Learning Topological Interactions for Multi-Class Medical Image Segmentation

Saumya Gupta*, Xiaoling Hu*, James Kaan, Michael Jin, Mutshipay Mpoy, Katherine Chung, Gagandeep Singh, Mary Saltz, Tahsin Kurc, Joel Saltz, Apostolos Tassiopoulos, Prateek Prasanna, and Chao Chen

Stony Brook University, Stony Brook, New York, USA
{saumya.gupta, xiaoling.hu, chao.chen.1}@stonybrook.edu

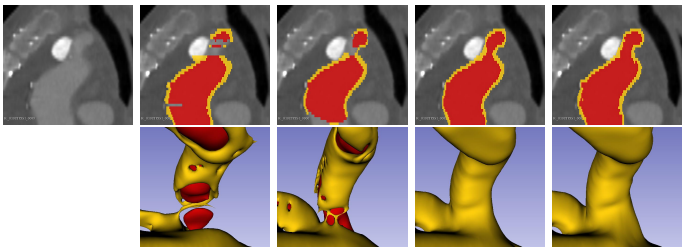
Abstract. Deep learning methods have achieved impressive performance for multi-class medical image segmentation. However, they are limited in their ability to encode topological interactions among different classes (e.g., containment and exclusion). These constraints naturally arise in biomedical images and can be crucial in improving segmentation quality. In this paper, we introduce a novel *topological interaction module* to encode the topological interactions into a deep neural network. The implementation is completely convolution-based and thus can be very efficient. This empowers us to incorporate the constraints into end-to-end training and enrich the feature representation of neural networks. The efficacy of the proposed method is validated on different types of interactions. We also demonstrate the generalizability of the method on both proprietary and public challenge datasets, in both 2D and 3D settings, as well as across different modalities such as CT and Ultrasound. Code is available at: https://github.com/TopoXLab/TopoInteraction

Keywords: Medical Imaging · Segmentation · Topological Interaction

1 Introduction

Instead of using hand-crafted features, state-of-the-art deep segmentation methods [4,5,6,16,29] learn powerful feature representations automatically and achieve satisfactory performances. However, standard deep neural networks cannot learn global structural constraints regarding semantic labels, which can often be critical in biomedical domains. While existing works mostly focus on encoding the topology of a single label [18,19,8,35], limited progress has been made addressing the constraints regarding interactions between different labels. Even strong methods (e.g., nnUNet [21]) may fail to preserve the constraints as they only optimize per-pixel accuracy. For example, in the segmentation of abdominal aorta, we know a priori that the aorta wall always encloses the lumen. Exploiting this constraint can help us segment the wall correctly, providing accurate

* Equal contribution.



(a) Input (b) UNet (c) nnUNet (d) Ours (e) GT

Fig. 1: Motivating examples for aorta segmentation. Red and yellow represent aortic lumen and wall, respectively. Anatomically, the lumen is always enclosed by the wall, separated from the background (illustrated in (e) ground truth GT). Even strong baselines, e.g., (b) *UNet* and (c) *nnUNet*, fail to respect this anatomically important topological constraint because often the intensity of the wall in the input is similar to that of the background. Our proposed method explicitly encodes the constraint, thereby improving the segmentation quality.

geometric measures (e.g., wall thickness and aorta volume) for the prediction of aortic aneurysm eruption risk [11]. See Fig. 1 for an illustration. Another kind of global constraint is mutual exclusion of different labels. For example, in multi-organ segmentation, ensuring different organs to not touch each other can help improve the segmentation quality.

In this paper, we investigate how to help deep neural networks learn these global structural constraints, which we call *topological interactions*, between different semantic labels. To encode such interaction constraints into convolutional neural networks is challenging; it is hard to directly encode hard constraints into kernels while keeping them learnable. Traditional methods [10,38,31,26,22,3] solve the segmentation problem as a combinatorial optimization problem (e.g., graph-cut or multicut) and encode these topological interactions as constraints of the solution. However, these approaches do not apply to deep neural networks, which do not rely on a global optimization for the inference. Even if one can encode the constrained optimization as a post-processing step, it will be very inefficient. More importantly, the optimization is not differentiable and thus cannot be incorporated into training.

We propose a novel method to learn the topological interactions for multi-class segmentation tasks. A desirable solution should be efficient. Furthermore, it should be incorporated into training to help the network learn. Our key observation is that a broad class of topological interactions, namely, enclosing and exclusion, boils down to certain impermissible label combinations of adjacent pixels/voxels. Inspired by such observation, we propose a *topological interaction module* that encodes the constraints into a neural network through a series of convolutional operations. Instead of directly encoding the constraint into the convolutional kernels, the proposed module directly identifies locations where the constraints are violated. Our module is extremely efficient due to the

convolution-based design. Furthermore, it can naturally be incorporated into the training of neural networks, e.g., through an extra loss penalizing the constraint-violating pixels/voxels. As shown in Fig. 1, incorporated with our module, the network can learn to segment aortic walls correctly even when strong baselines, such as nnUNet, fail.

We evaluate the proposed method by performing experiments on both proprietary and public challenge datasets, in both 2D and 3D settings, and across different modalities. The results show that our method is generalizable and can be employed in various scenarios where topological interactions apply. It not only enforces the constraints, but also improves the segmentation quality significantly in standard metrics such as DICE, Hausdorff distance, etc. This is as expected; a network that encodes the constraints also learns a better representation for segmentation. In summary, our contributions are as follows:

- We propose an efficient convolution-based module to encode the topological interactions in a multi-class segmentation setting.
- The proposed module is very efficient and generic. It can be incorporated into any backbone to encode the constraints in an end-to-end training pipeline.
- Through extensive experiments on multiple medical imaging datasets, we show our method effectively improves the segmentation quality without increasing computational cost.

2 Related Work

Multi-Class Image Segmentation. Numerous graph or energy based methods have been proposed to deal with multi-class image segmentation in the pre-deep learning era. Some of these methods integrate fuzzy spatial relations [9] or encode spatial interactions via inter-object distances [28]. Others encode spatial relationships for hierarchical segmentation [12,36]. For example, Strekalovskiy et al. [36] enforce geometric constraints by introducing a label ordering constraint. Li et al. [27] propose to segment nested objects with graph-based approaches. Delong et al. [10] propose to encode geometric constraints between different regions into a graph cut framework for multi-class image segmentation.

Geometric and Topological Constraints. Early works, using classic frameworks such as level set or Markov random field, enforce topological or geometric constraints while solving the energy minimization problem [14,25,3,10,38,31,26,22]. However, these methods cannot be easily incorporated into the training of deep neural networks. In recent years, new methods have been proposed to incorporate geometric/topological constraints into the training of deep neural networks (DNNs) [18,19,8,35,39]. These methods enable the DNNs to learn geometry-/topology-aware representations and to deliver better segmentation results. However, all these methods are focusing on the topology, e.g., connections, loops and branches, of a single foreground class. They cannot enforce topological interactions between different classes. For example, in aorta segmentation, forcing the aortic wall to be a tube in 3D cannot guarantee that the wall contains the lumen

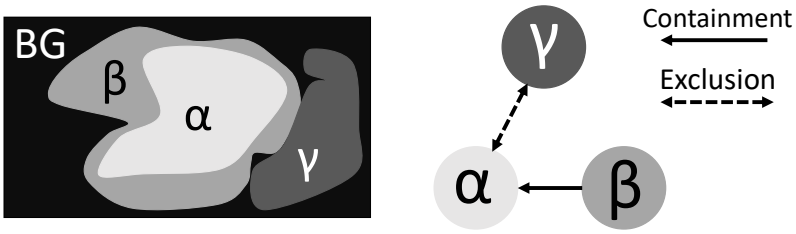


Fig. 2: Schematic illustration of the topological interactions: containment and exclusion. BG denotes the background class. **Containment**: β contains α. **Exclusion**: α and γ are mutually exclusive.

and separates it from the background. This gap motivates our investigation on encoding the inter-class topological interactions in DNN training.

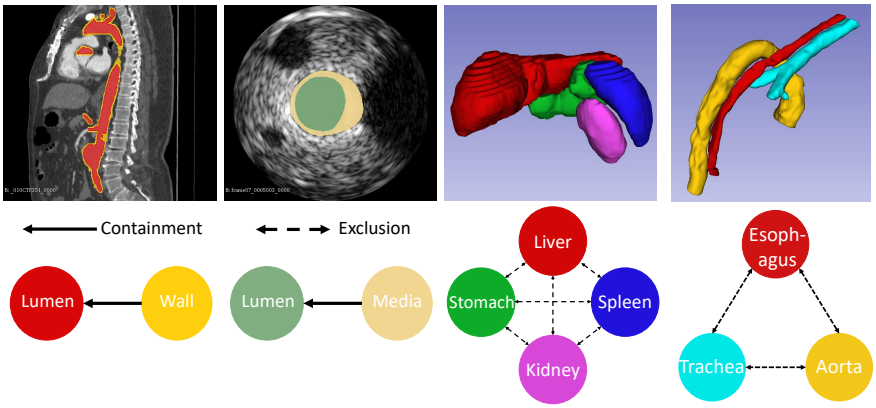
The method closest to ours is [2], which we refer to as TopoCRF. It encodes the mutual exclusion constraint as a constraint on the posterior probability (softmax layer output) at each pixel/voxel, without taking neighborhoods into account. Therefore, this approach cannot really exclude the case when adjacent pixels have a forbidden label combination. The explicit construction of 2^c constraint-encoding priors for a c-class problem is also very expensive and does not scale. Additional methods similar to TopoCRF are [32] which we refer to as MIDL, and [13] which we refer to as NonAdj. MIDL is a direct application of TopoCRF by simply adding a DICE loss term. NonAdj extends TopoCRF by taking the adjacent pixels into consideration, however, it requires a strong pre-trained model to perform well. Both MIDL and NonAdj focus on modeling joint distributions, and thus suffer from similar issues as TopoCRF.

3 Methodology

Broadly speaking, topological interactions between different foreground classes include two types, containment and exclusion. In Fig. 2, we illustrate these constraints using three class labels, α, β and γ.

- **Containment**: Class β contains α if β completely surrounds α. We use solid arrow from β to α to denote the containment relationship. In real applications, e.g., aorta segmentation, the aortic wall contains the lumen. See Fig. 3(a) for an illustration.
- **Exclusion**: Classes α and γ are mutually exclusive if the pixels/voxels of class α and class γ cannot be adjacent to each other. We use dashed double-arrow to denote the exclusion relationship. In multi-organ segmentation, there is clear separation between stomach and liver. They are mutually exclusive. See Fig. 3(c) for an illustration.

These constraints are quite general and can be observed in different medical imaging applications. See Fig. 3 for more examples. We can also enforce stronger



(a) Aorta (b) IVUS [1] (c) Multi-Atlas [24] (d) SegTHOR [23]

Fig. 3: Multi-class topological interactions for each dataset.

constraints. For containment, we may require the surrounding class (β in Fig. 2) to be at least d-pixel thick. For exclusion, we may require the gap between two mutually exclusive classes to be at least d-pixel wide. We call these generalized constraints *d-containment* and *d-exclusion*.

Overview of the Proposed Method. Though the aforementioned topological interactions are global constraints, we observe that they can be encoded in a localized manner. Specifically, both containment and exclusion constraints can be rewritten as forbidding certain label combinations for adjacent pixels/voxels. In the example in Fig. 2, β contains α equals to the constraint that a pixel/voxel of label α cannot be adjacent to a pixel/voxel of any label other than β and itself. Exclusion is more straightforward, α and γ are mutually exclusive if any two adjacent pixels/voxels do not have the label pair (α, γ) or (γ, α).

We enforce these constraints into DNN training by proposing a novel topological interaction module. The idea is to go through all pairs of adjacent pixels/voxels and identify the pairs that violate the desired constraints. Pixels belonging to these pairs are the ones inducing errors into the topological interaction. We will refer to them as *critical pixels*. Our topological interaction module will output these critical pixels. Then, we can incorporate the module into training by designing a loss paying extra penalty to these critical pixels.

An efficient implementation of the module, however, is not trivial. Simply looping through all pixels is too expensive to serve as a frequent operation during training. To this end, we propose an efficient implementation of the constraints purely based on convolutional operations (Sec. 3.1). The method is much more efficient and can easily generalize to more challenging d-containment and d-exclusion without much extra computational expense. Finally, in Sec. 3.2, we incorporate the proposed module into training by formulating a loss function penalizing the critical pixels. This ensures the DNNs learn better feature rep-

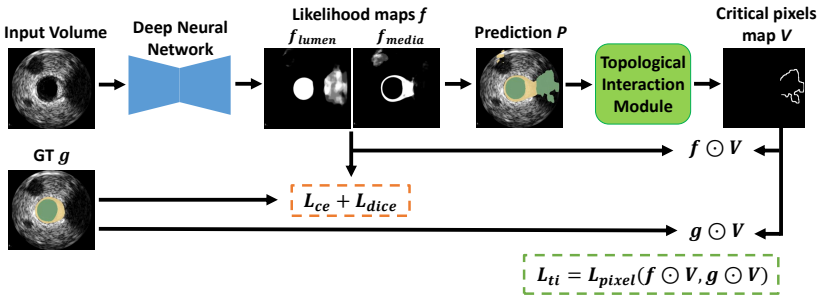


Fig. 4: An overview of the proposed method. The proposed module encodes the topological interactions between the different classes (e.g., *media* and *lumen* classes in the IVUS dataset follow the containment constraint). Critical pixels are identified and used for the new loss L_{ti}.

resentation while respecting the imposed constraints, as we will demonstrate empirically. Fig. 4 provides an overview of the proposed method.

3.1 Topological Interaction Module

The *topological interaction module* encodes the topological interactions defined above. Recall the key is to forbid certain label combinations appearing in any pair of adjacent pixels. Our module identifies the pairs that violate the constraints.

Next, we explain how to map the constraints into the local constraints regarding two labels that should not appear in adjacent pixels. For exclusion constraint, the forbidding label pair is obvious. In Fig. 2, labels α and γ are mutually exclusive. We create new labels $A = \alpha$ and $C = \gamma$, and forbid them to appear in adjacent pixels. For containment constraints, say label β contains label α (as in Fig. 2), we create a new label $A = \alpha$ and a new label C being the union of all other labels except for α and β. Then β containing α is equivalent to $A = \alpha$ not touching C.

For the rest of this section, we focus on how to create a module identifying adjacent pixel pair having the label pair (A, C) or (C, A). For ease of exposition, we assume a 2D 4-connectivity neighborhood (i.e., each pixel is only adjacent to 4 neighboring pixels), and so $d = 1$. The approach can be naturally generalized to other connectivities as formalized in the classic digital topology [34].

Naive Solution. Given a discretized segmentation map predicted by the network, the naive solution is simply looping over all pixels and for each pixel, scan all its neighbors. For every pair of adjacent pixels with the label pair (A, C) or (C, A), we flag both of the pixels as critical. The obvious issue with this naive solution is that it is very expensive. Furthermore, such computation can only run on a CPU, and so is rather slow; this is detailed in the Supplementary Material.[1]

[1] There is an alternate way to better implement this naive solution by creating extra maps representing neighboring pixels. The issue of such a method is it does not

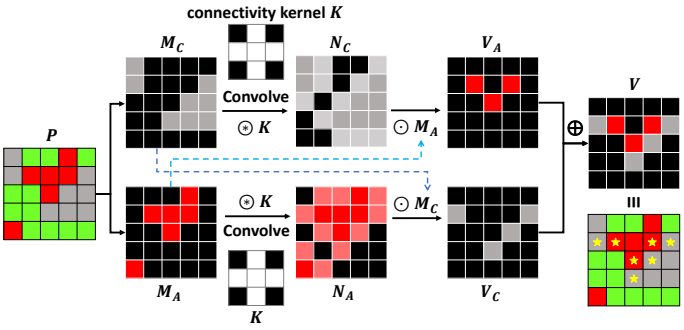


Fig. 5: 2D illustration of the **proposed** strategy to detect the set V of topological critical pixels. We use 4-connectivity kernel. The entire critical pixel map V is highlighted with $*$'s.

Convolution-Based Solution. Let $P \subseteq \mathbf{R}^d$ denote the d-dimensional discrete segmentation map predicted by the network. We want to generate a critical pixel map in which only those label-A pixels with a label-C neighbor are activated and vice-versa. We achieve this goal through manipulations of different semantic masks. First, to determine the critical pixels in A, we expand the C mask by d pixels, and then find out the intersection of the expanded mask with the A mask ($d = 1$ for 2D 4-connectivity). In this way, we obtain the set of all the critical A pixels: they fall within the expanded C mask, and thus must be a neighbor of some C pixels. In top row of Fig. 5, second to fourth columns, we show the C mask (denoted by M_C), its expansion, and the intersection with the A mask (denoted by M_A), resulting in the critical A pixels. In a similar manner, we can obtain the set of critical C pixels by expanding the A mask and finding its intersection with the C mask. This is illustrated in the bottom row, second to fourth columns of Fig. 5.

In practice, expanding a mask can be done efficiently using the *dilation* morphological operation [15]. In dilation, we convolve a given binary mask with a kernel K. The kernel defines the neighbors of a given voxel. Formally, let M_A and M_C be the class masks for A and C respectively. We then obtain neighborhood information N_A and N_C via dilation/convolution as follows:

$$N_A \coloneqq M_A \circledast K, \quad N_C \coloneqq M_C \circledast K \tag{1}$$

where we use $\circledast$ to denote the standard convolution operation. K is the convolution kernel which we refer to as the *connectivity kernel*. As we are dealing with 2D 4-connectivity case, the connectivity kernel used is as shown in Fig. 5. Notice that in map N_A, all the pixels which are in contact with class A get activated. We obtain N_C in a similar way. Now that we have the expanded neighborhood

scale well with larger neighborhood (which is necessary for more general constraints assuming a gap of width d between forbidden label pairs). See the Supplementary Material for more details.

information, and we use this to find which pixels of A and C fall in each other's neighborhood. If V denotes the entire critical pixel map, it can be further divided into V_A and V_C which contain the critical pixels in class A and C respectively. We can then quantify them as:

$$V_A \coloneqq M_A \odot N_C, \quad V_C \coloneqq M_C \odot N_A, \quad V \coloneqq V_A \oplus V_B \tag{2}$$

where $\oplus$ denotes the union operation, and $\odot$ denotes the Hadamard product.

Fig. 5 gives an overview of our method to compute topological critical pixels in the form of a binary mask V. Thus through the manipulation of maps obtained via standard convolution, we are able to augment existing information by deriving information relevant to topological interactions.

Remark on the Connectivity Kernel K. We remark that the connectivity kernel K depends on the definition of neighborhood. Our current choice of K corresponds to the 4-connectivity neighborhood (illustrated in Fig. 5). In general, we can choose different neighborhood definitions corresponding to different kernels. Following the classic digital topology [34], in 2D, we can have 4- and 8-connectivities. In 3D, we can have 6- and 26-connectivities. We can also specify different connectivity kernels for classes A and C. See Supplementary Material for illustrations.[2]

We also note it is natural to generalize the neighborhood definition and modify the kernel accordingly to enforce the more general/stronger constraints: d-containment and d-exclusion. These constraints essentially boil down to the constraint that labels A and C cannot appear on two pixels within distance d. To encode such constraints, we simply define the neighborhood of a pixel p to be all pixels within a $(2d+1) \times (2d+1)$ local patch centered at p. The connectivity kernel is then an all-one kernel of the same size.

Computational Efficiency. We analyze the computational efficiency of the proposed method by determining its complexity as a function of the input and neighborhood size. Let the image size be $N \times N$. Suppose we enforce a separation of d pixels, then the neighborhoods to be inspected for each pixel will be $k \times k$, where $k = 2d+1$. In the naive solution, we require scanning the neighborhood of each pixel via loops and so the time complexity is in the order of $O(N^2k^2)$, not really scalable. This is apart from the fact that such a solution can only run on a CPU. On the contrary, the convolution-based solution has a time complexity $O(N^2 \log N)$. Here $\log N$ is due to the FFT (Fast Fourier Transform) implementation of convolution. While the naive solution's running time is quadratic to k, our proposed is independent of k due to FFT. In practice, deep learning frameworks are highly optimized for convolution operations, and so they are several orders of magnitude cheaper than the naive solution. The memory requirement for both methods is similar in the order of $O(N^2)$ to store the map V.

[2] In digital topology, to ensure the Jordan curve theorem is correct, one needs to have either 4-conn. for foreground and 8-conn. for background, or the opposite. This is not in conflict with our method. A and C are both considered foreground labels. In 2D, they can use either 4-conn. or 8-conn. as long as they are the same. Similar rules apply to 3D.

3.2 Incorporating into End-to-End Training

To incorporate the proposed topological interaction module into end-to-end training, we propose a topological interaction loss to correct the violations by penalizing the critical pixels.

Let $f \in \mathbb{R}^{c \times H \times W}$ be the multi-class likelihood map predicted by the network, where c, H and W denote the number of classes, height and width of the image, respectively. $g \in \mathbb{R}^{H \times W}$ is the ground truth segmentation map with discrete labels, $0, 1, ..., c-1$. We use L_{pixel} to denote the pixel-wise loss function, such as, cross-entropy, mean-squared-error, or dice losses. We use the binary mask V obtained from Sec. 3.1, to define L_{ti}, denoting the additional topological interaction loss, as:

$$L_{ti} = L_{pixel}(f \odot V, g \odot V) \tag{3}$$

L_{ti} can essentially encode the topological interactions, correct the topological interaction errors, and eventually produce a segmentation that is topologically correct. The final loss of our method, L_{total}, is given by:

$$L_{total} = L_{ce} + \lambda_{dice} L_{dice} + \lambda_{ti} L_{ti} \tag{4}$$

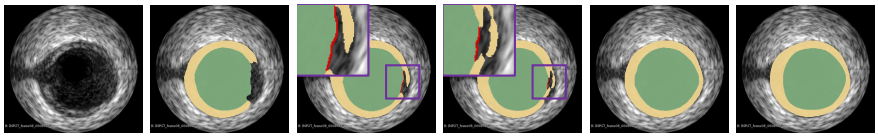
where L_{ce} and L_{dice} denote the cross-entropy and dice loss. The loss is controlled by the weights λ_{dice} and λ_{ti}.

4 Experiments

Datasets. We validate our method on four datasets: The proprietary **Aorta** dataset contains 3D CT scans of 28 patients from an institutional database of patients with thoracic and/or abdominal aortic aneurysm. The **IVUS** (IntraVascular Ultrasound) [1] is a 2D dataset of human coronary arteries and contains lumen and media-adventitia labels. The **Multi-Atlas BTCV** [24] is a multi-organ segmentation challenge, containing 3D CT scans of the cervix and abdomen. We use the abdomen dataset and segment four classes, namely, spleen, left kidney, liver, and stomach which appear in close proximity. We have clinically verified that the exclusion constraint holds among these four classes. The **SegTHOR** [23] 2019 challenge contains 3D CT scans of thoracic organs at risk (OAR). In this dataset, the OARs are the heart, trachea, aorta and esophagus. The exclusion constraint holds among three classes, that is, the trachea, the aorta, and the esophagus do not touch each other. We do not take the heart class into consideration.

The containment constraint holds for the Aorta and IVUS datasets, while the exclusion constraint holds for the remaining two. Fig. 3 gives an overview of the classes in each dataset and the topological interactions among them.

Baselines and Implementation Details. We use the PyTorch framework, a single NVIDIA Tesla V100-SXM2 GPU (32G Memory) and a Dual Intel Xeon Silver 4216 CPU@2.1Ghz (16 cores) for all the experiments. The comparison baselines consist of the UNet [33,7], FCN [30], nnUNet [21], TopoCRF [2], MIDL [32],



(a) Input (b) nnUNet (c) ≈Ep. 70 (d) ≈Ep. 140 (e) ≈Ep. 200 (f) GT

Fig. 6: Epoch (*Ep.*) progression of the proposed method. Critical pixel map identified by the module is marked in red.

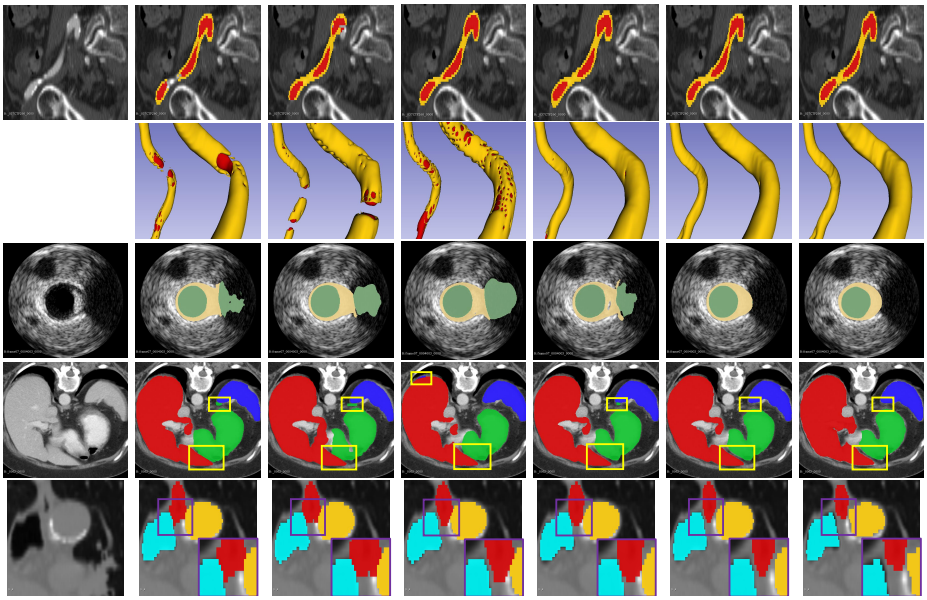
and NonAdj [13]. We use the publicly available codes for UNet, FCN, nnUNet, and NonAdj. For TopoCRF and MIDL, we implemented it by ourselves in PyTorch. Specifically, for TopoCRF, MIDL, NonAdj and our proposed method, we fine-tune the models pre-trained by nnUNet. To support our claim that our method can be incorporated into any backbone, we train our module on FCN and UNet backbones as well. More details and additional results are included in the Supplementary Material.

The connectivity kernel K, in 2D, is a 3×3 kernel filled with 1's to enforce 8-connectivity. Similarly in 3D, K is a $3 \times 3 \times 3$ kernel filled with 1's to enforce 26-connectivity. We also perform an ablation study on the connectivity kernel in the Supplementary Material.

Evaluation Metrics. Dice score [40], Hausdorff distance (HD) [20], and average symmetric surface distance (ASSD) [17] are used as the performance metrics. We introduce a new metric called the *% violations*. The % violations is calculated by the number of pixels violating the constraint as a fraction of the total number of foreground class pixels/voxels. We report the % violations for all the pixels/voxels together instead of separately per class. For all metrics, we report the means and standard deviations. We also perform the unpaired t-test [37] to determine the statistical significance of the improvement. The statistically significant better performances are highlighted with bold in all the tables. The t-test [37] used to determine the statistical significance of the improvement has a confidence interval of 95%. The best, while not statistically significant, performances are highlighted with italics.

4.1 Results

Tab. 1 shows the quantitative results for the containment constraint on the Aorta and IVUS datasets, while Tab. 2 shows the quantitative results for the exclusion constraint on the Multi-Atlas (Abdominal) and SegTHOR datasets. In Fig. 7, we show the qualitative comparison of different methods. The comprehensive quantitative and qualitative results of our method on the UNet and FCN backbones, along with different connectivity kernels can be found in the Supplementary Material. In general, we observe that learning the topological constraint leads to better feature representation and thus better segmentations both qualitatively and quantitatively. We discuss the results for both interactions below.



(a) Input (b) nnUNet (c) CRF (d) MIDL (e) NonAdj (f) Ours (g) GT

Fig. 7: Qualitative results compared with the baselines. Top three rows deal with the containment constraint, while bottom two rows deal with the exclusion constraint. Aorta: rows 1-2, IVUS: row 3, Multi-Atlas: row 4, SegTHOR: row 5. The second row is the 3D view of the first row. It is hard to visualize the input 3D volumetric image and so we leave it blank in the second row. Colors for the classes correspond to the ones used in Fig. 3.

Quantitative and Qualitative Results for Containment Constraint. From Tab. 1, we observe that the proposed method improves the quality of segmentations by improving all the metrics significantly. In Fig. 7, we see that the networks trained with the proposed method have considerably fewer topological violations compared to the other baseline networks. In the top two rows of the figure, we see that the proposed method fixes the topological interaction errors by enforcing the lumen always be enclosed by the wall. By enforcing this constraint, our method is able to reconstruct the broken lumen and wall structures, thereby significantly improving the segmentation quality. In the third row, we show results on the IVUS dataset. Due to artifacts in the input (like shadow), nnUNet erroneously classifies extraneous lumen regions beyond the media. Due to the smoothness loss component in TopoCRF, the boundaries of its segmentations are a lot smoother compared to nnUNet, however, it also fails to correct the lumen prediction. MIDL performs similarly as TopoCRF, and while NonAdj performs better than both of them, it still fails in several places. By enforcing the containment constraint, our method is able to learn better features and gets rid of such extraneous lumen regions.

Table 1: Quantitative comparison for containment constraint

Class	Model	Dice↑	HD↓	ASSD↓	% Violations↓
		Aorta dataset			
Lumen	UNet [7]	0.900 ± 0.016	64.392 ± 16.874	9.315 ± 1.749	13.994 ± 1.809
	FCN [30]	0.894 ± 0.013	57.974 ± 19.756	9.77 ± 1.421	15.675 ± 2.409
	nnUNet [21]	0.906 ± 0.020	36.368 ± 12.559	4.563 ± 0.675	5.424 ± 2.461
	Topo-CRF [2]	0.897 ± 0.057	40.162 ± 18.687	5.952 ± 0.999	8.358 ± 2.151
	MIDL [32]	0.912 ± 0.008	32.157 ± 16.270	6.405 ± 0.524	6.377 ± 1.661
	NonAdj [13]	0.916 ± 0.030	32.465 ± 18.848	4.771 ± 1.129	4.932 ± 1.479
	Ours	**0.922 ± 0.009**	**25.959 ± 13.574**	**3.920 ± 0.765**	**3.526 ± 1.244**
Wall	UNet [7]	0.677 ± 0.015	71.109 ± 24.653	12.497 ± 1.372	/
	FCN [30]	0.651 ± 0.015	66.059 ± 17.188	12.339 ± 0.959	/
	nnUNet [21]	0.741 ± 0.026	42.486 ± 15.139	8.005 ± 0.811	/
	Topo-CRF [2]	0.739 ± 0.010	46.873 ± 17.636	7.914 ± 0.877	/
	MIDL [32]	0.742 ± 0.028	43.132 ± 15.624	6.420 ± 1.242	/
	NonAdj [13]	0.748 ± 0.017	38.197 ± 19.598	4.887 ± 0.702	/
	Ours	**0.758 ± 0.017**	**31.137 ± 17.772**	**5.799 ± 0.737**	/
		IVUS dataset			
Lumen	UNet [33]	0.786 ± 0.144	6.643 ± 1.936	30.944 ± 11.631	5.970 ± 2.141
	FCN [30]	0.824 ± 0.071	5.319 ± 1.519	22.551 ± 7.882	3.766 ± 1.444
	nnUNet [21]	0.893 ± 0.066	3.464 ± 0.917	11.152 ± 3.954	2.708 ± 1.032
	Topo-CRF [2]	0.887 ± 0.096	4.138 ± 1.454	10.497 ± 2.487	2.371 ± 0.960
	MIDL [32]	0.891 ± 0.073	4.226 ± 1.390	10.641 ± 2.322	2.394 ± 0.918
	NonAdj [13]	0.897 ± 0.081	3.140 ± 1.154	9.628 ± 3.221	2.173 ± 0.994
	Ours	**0.949 ± 0.070**	**2.046 ± 1.079**	**6.057 ± 2.746**	**0.157 ± 0.808**
Media	UNet [33]	0.651 ± 0.130	7.391 ± 1.072	21.984 ± 6.634	/
	FCN [30]	0.782 ± 0.144	6.806 ± 1.147	13.863 ± 4.511	/
	nnUNet [21]	0.856 ± 0.090	5.646 ± 1.228	6.491 ± 2.314	/
	Topo-CRF [2]	0.843 ± 0.106	5.409 ± 1.166	5.929 ± 1.785	/
	MIDL [32]	0.841 ± 0.121	5.461 ± 1.214	6.071 ± 1.837	/
	NonAdj [13]	0.848± 0.117	5.983 ± 1.342	6.615 ± 1.937	/
	Ours	**0.910 ± 0.089**	**3.873 ± 0.933**	**3.171 ± 1.871**	/

For both the Aorta and IVUS datasets, by identifying the critical pixels, our method improves the learning capability of the network through the epochs. In Fig. 6, we show how our method improves the network predictions through the epochs on an IVUS data sample. Our results demonstrate that our proposed method is able to significantly improve the segmentation quality without the need for any additional post-processing.

Quantitative and Qualitative Results for Exclusion Constraint. For the Multi-Atlas dataset, our method brings in the greatest improvement for the stomach and liver classes. As can be seen in fourth row of Fig. 7, it is correctly able to separate these two classes while the other methods fail to do so. This correlates with the quantitative metrics as well. In the case of the spleen and kidney classes, nnUNet itself predicts separation between these two classes. Our method improves the dice score slightly, but significantly improves other metrics like HD and ASSD. For the SegTHOR dataset, our method brings in the greatest improvement for the esophagus and trachea classes which tend to come in contact at several points across their lengths. In the final row of Fig. 7, we show that our proposed method is able to impose the exclusion constraint among the three classes. For the aorta class, nnUNet is largely able to separate it from the other classes, and so our method's performance on this class is comparable to nnUNet. We include results for the aorta and heart classes in the Supplementary Material.

Table 2: Quantitative comparison for exclusion constraint

Class	Model	Dice↑	HD↓	ASSD↓	% Violations↓
		Multi-Atlas dataset			
Spleen	UNet [7]	0.919 ± 0.041	47.037 ± 17.365	4.323 ± 0.367	1.857 ± 0.123
	FCN [30]	0.909 ± 0.037	134.915 ± 65.623	17.646 ± 10.604	3.041 ± 0.181
	nnUNet [21]	0.950 ± 0.041	6.084 ± 1.078	0.573 ± 0.131	0.819 ± 0.064
	Topo-CRF [2]	0.947 ± 0.028	6.403 ± 1.039	1.844 ± 0.517	0.934 ± 0.032
	MIDL [32]	0.944 ± 0.015	5.597 ± 1.374	0.565 ± 0.124	0.725 ± 0.151
	NonAdj [13]	0.952 ± 0.058	5.621 ± 1.065	0.513 ± 0.175	0.521 ± 0.082
	Ours	*0.960 ± 0.009*	**5.340 ± 1.049**	**0.484 ± 0.109**	**0.464 ± 0.043**
Kidney	UNet [7]	0.908 ± 0.079	61.602 ± 13.168	9.992 ± 2.461	/
	FCN [30]	0.892 ± 0.018	187.472 ± 36.096	11.583 ± 2.396	/
	nnUNet [21]	0.931 ± 0.018	27.252 ± 5.406	5.352 ± 0.199	/
	Topo-CRF [2]	0.928 ± 0.059	30.209 ± 5.317	6.308 ± 0.905	/
	MIDL [32]	0.935 ± 0.071	25.208 ± 5.440	4.885 ± 0.421	/
	NonAdj [13]	0.934 ± 0.012	24.182 ± 5.561	4.692 ± 0.657	/
	Ours	*0.936 ± 0.026*	**20.013 ± 2.785**	**4.298 ± 0.798**	/
Liver	UNet [7]	0.912 ± 0.016	64.556 ± 13.894	2.324 ± 0.513	/
	FCN [30]	0.885 ± 0.034	183.870 ± 49.796	29.061 ± 13.484	/
	nnUNet [21]	0.951 ± 0.008	38.931 ± 12.161	1.922 ± 0.506	/
	Topo-CRF [2]	0.949 ± 0.006	46.449 ± 14.188	2.072 ± 0.313	/
	MIDL [32]	0.955 ± 0.005	34.276 ± 11.253	1.344 ± 0.431	/
	NonAdj [13]	0.957 ± 0.003	33.671 ± 13.543	1.185 ± 0.372	/
	Ours	**0.962 ± 0.005**	**30.341 ± 9.111**	**0.985 ± 0.386**	/
Stomach	UNet [7]	0.846 ± 0.084	76.000 ± 24.352	5.023 ± 1.508	/
	FCN [30]	0.708 ± 0.156	172.855 ± 43.735	11.328 ± 3.178	/
	nnUNet [21]	0.895 ± 0.015	45.767 ± 7.960	2.720 ± 0.430	/
	Topo-CRF [2]	0.888 ± 0.015	46.877 ± 9.861	3.675 ± 0.358	/
	MIDL [32]	0.899 ± 0.012	40.282 ± 6.437	2.567 ± 0.431	/
	NonAdj [13]	0.907 ± 0.028	41.749 ± 8.630	2.184 ± 0.325	/
	Ours	**0.910 ± 0.018**	**35.514 ± 10.295**	**1.644 ± 0.311**	/
		SegTHOR dataset			
Esophagus	UNet [7]	0.827 ± 0.038	11.357 ± 2.709	1.186 ± 0.113	3.212 ± 0.720
	FCN [30]	0.800 ± 0.031	10.770 ± 2.085	1.303 ± 0.128	3.616 ± 0.709
	nnUNet [21]	0.841 ± 0.014	8.018 ± 2.085	0.950 ± 0.070	1.947 ± 0.525
	Topo-CRF [2]	0.839 ± 0.029	8.602 ± 2.363	0.991 ± 0.081	2.070 ± 0.687
	MIDL [32]	0.840 ± 0.020	7.266 ± 2.132	0.921 ± 0.136	1.271 ± 0.912
	NonAdj [13]	0.843 ± 0.020	6.293 ± 2.703	0.897 ± 0.078	1.215 ± 0.211
	Ours	**0.858 ± 0.019**	**5.582 ± 2.250**	**0.798 ± 0.042**	**0.749 ± 0.428**
Trachea	UNet [7]	0.897 ± 0.027	10.656 ± 4.047	0.728 ± 0.146	/
	FCN [30]	0.891 ± 0.031	11.789 ± 5.291	0.953 ± 0.221	/
	nnUNet [21]	0.910 ± 0.018	9.423 ± 2.393	0.478 ± 0.152	/
	Topo-CRF [2]	0.909 ± 0.022	10.435 ± 2.334	0.473 ± 0.167	/
	MIDL [32]	0.914 ± 0.027	7.929 ± 2.305	0.456 ± 0.143	/
	NonAdj [13]	0.913 ± 0.028	7.866 ± 2.343	0.440 ± 0.113	/
	Ours	**0.929 ± 0.020**	**7.280 ± 2.109**	**0.316 ± 0.186**	/

4.2 Ablation Studies

To further demonstrate the efficacy of the proposed method, we conduct several ablation studies. The following ablation studies have been performed on the IVUS dataset (containment constraint). We perform identical ablation studies on the Multi-Atlas dataset (exclusion constraint) in the Supplementary Material.

Ablation Study for Loss Functions. Our additional topological interaction loss L_{ti} is a general term, and can adopt any existing pixel-wise loss function. We conduct an ablation study using three different loss functions for L_{pixel}, the cross-entropy loss (CE), the mean-squared-error loss (MSE), and the dice loss. The results are tabulated in the top half of Tab. 3, where the *None* entry denotes nnUNet trained without L_{ti}. Using CE for L_{pixel} gives the best performance.

Table 3: Ablation study for L_{pixel} and λ_{ti} (IVUS)

Class	L_{pixel}	Dice↑	HD↓	ASSD↓	% Violations↓
Lumen	None	0.893 ± 0.066	3.464 ± 0.917	11.152 ± 3.954	2.708 ± 1.032
	MSE	0.915 ± 0.073	3.162 ± 0.937	9.963 ± 3.086	0.835 ± 0.907
	DICE	0.937 ± 0.067	2.385 ± 1.065	6.520 ± 2.845	0.320 ± 0.811
	CE	**0.949 ± 0.070**	**2.046 ± 1.079**	**6.057 ± 2.746**	**0.157 ± 0.808**
Media	None	0.856 ± 0.090	5.646 ± 1.228	6.491 ± 2.314	/
	MSE	0.893 ± 0.087	4.042 ± 0.986	3.874 ± 1.912	/
	DICE	0.896 ± 0.088	3.964 ± 1.112	3.445 ± 1.681	/
	CE	**0.910 ± 0.089**	**3.873 ± 0.933**	**3.171 ± 1.871**	/
Class	λ_{ti}	**Dice↑**	**HD↓**	**ASSD↓**	**% Violations↓**
Lumen	0	0.893 ± 0.066	3.464 ± 0.917	11.152 ± 3.954	2.708 ± 1.032
	5.0e-5	0.913 ± 0.071	3.249 ± 0.998	9.338 ± 3.649	0.964 ± 0.893
	1.0e-4	**0.949 ± 0.070**	**2.046 ± 1.079**	**6.057 ± 2.746**	**0.157 ± 0.808**
	1.5e-4	0.941 ± 0.069	2.124 ± 1.062	6.426 ± 2.976	0.187 ± 0.814
	2.0e-4	0.938 ± 0.070	2.428 ± 1.041	6.558 ± 2.780	0.252 ± 0.830
Media	0	0.856 ± 0.090	5.646 ± 1.228	6.491 ± 2.314	/
	5.0e-5	0.877 ± 0.088	5.099 ± 0.997	5.024 ± 2.100	/
	1.0e-4	**0.910 ± 0.089**	**3.873 ± 0.933**	**3.171 ± 1.871**	/
	1.5e-4	0.905 ± 0.088	3.889 ± 0.919	3.257 ± 1.877	/
	2.0e-4	0.885 ± 0.089	4.319 ± 1.059	4.364 ± 1.943	/

However, using any of the choices for L_{pixel} results in improvement across all metrics compared to the vanilla nnUNet. Thus L_{ti} is a generic term which works towards its intended purpose of correcting topological errors irrespective of the choice of L_{pixel}.

Ablation Study for Loss Weights. Since the topological loss is the main contribution of this paper, we conduct another ablation study in terms of its weight λ_{ti}. We run the experiments with different weights for the additional topological interaction loss and report the results in the bottom half of Tab. 3. When λ_{ti}=1e-4, the proposed method achieves the best performance. However, a reasonable range of λ_{ti} always results in improvement. This demonstrates the efficacy and robustness of the proposed method.

5 Conclusion

We introduce a new convolution-based module for multi-class image segmentation that focuses on topological interactions. The module consists of an efficient algorithm to identify critical pixels which induce topological errors. We also introduce an additional topologically constrained loss function. By incorporating the module as well as the loss function into the training of deep neural networks, we enforce the network to learn better feature representations, resulting in improved segmentation quality. Results suggest that the method is generalizable to both 2D and 3D settings, and across modalities such as US and CT.

Acknowledgements. We thank the anonymous reviewers for their constructive feedback. The reported research was partly supported by grants NSF IIS-1909038 and NIH 1R21CA258493-01A1.

References

1. Balocco, S., Gatta, C., Ciompi, F., Wahle, A., Radeva, P., Carlier, S., Unal, G., Sanidas, E., Mauri, J., Carillo, X., et al.: Standardized evaluation methodology and reference database for evaluating ivus image segmentation. Computerized medical imaging and graphics **38**(2), 70–90 (2014)
2. BenTaieb, A., Hamarneh, G.: Topology aware fully convolutional networks for histology gland segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 460–468. Springer (2016)
3. Chen, C., Freedman, D., Lampert, C.H.: Enforcing topological constraints in random field image segmentation. In: CVPR (2011)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
6. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
7. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: MICCAI (2016)
8. Clough, J., Byrne, N., Oksuz, I., Zimmer, V., Schnabel, J., King, A.: A topological loss function for deep-learning based image segmentation using persistent homology. TPAMI (2020)
9. Colliot, O., Camara, O., Bloch, I.: Integration of fuzzy spatial relations in deformable models—application to brain mri segmentation. Pattern recognition **39**(8), 1401–1414 (2006)
10. Delong, A., Boykov, Y.: Globally optimal segmentation of multi-region objects. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 285–292. IEEE (2009)
11. Doweidar, M.H.: Advances in Biomechanics and Tissue Regeneration. Academic Press (2019)
12. Felzenszwalb, P.F., Veksler, O.: Tiered scene labeling with dynamic programming. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3097–3104. IEEE (2010)
13. Ganaye, P.A., Sdika, M., Triggs, B., Benoit-Cattin, H.: Removing segmentation inconsistencies with semi-supervised non-adjacency constraint. Medical image analysis **58**, 101551 (2019)
14. Han, X., Xu, C., Prince, J.L.: A topology preserving level set method for geometric deformable models. TPAMI (2003)
15. Haralick, R.M., Sternberg, S.R., Zhuang, X.: Image analysis using mathematical morphology. TPAMI (1987)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
17. Heimann, T., Van Ginneken, B., Styner, M.A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., et al.: Comparison and evaluation of methods for liver segmentation from ct datasets. IEEE transactions on medical imaging **28**(8), 1251–1265 (2009)

18. Hu, X., Li, F., Samaras, D., Chen, C.: Topology-preserving deep image segmentation. NeurIPS (2019)
19. Hu, X., Wang, Y., Fuxin, L., Samaras, D., Chen, C.: Topology-aware segmentation using discrete morse theory. ICLR (2021)
20. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence **15**(9), 850–863 (1993)
21. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods (2021)
22. Kappes, J.H., Speth, M., Reinelt, G., Schnörr, C.: Higher-order segmentation via multicuts. Computer Vision and Image Understanding **143**, 104–119 (2016)
23. Lambert, Z., Petitjean, C., Dubray, B., Ruan, S.: Segthor: Segmentation of thoracic organs at risk in ct images (2019)
24. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)
25. Le Guyader, C., Vese, L.A.: Self-repelling snakes for topology-preserving segmentation models. TIP (2008)
26. Leon, L.M.C., De Miranda, P.A.V.: Multi-object segmentation by hierarchical layered oriented image foresting transform. In: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). pp. 79–86. IEEE (2017)
27. Li, K., Wu, X., Chen, D.Z., Sonka, M.: Optimal surface segmentation in volumetric images-a graph-theoretic approach. IEEE transactions on pattern analysis and machine intelligence **28**(1), 119–134 (2005)
28. Litvin, A., Karl, W.C.: Coupled shape distribution-based segmentation of multiple objects. In: Biennial International Conference on Information Processing in Medical Imaging. pp. 345–356. Springer (2005)
29. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
30. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
31. Nosrati, M.S., Hamarneh, G.: Local optimization based segmentation of spatially-recurring, multi-region objects with part configuration constraints. IEEE transactions on medical imaging **33**(9), 1845–1859 (2014)
32. Reddy, C., Gopinath, K., Lombaert, H.: Brain tumor segmentation using topological loss in convolutional networks. In: International Conference on Medical Imaging with Deep Learning–Extended Abstract Track (2019)
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
34. Rosenfeld, A.: Digital topology. The American Mathematical Monthly **86**(8), 621–630 (1979)
35. Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P., Bauer, U., Menze, B.H.: cldice-a novel topology-preserving loss function for tubular structure segmentation. In: CVPR (2021)
36. Strekalovskiy, E., Cremers, D.: Generalized ordering constraints for multilabel optimization. In: 2011 International Conference on Computer Vision. pp. 2619–2626. IEEE (2011)

37. Student: The probable error of a mean. Biometrika pp. 1–25 (1908)
38. Ulén, J., Strandmark, P., Kahl, F.: An efficient optimization framework for multi-region segmentation based on lagrangian duality. IEEE transactions on medical imaging **32**(2), 178–188 (2012)
39. Yang, J., Hu, X., Chen, C., Tsai, C.: A topological-attention convlstm network and its application to em images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 217–228. Springer (2021)
40. Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M., Kaus, M.R., Haker, S.J., Wells III, W.M., Jolesz, F.A., Kikinis, R.: Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. Academic radiology **11**(2), 178–189 (2004)