where spatial information has different contributions to each of the two kernels, depending on the spatial standard deviations θ_α and θ_γ.

3.3. CRF with Predefined Features

Conventional CRFs use predefined features, such as the image intensity and spatial position shown in Equation 6. These features are commonly used in CRFs to encourage intensity and spatial coherence, based on the assumption that voxels that have a similar intensity or are close together are likely to belong to the same class.

We evaluate two state-of-the-art approaches to combine CRFs with predefined features with a CNN: 1. Apply the CRF

as post-processing to refine the CNN outputs (Section 3.3.1); 2. Implement the CRF as a neural network layer that can be trained together with the CNN in an end-to-end manner (Section 3.3.2).

3.3.1. CRF as Post-processing

After we train a CNN model and get its predictions, we can apply CRF as a post-processing method to refine the results (Chen et al., 2017). We refer to this method as *Postproc-CRF* (Figure 1a).

3.3.2. End-to-end Training CRF

The CNN and CRF can be combined more elegantly by optimizing them together in an end-to-end manner (Zheng et al., 2015) (Figure 1b), which allows the CRF to influence the CNN optimization. The end-to-end CRF uses the same pairwise potentials as that in the post-processing CRF (Equation 6). We refer to this variant as *Intensity-CRF*.

To investigate the spatial term in the end-to-end CRF, we can also use only the position features as the CRF feature space, which means that the CRF layer will only encourage nearby voxels to have the same class. We implement this CRF by setting the weight of the appearance kernel ω_1 to zero and make it not trainable. We refer to this method as *Spatial-CRF*.

3.4. Proposed CRF with Learning-based Features

Our proposed CRF uses a learning-based feature space. We replace the intensity feature vector I in the CRF kernel (Equation 6) with the new feature vector $F(\mathbf{I})$ from the CNN feature maps. The information in these CNN feature maps differs per level: in the first level of UNet the feature maps contain information close to the intensity, while in the last level of the UNet they contain more context for each voxel and potentially more class-discriminative information.

We refer to the CRF that uses features learned by CNN as *feature-learning-based CRF* (see Figure 1c) and refer to the specific form of CRF using the features in the last CNN softmax layer as *Posterior-CRF* (see Figure 3).

Unlike the CRFs with predefined features, our CRF takes CNN feature maps as the reference maps and updates the random field $\mathbf{X}$ based on $F(\mathbf{I})$ instead of on $\mathbf{I}$ directly. Compared to the original CRF pairwise potential in Equation 6, the feature I is replaced with $F(\mathbf{I})$ and the new pairwise potential becomes:

$$\begin{aligned}\varphi_p(x_i, x_j) =& \mu(x_i, x_j)[\omega_1 \mathbf{exp}(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \\ & \frac{|F_i(\mathbf{I}) - F_j(\mathbf{I})|^2}{2\theta_\beta^2}) + \omega_2 \mathbf{exp}(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2})]\end{aligned} \tag{7}$$

3.4.1. Back-propagation of the Learning-based CRF

The back-propagation of the proposed end-to-end feature-learning-based CRF is shown in Figure 4. There are five steps within one optimization iteration. Steps 1~3 are the forward process that generates the output of the CNN. In the 4th step, CRF weights will adapt to the outputs calculated by the reference maps and unary maps, both given by CNN feature maps before back-propagation. In the 5th step, CNN weights are updated to provide new unary maps and reference maps for CRF for the next iteration. When the optimization converges, both CNN and CRF weights become stable close to their optimal values. Note that the mean-field inference in CRF happens in the forward process (after step 2 and before step 3) and thus contributes to the gradient updates of both CNN and CRF weights. The derivation of the mean-field inference gradient is omitted due to the length of the paper and can be found in Section 4.2 of the paper by Zheng et al. (Zheng et al., 2015).

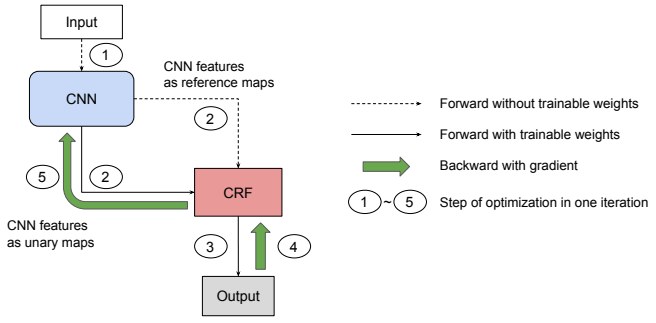

Fig. 4. One end-to-end optimization iteration of the proposed CRF method. Best viewed in color with zoom.

4. Experiments

In this section, we present experiments to evaluate the proposed method and compare it to the baseline methods: 3D UNet, Post-processing CRF, Intensity-CRF, and Spatial-CRF. Implementation details are discussed in Section 4.1, followed by the experimental settings (Section 4.2), the description of the datasets and pre-processing (Section 4.3), data augmentation and training details (Section 4.4) and evaluation metrics (Section 4.5).

4.1. Implementation

4.1.1. CNN Implementation

We implement all the algorithms in the TensorFlow framework. The detailed CNN architecture for the experiments is shown in Figure 3. All convolution layers use ReLU as the activation function except for the last output layer, which uses softmax to produce the final probability maps. For a fair comparison, the 3D UNet architecture that is tuned for the CNN baseline method is applied to all the CRF methods in Table 3. The 5-layer depth of UNet (tuned from 3 to 6) and 32 base feature maps (tuned from 8 to 64) are tuned based on all three datasets.

All segmentation models are optimized by minimizing the Dice loss (Isensee et al., 2020):

$$\mathcal{L}_{dc} = -\frac{2}{|C|}\sum_{c\in C}\frac{\sum_{i\in I} u_i^c v_i^c}{\sum_{i\in I} u_i^c + \sum_{i\in I} v_i^c} \tag{8}$$

where v_i^c is the predicted probability that voxel i belongs to the cth class. u_i^j is the true label. The loss is minimized using the Adam optimizer (Kingma and Ba, 2014).

4.1.2. CRF Implementation

In CRF, mean-field approximation can be used to calculate the maximum a posteriori probability (MAP) of the inference. We use an efficient approximation algorithm for mean-field inference (Krähenbühl and Koltun, 2011; Monteiro et al., 2018) built on a fast high-dimensional filtering using the permutohedral lattice (Adams et al., 2010) that allows voxel-wise fully-connected CRF to be iteratively computed in linear time. For a fair comparison, all the CRF methods in this paper are implemented in 3D fully-connected manner. The codes are publicly available: https://github.com/ShuaiChenBIGR/Posterior-CRF.

4.2. CRF Settings

4.2.1. Post-processing CRF

For *Postproc-CRF*, we fix the label compatibility μ in Equation 6 to the identity matrix, which means that the CRF does not model label-specific interaction. In the case of multi-modal input, each imaging modality has a specific θ_β to control the strength of the intensity term.

4.2.2. End-to-end CRF with Predefined Features

We consider two forms of end-to-end CRFs with predefined features: *Intensity-CRF* uses intensity of the input image $\mathbf{I}$ and position information as its feature space. *Spatial-CRF* uses only the position information (the smoothness term in Equation 6). The label compatibility is a $C \times C$ parameter matrix which is optimized during training to allow the CRF to learn the label compatibility automatically. The weights ω_1 of the appearance kernel for *Intensity-CRF* and ω_2 of the spatial kernel for *Spatial-CRF* are $C \times C$ matrices, which we restrict to diagonal matrices because the relationship between classes is already covered by the label compatibility matrix. Inner product is calculated by multiplying the matrices. For simplicity, only one θ_β is applied for all modalities.

4.2.3. End-to-end CRF with Learned Features

The proposed *Posterior-CRF* uses the last softmax layer of the CNN as its reference map. The hyperparameters are the same as end-to-end CRF with predefined features. Note that Posterior-CRF is a special case of the feature-learning-based CRF. We can also use early CNN feature maps as CRF reference maps. An ablation study investigating other CRF variants can be seen in Section 5.4.

4.2.4. CRF Parameters

Parameters in the post-processing CRF for each dataset were obtained by grid search on the validation set and are shown in Table 1. We computed results with 500 different configurations of Postproc-CRF on each dataset for grid-search. Parameters in the end-to-end CRFs (*Intensity-CRF*, *Spatial-CRF*, *Posterior-CRF*) are initialized with the same values as were used in post-processing CRF. Although the end-to-end CRF approaches have the ability to learn CRF weights automatically during training, we initialize all CRF approaches in the same way to facilitate visualization of the evolution of CRF parameters during training (see Figure 5). We study the sensitivity to different CRF parameter initializations in Section 5.3.

The initial label compatibility matrix is set to an identity matrix and can be optimized during training. In the multi-modality case, the initial value of θ_β is averaged over all modalities. The initial values for each dataset are shown in Table 2.

4.2.5. Computation Costs of CRF

The training and testing time of the proposed CRF method is the same as Intensity-CRF but a bit slower than Spatial-CRF, since there is no bilateral term in Spatial-CRF. Although the proposed CRF uses CNNs features to compute the pairwise potential, the gradients only flow through the unary map path but not the reference map path which is the same as that in traditional Intensity-CRF. Therefore, there is no additional time and memory cost of the proposed method compared to traditional end-to-end CRF approaches with fixed feature space. Post-processing CRF is after the CNN training and takes more time for inference compared to the end-to-end CRFs, since the inference is done by CPU but not GPU.

4.3. Datasets and Preprocessing

We evaluate the proposed method on three segmentation problems: CT arteries, MRI white matter hyperintensities, and MRI ischemic stroke lesions. We chose these problems to study the generalizability of the method as these applications differ a lot in object shapes and appearances, imaging modalities, and suffer from different problems (see Fig. 2).

4.3.1. CT Arteries Dataset

We use 25 non-contrast lung CT scans from 25 different subjects enrolled in the Danish Lung Cancer Screening Trial (DLCST) (Pedersen et al., 2009). The selection of the 25 subjects was completely random and it was done before the development of this algorithm for an unrelated study. The aorta and pulmonary artery were manually segmented by a trained observer (ZS). Images have an anisotropic voxel resolution of $0.78\text{mm} \times 0.78\text{mm} \times 1.00\text{mm}$ and are of size 512x512 with on average 336 slices (range 271-394). The 25 scans are split into three parts of 10, 5, and 10 scans for training, validation, and testing respectively. Due to the limitation of GPU memory, we first crop the original CT images and only keep the axial central part of 256×256 voxels for all slices. Then, 3D patches of the size $256 \times 256 \times 16$ are extracted from the cropped images. All training patches have 80% overlap in z-axis between neighboring patches to mitigate border effects. In total, there are 840 3D patches for training. We use the original CT intensities without normalization.

4.3.2. MRI White Matter Hyperintensities (WMH) Dataset

The White Matter Hyperintensities (WMH) Segmentation Challenge (Kuijf et al., 2019) provided images from 60 subjects (T1 and FLAIR) acquired from three hospitals and manually segmented for background and white matter hyperintensities. We randomly split these in 36 subjects for training, 12 for validation, and 12 for testing. For each subject, we cropped/padded MRI images into a constant size $200 \times 200 \times Z$, where Z is the number of slices in the image. We use Gaussian normalization to normalize the intensities inside the brain mask in each image

Table 1. Post-processing CRF parameters for each dataset. Search range indicates the range of parameter values explored during grid search.

Datasets	*CT Arteries*	*WMH*	*ISLES*	Search range
ω_1	6.39	3.85	9.75	(0.1, 10)
θ_α	4.09	4.46	8.74	(0.1, 10)
θ_β for CT	1.10	-	-	(0.1, 10)
θ_β for T1	-	7.01	9.26	(0.1, 10)
θ_β for T2	-	-	9.73	(0.1, 10)
θ_β for FLAIR	-	2.64	2.36	(0.1, 10)
θ_β for DWI	-	-	6.85	(0.1, 10)
ω_2	3.40	1.41	2.34	(0.1, 10)
θ_γ	4.83	0.11	1.35	(0.1, 10)
Iterations	3	1	2	(1, 5)

Table 2. Initial end-to-end CRF parameters for each dataset.

Methods	ω_1	θ_α	θ_β	ω_2	θ_γ	Iterations
CT Arteries						
Spatial-CRF	-	-	-	3.40	4.83	3
Others	6.39	4.09	1.10	3.40	4.83	3
WMH						
Spatial-CRF	-	-	-	1.41	0.11	1
Others	3.85	4.46	4.83	1.41	0.11	1
ISLES						
Spatial-CRF	-	-	-	2.34	1.35	2
Others	9.75	8.74	7.05	2.34	1.35	2

to zero mean and unit standard deviation. We extract training patches of size $200 \times 200 \times 16$ with 80% overlap in z-axis between patches. In total, there are 528 3D patches for training.

4.3.3. MRI Ischemic Stroke Lesions (ISLES) Dataset

The ISLES 2015 Challenge (Maier et al., 2017) is a public dataset of diverse ischemic stroke cases. There are 4 MRI sequences available for each patient (T1, T2, FLAIR, and DWI). We use the sub-acute ischemic stroke lesion segmentation (SISS) dataset (28 subjects) with the lesion labels for experiments and randomly split them as 14 for training, 7 for validation and 7 for testing. The images are cropped/padded to the size $200 \times 200 \times Z$. Gaussian normalization is applied for normalizing the intensities in each image. Training patches of the size $200 \times 200 \times 16$ with 80% overlap in z-axis are extracted. In total, there are 560 3D patches for training.

4.4. Data Augmentation and Training Details

The network is trained on all mini-batches (each mini-batch contains one 3D patch). For each 3D patch in the current mini-batch we apply 3D random rotation sampled from ([-5,5],[-5,5],[-10,10]) degrees, shifting ([-24,24],[-24,24],[-7,7]) voxels, as well as random horizontal (left and right) flipping. We stopped training when the validation loss is not decreasing anymore and chose the model that achieved the best validation performance. The experiments are run on an Nvidia GeForce GTX1080 GPU. The average training time is 5∼10 hours for one CNN baseline model and 1∼2 hours more when the CRF layer is added.

4.5. Evaluation Metrics

We use four voxel-wise metrics of segmentation quality: Dice similarity coefficient (DSC), indicating the relative overlap with the ground truth (larger is better); 95th percentile Hausdorff distance (H95), showing the extremes in contour distance from ground truth to the prediction (smaller is better); Average volume difference (AVD) as a percentage of the difference between ground truth volume and segmentation volume over ground truth volume (smaller is better), and Recall score (larger is better). For the lesion segmentations (WMH and ISLES), we additionally assess accuracy of lesion detection by computing the lesion-wise Recall and lesion-wise F1 score (larger is better). The lesion-wise metrics use the 3D connected components, while the voxel-wise metrics do not use 3D connected components. The correct detection of a lesion is determined by the overlap (at least one voxel) of the 3D components. F1 score is equivalent to lesion-wise Dice score and is calculated by 2*(precision*recall)/(precision+recall), where precision is calculated by true positives/(true positives+false positives).

5. Results

5.1. Segmentation Results

Table 3 shows the segmentation results for all three datasets. In most metrics, Posterior-CRF had the best performance in all datasets. For all datasets, CNN without CRF provides good baseline results, which indicates that 3D UNet is an efficient architecture to extract useful features for segmentation in these applications. Intensity-CRF performed worse on DSC than Posterior-CRF (statistically significant in aorta segmentation and WMH segmentation), which reveals the limitation of intensity features. Among all end-to-end CRF methods, Spatial-CRF performs worst for all datasets except ISLES. From these results, we conclude that spatial coherence alone is not sufficient and often detrimental to segmentation accuracy, and that the CNN features in the last layer are more informative for CRF than the intensity features in the original images.

CRFs that depend strongly on intensity-based features have difficulties detecting objects that are similar in intensity. Examples of this problem can be observed in the segmentations for the CT arteries and ISLES datasets (Figure 6). In CT arteries segmentation, the aorta and pulmonary artery have very similar intensities, which causes most of the methods in our experiments to sometimes misclassify part of the aorta as pulmonary artery. This is especially true for Post-processing CRF but also for Intensity-CRF.

Posterior-CRF achieves a DSC segmentation overlap of 95.4% and an H95 lower than 2.87mm in aorta segmentation, which is significantly better than all other methods on this dataset. We argue that this is because the features from the last CNN feature maps are more informative than the intensity-based features, which allows the CRF inference to focus on refining the object boundary without expanding into neighboring

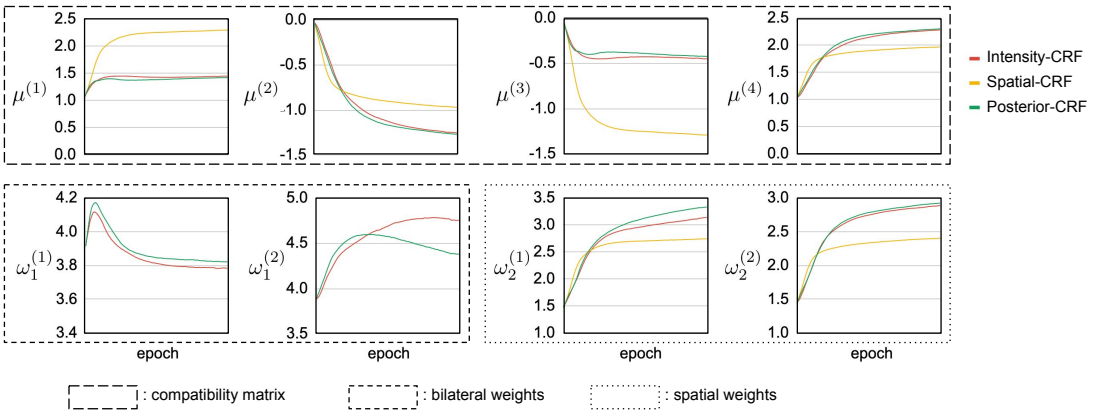

Fig. 5. CRF parameters during training in WMH dataset. The initial values of the CRF parameters can be found in Table 2. Best viewed in color with zoom.

class voxels with similar intensities. The Posterior-CRF also gives a performance improvement in the segmentation of the pulmonary artery, but this is not always statistically significant. One reason is that the blurred boundary between the aorta and pulmonary artery often results in the oversegmentation of pulmonary artery, the errors in pulmonary artery are emphasized because the overall pulmonary artery volume is lower. Another reason could be the curved shape of the pulmonary artery, which makes the results vary a lot between patients.

We see similar behavior on the ISLES dataset. The intensity boundaries of the large ischemic stroke lesions are ambiguous and their appearance varies a lot between lesions. Most of the methods fail to segment the boundaries accurately (see Figure 6 ISLES). Post-processing CRF hardly solves the problem and performs slightly worse than CNN. Posterior-CRF achieves better (while less significant due to the large prediction variance between samples) segmentation performance on DSC, AVD, lesion-wise F1.

A properly tuned spatial component of the post-processing CRF can benefits CT arteries and ischemic stroke lesion segmentation (Appendix Section 9, Figure 2 (a) and (c)). However, it can cause problems to white matter hyperintensities no matter how we try to tune it (Appendix Section 9, Figure 2 (b)), where we can see a positive ω_2 always leads to a decreased performance since the spatial smoothing contributes to remove both isolated true positives and false positives if they are small enough. The complete SHAP analysis will be discussed in Appendix Section 9.

The negative effect of the spatial smoothing results in the low average lesion-wise recall score in WMH segmentation for Postproc-CRF (34.8%) and can be observed in the WMH segmentation results (see Figure 6). In this case, Postproc-CRF is always worse than vanilla CNN (within our grid-search range). This is because the scenario where post-processing CRF has no influence (with both ω_1 and ω_2 set to zero) was not included in the grid search range (0.1,10). Intensity-CRF has a higher lesion-wise average recall than CNN baseline (68% to 64.8%) but a lower (not significantly) voxel-wise recall (77.5% to 79.8%): although it detects more correct lesions than CNN due to the intensity features, its use of spatial features causes it to undersegment individual lesions (see Figure 6). Spatial-CRF also suffers from this problem, with a high lesion-wise recall of 68.8% but low lesion-wise F1 of 65.7%.

For CT arteries, the proposed method performs better than the state-of-the-art (Sedghi Gamechi et al., 2018) in aorta segmentation (0.95 vs. 0.94) and worse in pulmonary segmentation (0.89 vs. 0.92). Note that five-fold cross-validation is applied in (Sedghi Gamechi et al., 2018) and in this paper we apply five random data splits, which may lead to different test data. Unlike in (Sedghi Gamechi et al., 2018), we do not cut the pulmonary artery prediction from the bottom level. In some cases, our method produces segments that extend beyond the manual annotations, which leads to a lower Dice performance. For WMH, the proposed method performs slightly worse than the best performance in the leaderboard using 5 2D UNet ensembles (0.78 vs. 0.81) using the same test data. The top 3 methods in the leaderboard are all 2D UNet ensembles (0.81 vs. 0.80 vs. 0.80), which shows a well-tuned UNet can provide strong baseline performance for WMH segmentation. The best non-ensemble approach is brain atlas guided attention UNet which is more comparable to the proposed method (0.79 vs. 0.78). For ISLES, note that the test sets used in this paper are different from the ones that are used to calculate the leaderboard performance. The performance of the proposed method using 14 training images is quite comparable to the best performance in the leaderboard (0.61 vs. 0.59), which is the only CNN-based method (Kamnitsas et al., 2017) among the top-3 methods in Dice metrics (0.59 vs. 0.55 vs. 0.47).

5.2. Optimization of the End-to-end CRF

We show the evolution of the trainable CRF parameters in one data split of WMH dataset in Figure 5. For the four parameters in the 2×2 compatibility matrix μ and the two diagonal spatial kernel weights ω_2, Spatial-CRF falls into different local optimal values compared to other CRF methods, probably because different parameter scaling due to the lack of the appearance kernel. In contrast, Intensity-CRF and Posterior-CRF converged to similar optimal values for μ and ω_2. For the two diagonal bilateral kernel weights in ω_1 that control the appearance kernel, Intensity-CRF and Posterior-CRF converged to two dif-

Table 3. Results. Mean (standard deviation). The best results are marked in bold. Each experiment is repeated 5 times with different random data split. The last two colomns are lesion-wise metrics. *: significantly better than CNN baseline (p<0.05). °: significantly worse than Posterior-CRF (p<0.05). P-values are calculated by two-sided paired t-test. All CRF methods are implemented in 3D fully-connected manner and share the same CNN architecture and hyperparameters.

Methods	DSC	H95(mm)	AVD(%)	Recall	Recall(lesion)	F1(lesion)
CT Arteries: Aorta						
CNN baseline	0.9291(0.02)°	5.5560(1.96)°	6.8780(4.17)°	0.8993(0.03)°	N/A	N/A
Postproc-CRF	0.9264(0.02)°	5.1591(1.59)°	8.5326(4.81)°	0.8878(0.04)°	N/A	N/A
Intensity-CRF	0.9457(0.01)*°	3.2802(0.77)*°	3.1967(2.58)	0.9548(0.02)*	N/A	N/A
Spatial-CRF	0.9188(0.02)°	7.6562(3.98)°	6.1013(5.13)°	0.8939(0.05)°	N/A	N/A
Posterior-CRF	**0.9538**(0.01)*	**2.8699**(0.86)*	**2.3688**(2.29)*	**0.9555**(0.02)*	N/A	N/A
CT Arteries: Pulmonary Artery						
CNN baseline	0.8510(0.05)°	10.3000(4.87)°	16.7687(12.60)°	0.8867(0.09)	N/A	N/A
Postproc-CRF	0.8561(0.05)	10.0052(5.22)°	13.7071(10.26)°	0.8698(0.09)°	N/A	N/A
Intensity-CRF	0.8773(0.04)*	8.9208(3.09)*	11.8671(8.66)*	**0.9079**(0.06)	N/A	N/A
Spatial-CRF	0.8558(0.06)°	10.5672(5.19)°	13.7399(13.47)	0.8603(0.09)°	N/A	N/A
Posterior-CRF	**0.8935**(0.04)*	**7.6635**(3.92)*	**8.9245**(7.07)*	0.8979(0.07)	N/A	N/A
WMH						
CNN baseline	0.7557(0.13)°	6.5015(9.87)°	28.3351(45.64)°	**0.7977**(0.14)	0.6476(0.14)	0.6648(0.11)°
Postproc-CRF	0.6970(0.17)°	8.8659(7.79)°	35.0786(22.69)°	0.5947(0.20)°	0.3476(0.16)°	0.4831(0.16)°
Intensity-CRF	0.7706(0.10)°	4.9403(4.58)	15.6263(16.44)*	0.7751(0.12)	0.6803(0.15)*	0.6705(0.10)°
Spatial-CRF	0.7602(0.11)°	5.8469(5.82)°	23.5154(25.76)°	0.7831(0.13)	**0.6876**(0.14)*	0.6569(0.11)°
Posterior-CRF	**0.7887**(0.09)*	**4.2972**(3.87)*	**14.8427**(12.66)*	0.7707(0.12)	0.6670(0.14)	**0.6952**(0.10)*
ISLES						
CNN baseline	0.5795(0.28)	27.6725(25.58)	72.3048(121.12)	0.6590(0.31)	0.7586(0.33)	0.4941(0.35)
Postproc-CRF	0.5621(0.31)	**19.5302**(20.72)	59.1030(85.99)	0.6132(0.34)	0.6518(0.39)	0.5545(0.36)
Intensity-CRF	0.5758(0.26)	46.6002(32.17)°	65.9278(68.98)	0.6397(0.30)	0.7350(0.33)	0.4094(0.31)°
Spatial-CRF	0.5898(0.26)	31.1519(29.50)	93.1006(171.83)	**0.6794**(0.28)	**0.7848**(0.31)	0.4945(0.34)
Posterior-CRF	**0.6075**(0.24)	25.1834(23.27)	**47.5171**(38.34)	0.6501(0.29)	0.7443(0.31)	**0.5625**(0.32)

ferent optimal values. This suggests that different CRF feature spaces contribute mostly through the appearance kernel and less through the compatibility matrix or the spatial kernel. Interestingly, for the second diagonal bilateral weight $\omega_1^{(2)}$, there is a different trend of Posterior-CRF compared to Intensity-CRF, which may indicate that at the early training stage Posterior-CRF uses similar feature space like that in Intensity-CRF, but at the later stage it finds and learns another set of features that may help categorize the lesion class better, which are more reliable than the original intensity features.

5.3. Influence of CRF Hyperparameters

We conduct experiments to investigate the influence of CRF hyperparameters on both end-to-end CRF with predefined features and the proposed CRF with learned features.

Trainable CRF parameters. The CRF weights μ, ω_1, and ω_2 in the end-to-end CRF learning can be automatically updated together with CNN weights. We run Intensity-CRF and Posterior-CRF using WMH datasets with five different initializations of CRF weights randomly sampled from the search scale with all other parameters the same as in Table 2. The CNN initializations are the same for all experiments. The results in Table 4 show that Intensity-CRF and Posterior-CRF converge to similar optimal points across different initializations. Spatial-CRF shows higher variances across experiments and is less stable to the change of initializations. Posterior-CRF is more robust to changes in initialization, achieving higher average performance and smaller standard deviations compared to Intensity-CRF and Spatial-CRF.

Table 4. Performance (Dice score) across 5 different initializations of CRF weights on WMH dataset.

Methods	Intensity-CRF	Spatial-CRF	Posterior-CRF
Mean (std)	0.7570 (0.008)	0.7507 (0.02)	**0.7833** (**0.003**)

Empirically tuned parameters. The CRF standard deviation parameters θ_α and θ_γ, controlling the spatial terms, and θ_β controlling the appearance term, were tuned empirically to give the best results for post-processing CRF. We here test, for WMH segmentation, five different values of θ_α, θ_β, and θ_γ for Intensity-CRF and Posterior-CRF and five different values of θ_γ for Spatial-CRF within the search scale. All other parameters are the same as in Table 2. The results are shown in Figure 7. We can see that Posterior-CRF is more robust to θ_α and θ_β and has consistently better performance than Intensity-CRF within the search scale, suggesting that Posterior-CRF parameters are more easy to tune. All CRF methods degenerate performance when θ_γ becomes larger and show the best performance when using a similar value as that in the grid search for post-processing CRF. Spatial-CRF is more robust to θ_γ compared to other CRF methods and has similar performance as CNN baseline with larger θ_γ. This indicates that large θ_γ reduces the CRF effect and the spatial term may introduce more incorrect segmentation when there is also an appearance term in the end-to-

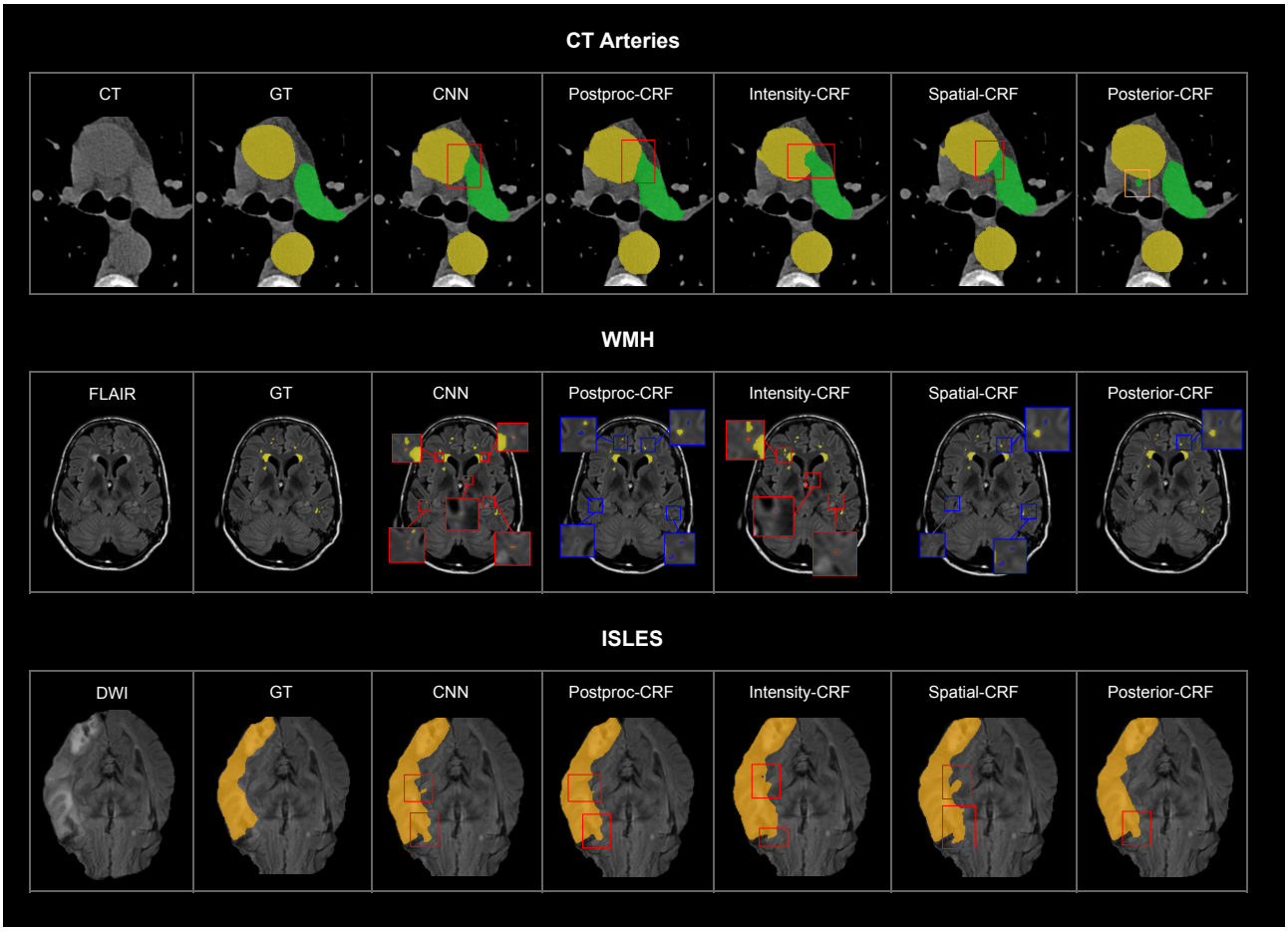

Fig. 6. Example segmentation results. ***From left for each row*****: (1) Original image (2) Manual annotation (3) CNN baseline (4) Postproc-CRF (5) Intensity-CRF (6) Spatial-CRF (7) Posterior-CRF. Aorta is colored with yellow and the pulmonary artery is green, white matter hyperintensities and ischemic stroke lesions in yellow. Red/blue rectangles indicate areas with over/under segmented voxels and the orange rectangle indicates another branch of pulmonary artery whose annotation starts in the next few slices and merged with the main branch gradually. In the WMH example (second row), only detections that do not overlap with any ground truth voxel (false positive lesions) or ground truth lesions for which no voxel is detected (false negative lesions) are highlighted, and in the zoomed patches red and blue voxels indicate false positive and false negative lesions respectively. Better viewed in color with zoom.**

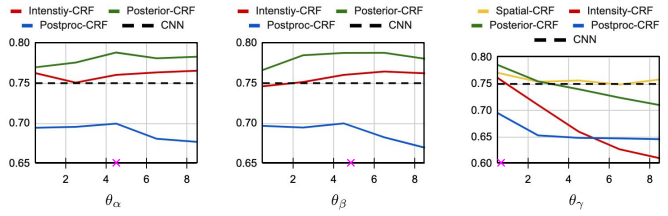

Fig. 7. Dice performance of varying θ for CRF methods on WMH dataset. CNN result is shown as the black dash line. Purple crosses indicate the values used in Table 4. Best viewed in color with zoom.

end CRF like Intenity-CRF and Posterior-CRF.

5.4. Influence of Hierarchical CNN Features as CRF Reference Maps

We conduct experiments to investigate which level of features – earlier or deeper in the network – are more useful for the feature-learning-based CRF. We implement nine variants of feature-learning-based CRF with different levels of CNN feature maps as reference maps in the same 3D UNet architecture. For example, the method *FL-CRF-e-1* indicates the feature-learning-based CRF using the level 1 feature maps in the UNet encoder path as CRF reference maps. The implementation detail of FL-CRF-e-1 is shown in Figure 3. To reduce the computational cost and keep the same layer capacity as Posterior-CRF, the 32-channel (or more in deeper layers) feature maps are encoded into C-channel feature maps and go through a softmax layer as the CRF reference maps. Since there is no gradient flowing back through the reference map path, we optimize the softmax layer with the segmentation loss directly in order to preserve as much semantic information as possible. Note that for CRF methods that use deeper CNN layers as reference maps, such as FL-CRF-e-2 to FL-CRF-d-2, we upsample the reference maps to the original image scale using nearest neighbor interpolation and optimize them with the segmentation loss, similar to FL-CRF-e-1.

The results are shown in Figure 8. Note that if we use the CNN input as CRF reference maps, it turns into Intensity-CRF; if we use the last CNN layer as CRF reference maps, it turns into Posterior-CRF. In the figure, we can see that all feature-learning-based CRF approaches (including Posterior-CRF) out-

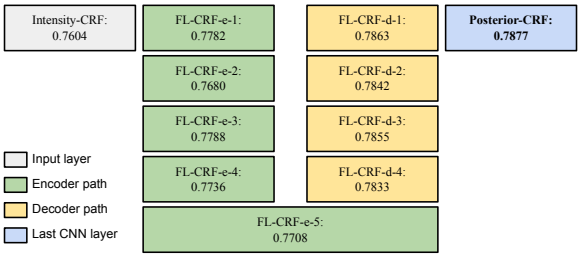

Fig. 8. Dice performance of end-to-end CRFs using different CNN feature maps in an independent run on WMH dataset. Different blocks indicate different level of CNN feature maps used as CRF reference maps. Best viewed in color with zoom.

perform Intensity-CRF and the overall Dice performance in the decoder path is better than that in the encoder path, indicating that CNN learned features are more useful to the CRF inference than intensity is and later CNN features are more useful than early features. The performance degenerates towards the middle part of the UNet (from FL-CRF-e-1 to FL-CRF-e-5 and FL-CRF-d-1 to FL-CRF-d-4) but fluctuates at the 2nd/3rd level. We argue that this may be due to the pooling effect which enables CNN to extract higher-level features but loses the spatial information at the same time. Posterior-CRF achieves the best performance among all variants and we argue that this is because the last CNN layer are more likely to contain more useful information for CRF inference and it still keeps the same spatial scale as the original image.

5.5. Evolution of CNN and CRF Outputs

The concurrent optimization of CNN and CRF in our end-to-end models allows the CNN and CRF to interact during training. We observed that this has a strong effect on what the CNN learns in the early training epochs. Figure 9 shows the evolution of CNN and CRF outputs for three typical examples. The baseline CNN without CRF converges quickly and focuses on the large lesions, already producing a fairly sparse output after the first epoch. The end-to-end models converge more slowly, and in this case the output of the CNN is influenced by the choice of CRF mostly in the early stage of training. For example, the CNN in the Intensity-CRF model initially tends to highlight voxels with similar intensity as the foreground (1 to 20 epoch), while the CNN in the Spatial-CRF model preserves the spatial coherence between voxels and outputs many small groups of voxels (5 epoch). The CNN in the Posterior-CRF model first focuses on the coarse area that might contain the target lesions (1 to 5 epoch) and then refine the prediction gradually to the ground truth (5 to 20 epoch). Eventually, all models converge to a result close to the ground truth.

6. Discussion

In this paper, we explored efficient methods to combine the global inference capabilities of a CRF with the feature extraction from a CNN. Our end-to-end approach optimizes the CRF and CNN at the same time, and allows the two components of the approach to cooperate in learning effective feature representations. This gives our method an advantage over traditional CRFs that only use the original image intensities and position information. Intensity-based features can be suboptimal for problems where the intensity does not provide sufficient information to find the object boundaries, for example because the contrast between objects is too small.

Unlike other CRF methods, our Posterior-CRF uses adaptive learning-based features that are learned by the CNN and can combine spatial and appearance information in a way that suits the CRF. The results show our method can achieve stable, good performance across a range of segmentation applications and imaging modalities. FL-CRF variants that use early CNN features in Section 5.4 achieve in-between performance between Intensity-CRF and Posterior-CRF, using learning-based features that range from more similar to intensity to more similar to posterior probability maps. Finally, we found that integrating learned features into the CRF model reduces the need to fine-tune CRF parameters, making the method easier to apply than CRF methods with predefined features.

6.1. Interaction between CRF and CNN

Figure 9 leads to the counter-intuitive observation that, at least initially, the CNNs in end-to-end models seem to imitate the CRF instead of complementing it. For example, the CNN output in Intensity-CRF highlights the ground truth, but also finds areas with similar intensities, producing something that looks very similar to the original image (20 epoch). The CNN output in Spatial-CRF selects the ground truth but also includes clusters of voxels in other areas (5 epoch).

This effect can be explained by the way the CNN and CRF interact during training. In Intensity-CRF and Spatial-CRF, the only interaction between CRF and CNN takes place through the unary map (Figure 4, step 5, green arrow). For example, consider how this works in the Intensity-CRF. In WMH segmentation, the ground truth is usually high-intensity area. However, for the voxels with high intensities but not the target lesions, it is difficult to get both low pairwise CRF potentials and low segmentation loss, since labeling them as non-lesion goes against the CRF assumption that voxels with similar high-intensities are more likely to be the lesion class. For convenience, we call these voxels as *hard voxels*, indicating the voxels that do not fit the CRF assumption. In order to keep the correctly segmented lesions and reduce the CRF effect on the hard voxels at the same time, the CNN tends to provide unary maps that 1) highlight the ground truth area for lower segmentation loss, and 2) look similar to the CRF reference maps on the hard voxels for lower pairwise CRF potentials. In the later stage of training, CNN is encouraged to push the confidence of its outputs even further to minimize unary potentials and thus prevent CRF from undoing segmentation improvement on the hard voxels. From Figure 9, we can see that there are many hard voxels in Intensity-CRF (1 to 20 epoch, areas that look like the original image) and Spatial-CRF (5 epoch, clusters of voxels that do not belong to the ground truth) which may harm the segmentation. This indicates that the predefined features may not be the optimal feature space for the end-to-end CRF.

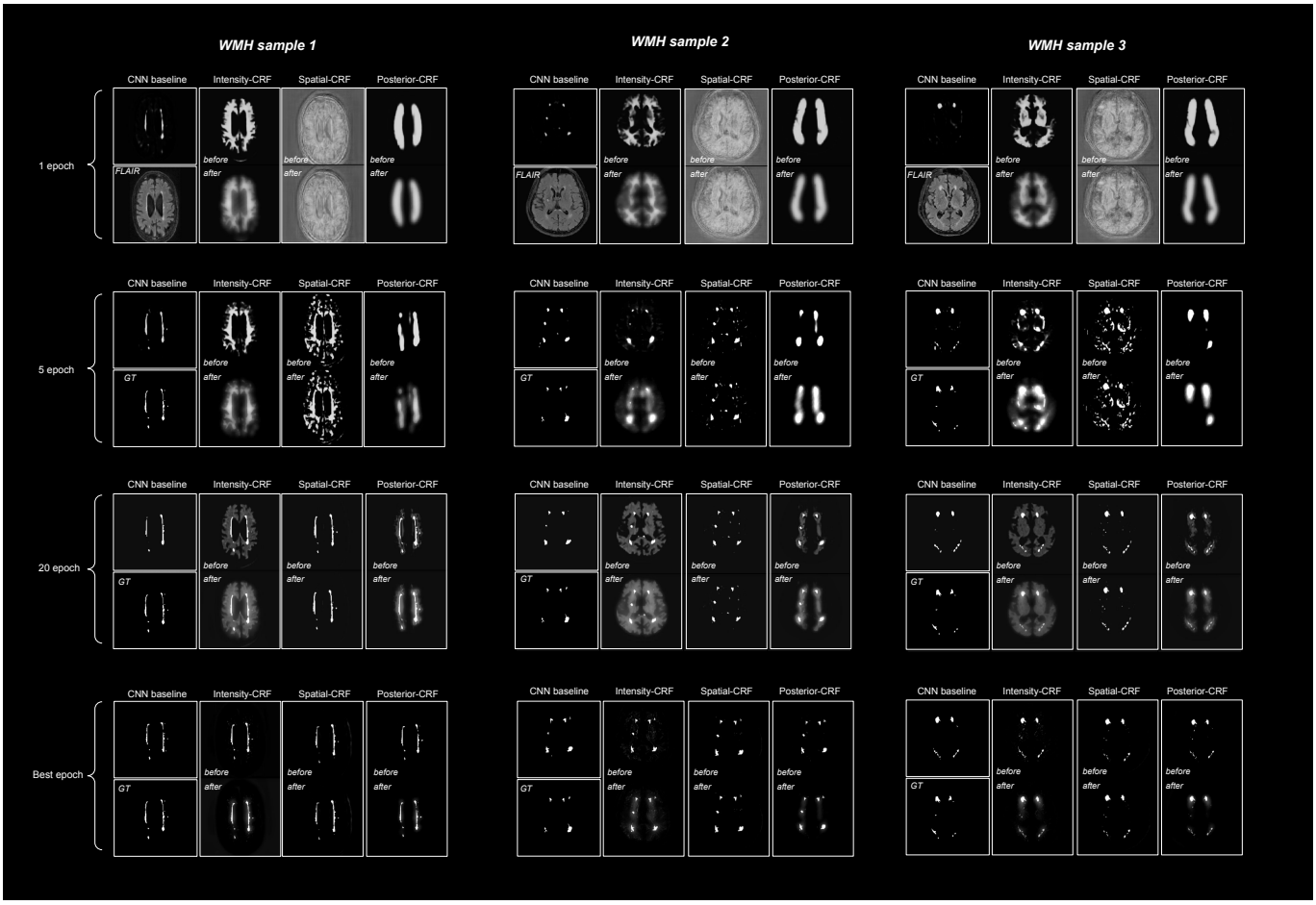

Fig. 9. Evolution of CNN and CRF outputs during training. The CNN output maps and CRF results for WMH segmentation in 3 different MRI images (columns) are shown at, from top row to bottom row, epoch 1, 5, 20, and the best epoch. The best epoch is chosen when the model shows the best validation performance till the end of training (usually at 50~80 epoch). *FLAIR:* the input FLAIR image of the current training sample. *GT:* ground truth. *CNN baseline:* the last layer (softmax output) of CNN. *Intensity-CRF, Spatial-CRF, Posterior-CRF:* the probability maps before/after the CRF layer at different epochs during training. Best viewed with zoom.

In the Posterior-CRF model, the CRF inference happens within the CNN feature space, which can improve the interaction between CNN and CRF. First, the features learned by CNN during training may contain information that is more useful for segmentation than that in the predefined features, which makes CRF benefit most from the CNN features. Second, using the learning-based features as CRF reference maps avoids the CRF assumption of the predefined features which may introduce many hard voxels, e.g., Intensity-CRF and Spatial-CRF, as discussed in the previous paragraph. With fewer hard voxels, the CNN in Posterior-CRF may provide better unary maps for the CRF inference.

6.2. Posterior-CRF vs. Mean-field Network

The mean-field approximation (MFA) in Posterior-CRF is somewhat similar to that in Mean-field networks (MFN) (Li and Zemel, 2014), since both methods use it to get the posterior probabilities of the variables. Therefore, MFN could be a promising alternative to the MFA process in our method. MFN has the advantage that it utilizes each layer of the network as an iteration of MFA, which has the advantage of allowing more relaxation on parameters and provides some efficiency improvements. This makes the idea of formulating Posterior-CRF as a feed-forward network like MFN very attractive. There are, however, a few limitations that would need to be solved.

The first limitation is in training. MFN is designed to provide a faster and more flexible way to obtain the prediction of MFA, by fitting a powerful function that predicts the real MFA result. To train an MFN, we first need to acquire the ground truth calculated by conventional mean-field iterations, which takes time during training but saves time during inference. On the other hand, Posterior-CRF provides a flexible and adaptive feature space for the conventional MFA, speeding up the procedure by applying Gaussian convolution in the message passing updates. As a result, the thing Posterior-CRF does is difficult to replicate with a MFN because the feature space of a Posterior-CRF changes during training, while MFN requires a predefined feature space to get the ground truth.

The second limitation is the tradeoff between dense inference and computation cost in the MFN. In its feed-forward network implementation, the computation cost increases exponentially when more neighbor nodes and number of layers are included, which limits its ability to model dense prediction problems such as segmentation tasks.

6.3. Posterior-CRF vs. Graph Neural Networks

The proposed Posterior-CRF shares some similarities with graph neural networks (GNN) (Scarselli et al., 2008; Selvan et al., 2018): both approaches aim to model interactions between variables within a graph model. The difference is that Posterior-CRF pre-defines the global relationship between variables through the mean-field assumptions and solves the maximum a posteriori problem, whereas GNN learns the global variable relationship by applying graph convolution filters and mapping the input graph to the output graph (Selvan et al., 2018).

It could be interesting to combine the global view of the Posterior-CRF and the more local view of the GNN. The Posterior-CRF might benefit from using a GNN to replace its CNN component for feature extraction. The graph-based network may extract better features for Posterior-CRF than a CNN, which is not designed to extract unary and pairwise features for a graphical model. Similarly, the GNN may benefit from the efficient message passing of the Posterior-CRF, which would allow it to use the local graph-based features as CRF features for global interactive modeling in a computationally efficient way.

6.4. Limitations

In this paper, we show that the proposed Posterior-CRF method has benefits in the three medical imaging applications. Considering the medical imaging datasets are usually small largely because the manual annotations are very expensive to make, difference between Posterior-CRF and UNet may be smaller in larger training sets. But we know from literature that Intensity-CRF helps in some computer vision applications with large training sets (e.g., 10k 2D images or even more), it would be promising to test our method on these datasets. This is considered as our future work.

In Section 5.3, we show that Posterior-CRF is robust to different CRF initializations and hyperparameters. However, the standard deviation parameters still require careful tuning, especially for θ_γ in the spatial term. θ_γ is sensitive to the image scale of different datasets and the size of the target object in different applications. Nevertheless, we recommend the researchers to use the default (or optimal if it is available) setting of post-processing CRF as a reference for tuning Posterior-CRF rather than random initialization. Posterior-CRF is more robust to θ_α and θ_β compared to Intensity-CRF, which facilitates exhaustive tuning of these parameters.

The computational expense of the CRF also restricts the choice of applications. Compared to UNet (∼5 mins for 1 epoch in WMH experiment), there is around 20% training time increased on average when applying a CRF layer on top of the network (∼6 mins for 1 epoch). All end-to-end CRFs share similar computational costs. Given that Posterior-CRF uses posterior probability maps as its reference maps, it can become computationally expensive in multi-class segmentation problems. For a similar reason, Intensity-CRF and Postproc-CRF can become expensive when there are too many imaging modalities in the input channels M.

In the experiments, we use a plain 3D UNet as the backbone network for all methods. The training pipeline and hyperparameters are determined empirically and kept the same for all datasets, which could be suboptimal compared to elaborate automatic configuration strategies like nnU-Net (Isensee et al., 2020). On the WMH dataset we therefore checked the performance of nnU-Net (3D version without ensembling). Average Dice score of nnU-net (0.77) was slightly higher than our CNN baseline (0.76, difference not statistically significant) but lower than the proposed posterior CRF using the CNN baseline as a backbone (0.79), which performed significantly better than the CNN baseline (see Table 3). Though our experiments have been limited to a standard 3D U-net architecture, We expect that posterior CRF can improve results of other segmentation architectures and other hyperparameter settings (such as nnU-net) as well.

7. Conclusions

In conclusion, we present a novel end-to-end segmentation method called Posterior-CRF that uses learning-based, class-informative CNN features for CRF inference. The proposed method is evaluated in three medical image segmentation tasks, including different MRI/CT imaging modalities and covering a range of object sizes, appearances and anatomical classes. In the quantitative evaluation, our method outperforms end-to-end CRF with early CNN features, end-to-end CRF approaches with predefined features, post-processing CRF, as well as a baseline CNN with similar architecture. In two of the three applications, our method significantly improves the segmentation performance. The qualitative comparison demonstrates that our method has good performance on segmenting blurred boundaries and very small objects.

Acknowledgments

The authors would like to thank Raghavendra Selvan, Gerda Bortsova for their constructive suggestions for the paper, Dr. Zaigham Saghir from DLCST for providing us with the chest CT scans, and organizers of WMH 2017 and ISLES 2015 Challenges for providing the public datasets. This work was partially funded by Chinese Scholarship Council (File No.201706170040), Iranian Ministry of Science, Research and Technology, and The Netherlands Organisation for Scientific Research (NWO).

References

Adams, A., Baek, J., Davis, M.A., 2010. Fast high-dimensional filtering using the permutohedral lattice, in: Computer Graphics Forum, Wiley Online Library. pp. 753–762.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 .

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40, 834–848.

Chen, S., de Bruijne, M., 2018. An end-to-end approach to semantic segmentation with 3d cnn and posterior-crf in medical images. medical imaging meets neurips workshop, NeurIPS 2018 .

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM. pp. 785–794.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 424–432.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.A., 2017. 3d deeply supervised network for automated segmentation of volumetric medical images. Medical image analysis 41, 40–54.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2020. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods , 1–9.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical image analysis 36, 61–78.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

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.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Kuijf, H.J., Biesbroek, J.M., de Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., et al., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities; results of the wmh segmentation challenge. IEEE transactions on medical imaging .

Li, Y., Ping, W., 2018. Cancer metastasis detection with neural conditional random field. arXiv preprint arXiv:1806.07064 .

Li, Y., Zemel, R., 2014. Mean-field networks. arXiv preprint arXiv:1410.5884 .

Lin, G., Shen, C., Van Den Hengel, A., Reid, I., 2016. Efficient piecewise training of deep structured models for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3194–3203.

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.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, pp. 4765–4774.

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. Medical image analysis 35, 250–269.

Maier, O., Wilms, M., von der Gablentz, J., Krämer, U.M., Münte, T.F., Handels, H., 2015. Extra tree forests for sub-acute ischemic stroke lesion segmentation in mr sequences. Journal of neuroscience methods 240, 89–100.

Monteiro, M., Figueiredo, M.A., Oliveira, A.L., 2018. Conditional random fields as recurrent neural networks for 3d medical imaging segmentation. arXiv preprint arXiv:1807.07464 .

Pedersen, J.H., Ashraf, H., Dirksen, A., Bach, K., Hansen, H., Toennesen, P., Thorsen, H., Brodersen, J., Skov, B.G., Døssing, M., et al., 2009. The danish randomized lung cancer ct screening trialoverall design and results of the prevalence round. Journal of Thoracic Oncology 4, 608–614.

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.

Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G., 2008. The graph neural network model. IEEE Transactions on Neural Networks 20, 61–80.

Schwing, A.G., Urtasun, R., 2015. Fully connected deep structured networks. arXiv preprint arXiv:1503.02351 .

Sedghi Gamechi, Z., Arias-Lorza, A.M., Pedersen, J.H., De Bruijne, M., 2018. Aorta and pulmonary artery segmentation using optimal surface graph cuts in non-contrast ct, in: Medical Imaging 2018: Image Processing, International Society for Optics and Photonics. p. 105742D.

Selvan, R., Welling, M., Pedersen, J.H., Petersen, J., de Bruijne, M., 2018. Mean field network based graph refinement with application to airway tree extraction, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 750–758.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

Vemulapalli, R., Tuzel, O., Liu, M.Y., Chellapa, R., 2016. Gaussian conditional random field network for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3224–3233.

Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al., 2018a. Deepigeos: a deep interactive geodesic framework for medical image segmentation. IEEE transactions on pattern analysis and machine intelligence 41, 1559–1572.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794–7803.

Wang, Z., Zou, N., Shen, D., Ji, S., 2020. Non-local u-nets for biomedical image segmentation., in: AAAI, pp. 6315–6322.

Xie, Y., Padgett, J., Biancardi, A.M., Reeves, A.P., 2014. Automated aorta segmentation in low-dose chest ct images. International journal of computer assisted radiology and surgery 9, 211–219.

Yu, L., Yang, X., Qin, J., Heng, P.A., 2016. 3d fractalnet: dense volumetric segmentation for cardiovascular mri volumes, in: Reconstruction, segmentation, and analysis of medical images. Springer, pp. 103–110.

Yuan, H., Zou, N., Zhang, S., Peng, H., Ji, S., 2019. Learning hierarchical and shared features for improving 3d neuron reconstruction, in: 2019 IEEE International Conference on Data Mining (ICDM), IEEE. pp. 806–815.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H., 2015. Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE international conference on computer vision, pp. 1529–1537.

Supplementary Material

8. Mean-field Inference

Mean-field inference is an efficient approximation to computing distribution $Q(\mathbf{X})$ instead of the real CRF distribution $P(\mathbf{X})$, which could be done in an iterative algorithm 1 (see also Figure 10). X is the random field w.r.t the current 3D image patch $\mathbf{I}$.

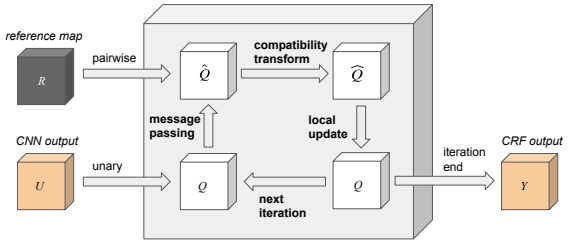

Fig. 10. Mean-field approximation in the end-to-end CRF layer. There are two inputs of the CRF layer, where U is the CNN probability maps as the unary maps and the pairwise distribution are calculated by the initialized distribution Q and the reference map I. The updated distribution Y is the output of the layer at the end of the iteration. Best viewed in color with zoom.

There are three main steps inside the inference iteration. First is message passing, which is the most calculation-intense step that could be expressed as a convolution operation on all the pairwise kernels k and the initialized $Q(\mathbf{X})$. An efficient way

Algorithm 1 Mean-field inference in fully-connected CRF

$Q_i(x_i) \leftarrow U_i(x_i), i = 1, 2, ..., N$ ▷ **Initialize** $Q(\mathbf{X})$
while not reach max iteration number **do**
  $\hat{Q}_i^{(m)}(x_i) \leftarrow \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(x_i)$ for all m ▷ **Message Passing**
  $\widehat{Q}_i^{(m)}(x_i) \leftarrow \sum_{l \in \mathcal{L}} \mu^{(m)}(x_i, l) \sum_m \omega^{(m)} \hat{Q}_i^{(m)}(l)$ ▷ **Compatibility Transform**
  $Q_i(x_i) \leftarrow \exp\{-\varphi_u(x_i) - \widehat{Q}_i(x_i)\}$ ▷ **Local Update**
  normalize $Q_i(x_i)$
end while

to perform high-dimensional convolution is using permutohedral lattice algorithm (Adams et al., 2010). In compatibility transform as the second step, all the convolution results $\hat{Q}_i^{(m)}(x_i)$ are weighted by $\omega^{(m)}$ in different sort of kernels and shared between labels to a varied extent, depending on the compatibility μ between these labels. At last, $Q(\mathbf{X})$ will be updated by the calculated pairwise potential and used as the input for the next iteration.

9. SHAP Analysis of Post-processing CRF

We conduct SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) analysis on the post-processing CRF grid search results to investigate the contribution of each individual CRF parameter to the segmentation performance. With this analysis, we show that it is difficult to tune traditional CRF parameters to achieve a consistent performance improvement on different applications, and our proposed method does not require tuning parameters. Moreover, the analysis shows the importance of each modality to each dataset, which can be automatically adapted in the proposed method but not in traditional methods. The model is trained using *XGBoost* (Chen and Guestrin, 2016) for 100 iterations using a learning rate of 0.5, 0.01, and 0.01 for CT Arteries, WMH, and ISLES respectively. Note that the SHAP analysis results can only be explained under the assumption of the current parameter search scales and XGBoost models.

The results are shown in Figure 11. The summary plot in the left sub-graph shows an overview of all parameter sets with the most important parameters on top of the list. For each dataset, the best and worst parameter settings are shown in the right subgraph. For all datasets, the post-processing quality is affected most by the spatial parameters ω_2 and θ_γ, and less by the intensity parameters per modality θ_β.

The results on the CT arteries data (Figure 11a *left*) are more stable (with smaller SHAP values) than the results for WMH and ISLES, indicating that the post-processing CRF can hardly change the CNN output of the artery segmentation (see Figure 6 in the paper as an example).

In the WMH dataset, looking at independent parameter contributions, low values for spatial parameters ω_2, θ_γ (less smoothing), and a smaller number of iterations lead to an improved performance. This is not unexpected, because white matter lesions are sparsely distributed and spatial smoothing tends to remove small lesions. Too strong spatial correlations (either large weight ω_2 or small θ_γ) will remove true positives as well (see Figure 6 in the paper). The summary plot (Figure 11b *left*) shows, as expected, that the FLAIR image has a larger impact on the model than the T1 image. Table 1 also shows a smaller θ_β selected (corresponding to higher influence) for FLAIR.

Similar trends can be found for the ISLES dataset (Figure 11c). Spatial parameters ω_2 and θ_γ are important to tune and high values can strongly harm the performance. The summary plot shows that the DWI image has a larger impact on the model than T1, T2, and FLAIR. In Table 1, θ_β for FLAIR and DWI are smaller than θ_β for T1 and T2, which means that FLAIR and DWI images are more informative for the segmentation of ischemic stroke lesions.

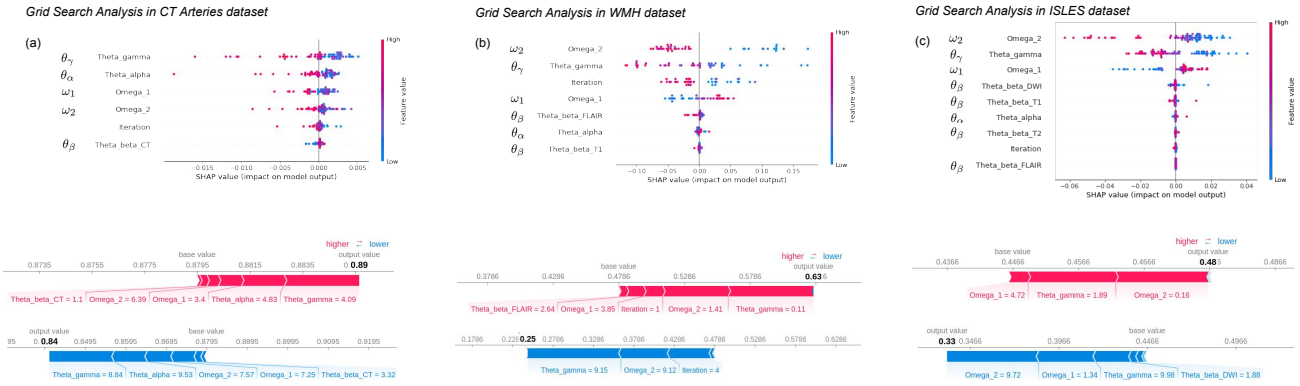

Fig. 11. SHAP analysis of the grid search results. See Section 9 for an explanation. *Upper sub-graphs*: summary plots of all parameter sets evaluated during grid search. Positive SHAP values indicates a positive contribution to the performance and vice versa. The legend (feature value bar) shows the search range for each parameter. This reveals for example that lower values of ω_2 lead to better segmentation performance for all datasets. *Lower sub-graphs*: the best (1st row) and worst (2nd row) parameter sets for each dataset. Red bar represents positive contribution to the performance and blue bar is negative contribution. Base value is the average DSC of all grid search results and output value is the DSC in the parameter set depicted. Best viewed in color with zoom.