CoordConv-Unet: Investigating CoordConv for Organ Segmentation

R. El Jurdi, C. Petitjean, Paul Honeine, Fahed Abdallah

▶ **To cite this version:**

R. El Jurdi, C. Petitjean, Paul Honeine, Fahed Abdallah. CoordConv-Unet: Investigating CoordConv for Organ Segmentation. Innovation and Research in BioMedical engineering, 2021, ⟨10.1016/j.irbm.2021.03.002⟩. ⟨hal-03410507⟩

HAL Id: hal-03410507

https://normandie-univ.hal.science/hal-03410507v1

Submitted on 5 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC 4.0 - Attribution - Non-commercial use - International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S1959031821000324
Manuscript_deb26845b261e10c28b69d427fb3405f

CoordConv-Unet: Investigating CoordConv for Organ Segmentation

Rosana El Jurdi[a,b], Caroline Petitjean[a], Paul Honeine[a], Fahed Abdallah[b,c]

[a] *LITIS Lab, Université de Rouen Normandie, Saint-Etienne-du-Rouvray, France*
[b] *Université Libanaise, Hadath, Beyrouth, Liban*
[c] *ICD, M2S, Université de technologie de Troyes, Troyes, France.*

Abstract

Objectives:
Convolutional neural networks (CNNs) have established state-of-the-art performance in computer vision tasks such as object detection and segmentation. One of the major remaining challenges concerns their ability to capture consistent spatial and anatomically plausible attributes in medical image segmentation. To address this issue, many works advocate to integrate prior information at the level of the loss function. However, prior-based losses often suffer from local solutions and training instability. The CoordConv layers are extensions of convolutional neural network wherein convolution is conditioned on spatial coordinates. The objective of this paper is to investigate CoordConv as a proficient substitute to convolutional layers for medical image segmentation tasks when trained under prior-based losses.
Methods:
This work introduces CoordConv-Unet which is a novel structure that can be used to accommodate training under anatomical prior losses. The proposed architecture demonstrates a dual role relative to prior constrained CNN learning: it either demonstrates a regularizing role that stabilizes learning while maintaining system performance, or improves system performance by allowing the learning to be more stable and to evade local minima.
Results:
To validate the performance of the proposed model, experiments are conducted on two well-known public datasets from the Decathlon challenge: a mono-modal MRI dataset dedicated to segmentation of the left atrium, and a CT image dataset whose objective is to segment the spleen, an organ characterized with varying size and mild convexity issues.
Conclusion:
Results show that, despite the inadequacy of CoordConv when trained with the regular dice baseline loss, the proposed CoordConv-Unet structure can improve significantly model performance when trained under anatomically constrained prior losses.

Keywords: Medical Image Segmentation, Fully Convolutional Networks, Prior-based Losses, CoordConv, MRI, CT

1. Introduction

Medical image segmentation is the process of classifying each pixel within an image into an instance corresponding to an anatomical object of interest. Generally, medical images are largely versatile in nature, depending on the acquisition process and the type of object to be segmented. They can be acquired with magnetic resonance imaging (MRI), computed tomography (CT), nuclear medicine functional imaging, ultrasound imaging, fundus photography, to name a few. Hence, they vary in characteristics and nature and are broad with regards to the anatomical object of interest. As such, guaranteeing high performance within medical image segmentation can be considered very challenging when compared to other types of images or segmentation tasks. Regardless, segmentation within the medical domain is considered a key step in performing non-invasive diagnostic procedures, assisting early disease detection or surgical planning.

Early attempts to automate or semi-automate the process of medical image segmentation started with optimization-based approaches. Such methods generally involve optimization of an energy functional, where the image can be considered continuous [1] or discrete (eg. a graph [2, 3]). In order to counter-react noise and low contrast, works have aimed to perform optimization processes in such a way that the automated predictions conform with particular rules relative to the anatomical object characteristics [4]. This information may include the object appearance, size, smoothness or compactness [5, 6, 7]. These rules and characteristics are known as prior knowledge of the datasets. Anatomical priors refer to medical knowledge and domain expertise that capture spatial as well as topological guidelines and character-

© 2021 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

istics with respect to the understudied anatomical objects.

In the recent era, deep learning has registered a pivotal milestone in many fields including pattern recognition, object detection, natural language processing, with medical image segmentation being no exception to the rule. Convolutional neural networks (CNNs), a class of deep learning models, have been known to register considerable results due to their generalization ability and powerful predictive notions. Due to the fact that not only what is inside the image is to be specified, but also where, semantic segmentation through CNNs must consider a trade-off between contextual and spatial understanding.

First CNN architectures known for their success in image segmentation are the fully convolutional neural networks (FCNs) [8]. FCNs are structures derived from typical deep classification models such as VGG16, AlexNet or GoogLeNet by removing the corresponding classification layers, replacing their fully connected layers with convolutional ones and adding an upsampling layer that is dedicated to transforming coarse outputs into dense predictions. FCNs are very well-known in the medical [9, 10] and non-medical [11] fields as they have paved the way for encoder-decoder networks for segmentation problems.

Many works within the field advocate to go deeper with FCN layers in order to increase the depth and precision of the learnt contextual features [12, 13]. However, increasing model's prediction ability by adding additional layers is not an easy task. Thus, as one goes deeper within the layers, insight on location features are hence lost. As a result, deep FCNs often fail to consider global and spatial information and are prone to producing fuzzy coarse-grained predictions [14]. Moreover, deepening the convolutional network will often increase the model's complexity, thus subjecting the training to additional challenges such as vanishing gradients. As a result, deep FCN may suffer from performance saturation or degradation while training.

One powerful architecture known for the ability to preserve semantic information while achieving promising segmentation performance is the well-known U-Net [15]. U-Net is a symmetrical encoder/decoder structure composed of a contracting path of stacked convolutional and max pooling layers representing the encoder branch and a corresponding expanding path composed of deconvolutional layers representing the decoder. U-Net is able to achieve a trade-off between extracting contextual features from the encoder convolutions on one hand and semantic features from the decoder convolutions on the other hand by concatenating their respective feature maps from different levels of the corresponding symmetric encoder and decoder layers via skip-connections. In this way, U-Net combines low-level detail and contextual information with high-level semantic and location attributes, thus achieving a trade-off between the two. Several variants of U-Net consist in changing the backbone model used for encoding, e.g. VGG and DenseNet, and/or replacing deconvolution layers with super-resolution ones for more concise localization ability [16, 17, 18].

There is no doubt that, thanks to the complex and powerful architectures such as U-Net and its variants, the segmentation performance has reached a serious breakthrough. Even so, multiple challenges still remain within medical imaging. Thus, these deep networks often require large amounts of annotated training data, which is not easy to obtain given the medical field. Rather, unannotated or partially labeled data are more easily available or less computationally expensive. Moreover, even with sufficient data, automated systems generally and CNNs particularly still lack the anatomical plausibility that a medical expert has.

Prior to deep learning, prior information such as shape and the topology of organs have often been investigated within variational approaches in order to increase the anatomical correctness of automated segmentation. Recent advances in the domain have attempted integrating these prior onto CNN training in order to overcome the problem of lack of data and to evade production of anatomically aberrant errors. For example, the method of [19] constrains segmentations produced by a regular U-Net to conform with particular upper and lower bound sizes of the heart within a cardiac dataset. Another example is in [20] where connected component numbers were preserved by introducing bounding box prior at the level of the skip connections. These bounding boxes allow the network to learn focused features concerning small components of the heart that are dissolved within the deep layers of a normal U-Net baseline network. Nevertheless, prior constraint neural networks still face difficulties regarding which information to model, how they are modeled and integrated into the deep neural networks.

Integration of the prior onto CNN learning can be conducted at the level of the network structure, hereby the name structural constraints [20, 21] or at the level of the loss function [22, 23] or a combination of both [24, 25]. Structure prior consists in designing parts of the network in order to take into consideration these external specifications [20, 26, 27, 21, 25]. Imposing constraint via loss functions, on the other hand, consists of formalizing the prior as an additional term in the loss function. The loss term computes the error demonstrating the degree of violation of the prior constraint. Given the state of the art, prior-based losses have been widely adopted as means of enhancing consistency and plausibility of segments produced by powerful neural network architectures within fully supervised learning [28, 29, 30].

On the other hand, they could be used to counteract the problem of lack of complete data annotations [19].

The types of prior knowledge cover a breadth of notions in the literature [4]. Aside from the low-level prior which includes ground-truth transformation such as distance maps [22, 30] or Laplacian filters [31], high-level prior integrate actual expert knowledge into the learning system. This type of prior can include the size of the organ [19], its compactness [32], its topological structure [33] or its convexity [34] among others. High-level prior losses is currently a growing trend that has captured the attention of many researchers within the field since they offer a versatile way to integrate external knowledge in a generic manner and can be plugged into any backbone.

The CoordConv layers, recently introduced in [35], are extensions of convolutions that allow convolution filter to take into account the spatial coordinates of the pixels. The goal of CoordConv is to learn a mapping between coordinates in the Cartesian space and coordinates in the one-hot pixel space. CoordConv has shown promising potential for object localization [35, 36], and has rightfully raised interest for image segmentation [37, 38]. However, the CoordConv's added value has not been yet assessed in image segmentation generally and in prior guided segmentation particularly.

This paper investigates CoordConv as a proficient substitute to convolutional layers in U-Net models for medical image segmentation. We explore the effect of CoordConv on model performance and rate of convergence when learning is conducted under anatomical prior-based losses, particularly the size loss proposed by [19] and the skeleton loss proposed by [33]. We propose a new U-Net variant based on the CoordConv layer, the CoordConv-Unet, and demonstrate its role in enhancing stability and performance of the above losses. Finally, we expose the dual role of CoordConv-Unet as a regularizer with the ability to stabilize network training under prior-based losses on one hand, and the ability to increase system performance significantly by evading local solutions on the other.

The use of CoordConv with UNet has been used in the papers of [39] and [40]. However, in this work, we investigate the significance of U-Net/CoordConv combinations when trained under prior-based losses which according to our knowledge has not been investigated prior to this paper. The significance of CoordConv is investigated on two datasets: a cardiac dataset which consists of MR images covering the entire atrium. The understudied organ within this dataset is characterized with multi-connected components and large size variability. The spleen dataset is a CT image dataset where the spleen is characterized by a largely varying size and shape convexity issues at boundary level. The contributions of the paper are summarized as follows:

- We investigate the significance of CoordConv solution given organ segmentation under prior-based loss training.
- We propose a novel architecture, the CoordConv-Unet as a proficient substitute to U-Net given prior constrained problems.
- We shed light on the dual role that CoordConv-Unet has in increasing and stabilizing system performance.

The rest of the paper is organized as follows. Section 2 provides a brief overview of the state of the art regarding incorporating constraints onto deep learning networks. Section 3 elaborates on the proposed CoordConv-Unet model as well as the multiple frameworks and paradigms explored. Section 4 presents the datasets and experimental settings. Section 5 evaluates the significance of the proposed CoordConv-Unet relative to the proposed datasets. Finally, Section 6 concludes with future works and perspectives.

2. Related works on Constrained Convolutional Neural Networks

Prior knowledge can be integrated into the CNN learning in the form of structural constraints or at the level of the loss function.

2.1. Structural Prior Constraint

Among structural constraint methods, integration of prior can be done either externally in conjunction with the segmentation network [29, 41, 14, 42, 43] or at the intermediate level [20, 27, 29].

In [41], the authors propose a cascade of several deep CNN architectures that consider multi-scale patches in order to incorporate anatomical location in their decision making process. Thus, spatial location of patches extracted from the image into a CNN model is injected posterior to the convolutional layers. Similar to [41], authors of [29] demonstrate two collaborative architectures in order to iteratively refine the posterior probability and provide information about neighboring organs. In this work, anatomical constraints are obtained from an auxiliary network that are later used by the segmentation network (U-Net) in order to refine ill-defined organ boundaries. Similarly, the SR-UNet in [14] jointly adds an external network to the segmentation one in order to take into consideration the incomplete, over- or under-segmented shape masks provided by the U-Net. However, unlike [29] that fine-tunes ill-defined segmentations by U-Net using the constraints obtained from the external network, [14] maps the segmentations to conform to a manifold of permissible training shapes.

Instead of integrating prior via external networks, BB-UNet as proposed by [20] aims at integrating location prior represented by bounding boxes onto a U-Net at the level of the skip connections in order to capture focused features and preserve connected components properties that are lost in deep layers given normal U-Net functioning. Whereas BB-UNet imposes external constraints at the level of the skip connection, Attention-UNet [21] do so by extracting these constraints from the bottle-neck layers of the baseline U-Net.

2.2. Loss prior constraints

Incorporating prior at the level of the loss function offers a versatile way to constraint neural network predictions while preserving computational complexity and generality. Prior integrated can be low-level, which resembles reformulated ground-truth representation such as: distance maps [22, 30], Laplacian filters [31] or internal layer feature maps [44]. Prior could also be high-level representing actual external medical information such as the shape of the organ, compactness [32] or size [19] and are optimized directly based on ground-truth prior tags.

Low-level Prior knowledge. Both works of [30] and [22] exploit distance maps to improve boundary consistency. Whereas [22] do so by the extraction of shape bio-markers and allow the network to differentiate between hard-to-segment boundaries of different anatomical classes, [30] aims to fine-tune probability outputs by these distance maps in order to overcome the problem of class imbalance between empty/full images with high boundary precision. In the same context, [31] demonstrates a shape aware loss function that constraints predictions to conform to permissible manifold in vertebrae segmentation. In order to do so, the method takes into consideration the average point to curve Euclidean distance factor between predicted and the ground-truth contours. Moreover, authors of [45] exploit Laplacian filters in order to develop a boundary enhanced loss term that invokes the network to generate strong responses around boundary areas while producing a zero response in pixels that are at the periphery. Afar from boundary criteria, [44] is able to close small gaps in neuronal membranes and alleviate topology mistakes by leveraging the topological information or shape descriptors present within the internal layers of VGG16 networks when introduced to both label and predicted segments. All the above-mentioned approaches enhance segmentation consistency by reformulating and refining transformations of posterior probabilities. However, they do not optimize the prior attribute directly between ground-truth and predicted segments. As a result, there is no guarantee that the ground-truth prior specification will be met.

High-Level Prior Loss. In [19] and [46], organ size is taken into consideration where the size prior is directly optimized relative to known ground-truth size bounds. Whereas [19] integrates the prior via an additional penalty loss term and computes the errors relative to approximate upper and lower bounds, [46] relates to the discrete nature of size prior and optimizes the network via discrete based optimization techniques. Another type of prior is topology which is concerned with the properties of spatial objects by abstracting their connectivity, while ignoring their detailed form [47]. Both works of [23] and [48] use notions of persistent homologies, which is a method for capturing topological structures via a series of thresholding on prediction maps in order to evade prohibited broken vessels and connections. Skeletonization is yet another way to represent topological properties of objects and is exploited in [33] in order to conduct vessels and neuron segmentations. Other properties, such as compactness [32], star-shape [34], inter-region relations [49], are also investigated as means of improving segmentation consistency via prior-based losses.

2.3. Combined Structural and Loss Constraints

In many works, authors introduce interchangeably both structural and loss constraints. In [14], a non-linear shape regularization model is trained jointly along U-Net. The main function of their adjoint network is to learn projections of arbitrary shapes onto a manifold space. The method then incorporates a loss function that updates the segmentation network (U-Net) parameters based on the regularized predicted segments, the rough predicted segments as well as the ground-truth labels. The authors of [50] adopt a similar regularization approach to that in [14]. However, they target the decoder layer with their U-Net-like structure and train the up-sampling layers through super resolution ground-truth maps. On the other hand, authors of [24] introduce a boundary enhanced loss function similar to that of [22] and [31]. Instead of weighting by the errors through distance maps, [24] adds an extra decoder branch to the U-Net network in order to predict hard to segment boundaries. In the same manner, the method in [51] integrates center of mass and contour prior into the loss function which was obtained from an encoder/decoder structure trained end-to-end along the segmentation network.

3. Proposed CoordConv-Unet Architecture

In this section, we present the CoordConv layer and its implementation within the segmentation framework. We further elaborate on the proposed CoordConv-Unet model, the corresponding building blocks, as well as the different integration strategies.

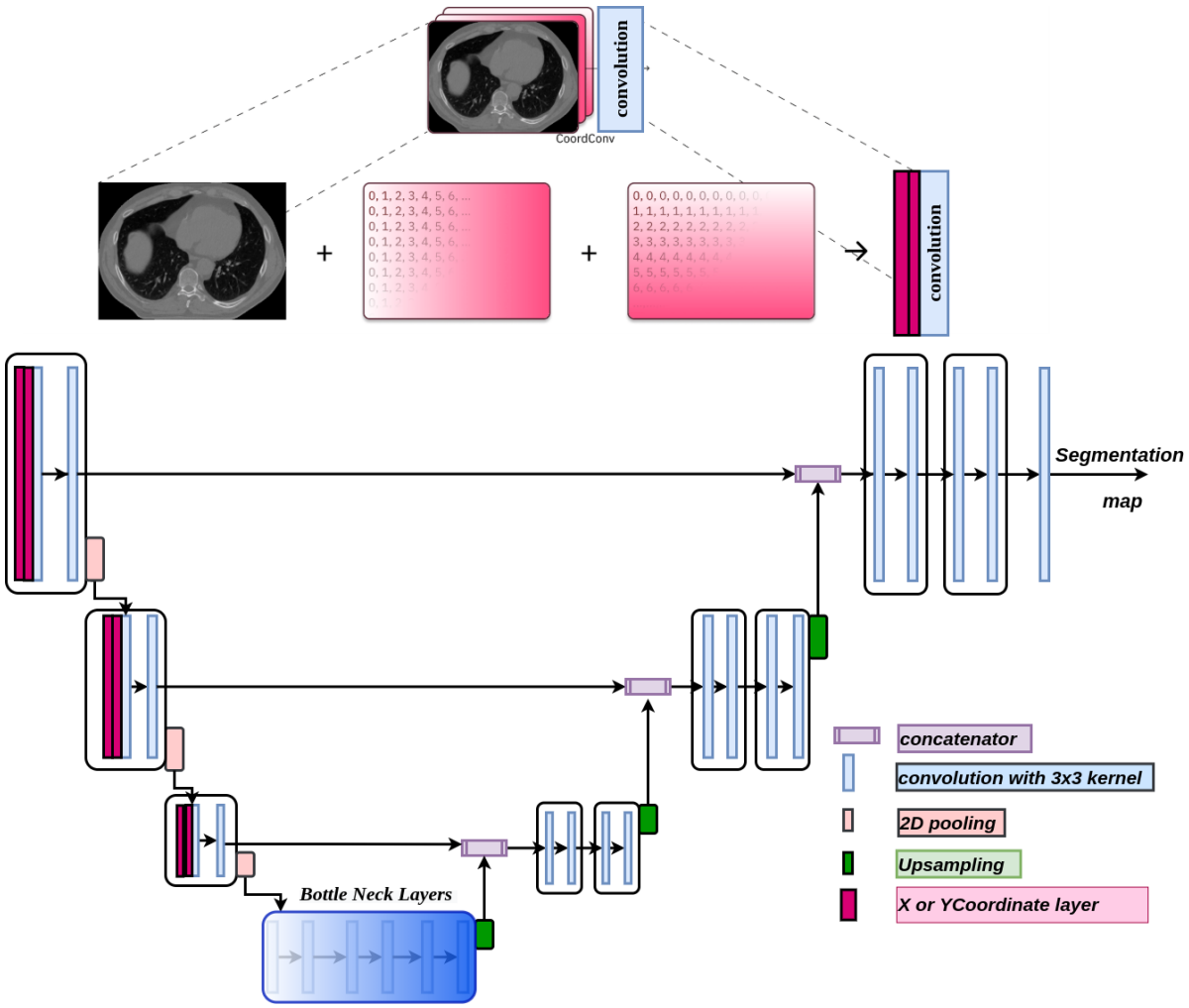


Figure 1: Proposed CoordConv-Unet model. In the top panel, the CoordConv layer consists in concatenating the x-layer and y-layer to the convolutional layer. CoordConv-Unet consists of replacing the first convolutional layer of each stage with the CoordConv layer.

3.1. CoordConv Component

The CoordConv layer is a simple extension of the standard convolutional layer wherein convolution is condition by spatial coordinates. The goal is to establish mappings between the Cartesian space and the pixel space, by enabling the filters to know where pixels are located. In a general sense, convolutions are mainly characterized by three specific characteristics: few training parameters, fast optimization via modern GPUs, and translational invariance. However, given many tasks, there is a controversy with regards to whether translational invariance will truly help model performance or not. CoordConv allows the network to keep or drop the property of translational invariance according to what is needed in the task at hand. In doing so, CoordConv ensures the best of both convolutional and spatial features. The implementation of CoordConv is done by concatenating two additional x and y channels to the input channel as shown in Figure 1(top figure). In such a way, CoordConv allows the learning of a function characterized by a certain degree of translational dependence, if the weights connecting the coordinate layers of the CoordConv with the convolutional are non-zero or could mimic a regular convolutional layer if they were set to zero. In the proposed experiments, CoordConv is implemented via a PyTorch library [1] where a linear scaling is applied in order to bound the values of the coordinate layers between -1 and 1.

3.2. CoordConv-Unet

In the proposed architecture, we extend upon U-Net by replacing convolutional blocks with the CoordConv ones. In such a way, we allow the network to take into consideration spatial and geometric aspects while training. As previously stated, U-Net is a symmetric encoder/decoder structure with equivalent distribution of convolutional and de-convolutional blocks connected via skip connections. Each convolutional block is composed of two consecutive ensembles of convolutional layers and batch normalization, whereas

[1] https://github.com/walsvid/CoordConv

the decoder block adds a bilinear upsampling layer to the previous ensemble. Our main contribution targets the first convolutional layer of the convolutional blocks consisting the U-Net model. Thus, instead of directly convoluting the input of the convolution layer with that of the one before it, rather coordinates for each feature are taken into consideration. The proposed network is represented in Figure 1(bottom figure).

3.3. Computational complexity

Given a U-Net, parameters generally involve the number of learnable quantities or weights connecting a convolutional layer with the corresponding precedent or following layer. In order to quantify the computational overhead induced by the CoordConv component, we will adopt the following mathematical notations. Let m represent the width of convolutional filter and n its height. Let d be the number of filters in the preceding layer and k the number of filters in the current one. Then, the number of parameters involving the conventional convolutional layer is hence computed according to $(mnd + 1)k$, where 1 represents the bias term for each filter. Given CoordConv component, since 2 additional channels are added at the level precedent to the convolutional layer, the number of filters in the preceding layer is hence d+2, thus resulting in a new number of parameters of $(mn(d + 2) + 1)k$. In such a way, each convolutional layer adds to the computational complexity $2mnk$ operations.

3.4. Integration Strategies

We have investigated various integration strategies of the CoordConv block onto the U-Net architecture. The first setting is one that mimics the state of the art, where the x and y channels are added only to the first convolutional layer. We call this model CoordConv(+1). The proposed CoordConv-Unet consists of replacing the first convolutional layer of each convolutional block within the encoding path with the CoordConv layer. We call the proposed method CoordConv-Unet.

3.4.1. Loss functions

The CoordConv-Unet is trained with two prior-based losses: the size loss [19] and the skeleton loss [33].

The **size loss** is a penalty loss function that integrates size information by computing the mean squared error between the grouping of pixel probabilities indicating predicted organ size and a predefined upper or lower bound indicating ground-truth size. The significance of the loss lies in its ability to impose some constraint on the size of the predicted segments.

The **clDice** loss exploits the topological notion of skeletonization in order to reveal subtle topological properties, such as the shape and connected components of anatomical objects within the dataset. Skeletonization is the process of obtaining compact representations of images and objects while still preserving topological properties. The aim of the skeletonization is to extract a region-based shape feature representing the general form of an object.

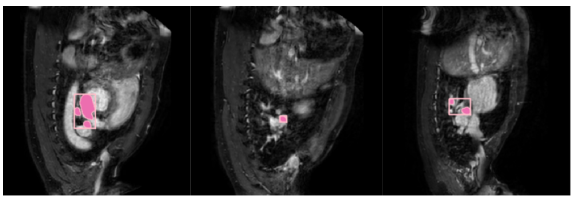


Figure 2: Atrium dataset from Decathlon challenge with multiple components of variable size that are in close proximity of each other

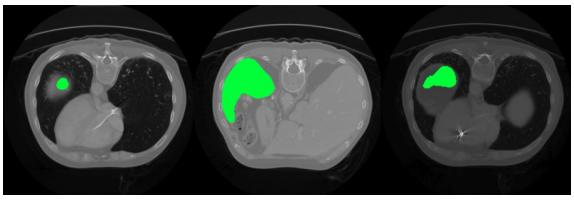


Figure 3: Spleen dataset from Decathlon challenge showing organ or large size variability and convexity issues at boundary level

Table 1: Dataset Description: Patient Dist.: Patient distribution. Org. Size: percentage of pixels occupied by the anatomical object relative to the entire image. Mod.: number of dataset modality. Con. Comp: number of connected components.

	Patient Dist.		Org. Size	Class	Mod.	Con.
	Train	Test	(% of image)			Comp
Atrium	16	4	≤ 1.42	1	1	$0 \sim 4$
Spleen	32	9	≤ 4.36	1	1	$0 \sim 1$

3.5. Comparison to the state of the art

The CoordConv concept in essence is not a new topic proposed in the paper. Introduced by [35], the CoordConv layer was initially designed to investigated supervised Coordinate classification, object detection, supervised coordinate regression, and generative adversarial modeling. The method in [35] raises questions with regards to CoordConv's ability to make the training more stable given the coordinate classification task. However, implementations in [35] did not address the segmentation problem, which is quite important in current research. To add to this, the method includes adding the CoordConv layer solely at the primary convolution of the entire architecture

Since segmentation is a classification problem done at pixel-level, investigating the efficiency of CoordConv in aiding prior-constrained medical

segmentation is an interesting research direction to be explored. In this context, works within the state-of-the-art most relevant to our method are [39] and [40]. In [39], the authors investigate the role of CoordConv in conducting 3D segmentation of pulmonary lobes via a V-Net. Instead of replacing the encoder layers of the network as our proposed CoordConv-Unet model, authors of [39], perform this interchange at the level of the up-sampling layers. Thus, the method first exploits a 2D automated lung segmentation model followed by the CoordConv embedded V-Net architecture. On the other hand, authors of [40] propose a 3-stage framework in order conduct brain mid-line delineation. Within the segmentation step, they introduce CoordConv component at the input level of the network at the intermediate segmentation step, midway between alignment (first step) and delineation (third step).

Whereas [40] integrates CoordConv solely at the input level (as is originally implemented in [35]), our proposed CoordConv-Unet in this paper interchanges the first convolution of each stage of the U-Net architecture with that of the CoordConv layer. Moreover, we do not conduct pre- or post processing steps as in [40]. Unlike [39], where CoordConv component is incorporated at the level of the upsampling layers, the proposed method do so at the encoder convolution level.

4. Experimental Setting

4.1. Datasets

Experiments are conducted on two well-known public datasets from the Decathlon challenge for medical image segmentation [2]. Namely, a cardiac dataset and a spleen dataset. In Table 1, a brief summary of data characteristics is specified.

Atrium dataset is a mono-modal MRI cardiac dataset dedicated to segmentation of the left atrium. The heart as shown in Figure 2 is a multi-connected component object with up to 4 elements of varying sizes and lying in close proximity to each other. The dataset is also characterized by a huge class imbalance with respect to background and foreground pixel distribution.

Spleen dataset is a CT dataset as presented in Figure 3. The objective is to segment a single organ (the spleen), characterized with largely varying size and mild convexity issues at boundary levels.

For pre-processing, we have resized the images to a size of 256×256 and normalized them to a pixel value between 0 and 1. Deploying the framework presented by [30], we have kept negative samples for training. Negative samples are empty images, meaning that the organ of interest is not present. The datasets were split into train and validation based on an 80 % , 20 % partition respectively. Cross-validation was done on three folds of the data based on three Monte-Carlo simulations [52].

[2] http://medicaldecathlon.com/

4.2. Model architecture and training

To insure reproducibility, we deploy a well-known experimental framework presented by [30]. The U-Net [15], of which we integrate the CoordConv layers onto, is a 3-stage U-Net composed of convolutional and de-convolutional blocks, bottleneck and skip connections. Each stage within the encoder is composed of convolutional blocks containing an ensemble of convolutional and batch normalization layers. On the other hand, each stage within the decoder path is composed of 2 consecutive convolutional blocks followed by an upsampling layer. The bottleneck is constituted of 2 convolutional blocks separated by a residual block [17].

Since prior losses essentially suffer from training instability and local solutions [30], we have trained both size loss [19] and clDice loss [33] in conjunction with the Dice baseline loss weighted by a hyper-parameter α, where α is dynamically updated through training according to the following equation

$$L = (1 - \alpha)L_{Dice} + \alpha.L_{prior}$$

Thus, starting from a value of $\alpha = 0.01$, α is increased by a value of 0.01 at each training epoch.

Models were evaluated using the Dice index and Hausdorff distance. Training was conducted via the Adam optimizer with a batch size of 8 over 200 epochs. The learning rate was set to 5×10^{-4} and halved each 20 epochs if the validation performance did not improve.

5. Investigation of CoordConv-Unet performance relative to the Datasets

In this section, we present results for the CoordConv-Unet model when trained on the atrium and the spleen datasets via just the Dice loss, the Size loss [19] and the clDice loss [33]. We note that the Atrium is characterized by multi-connected small components that are very close to each other, whereas the spleen is a complex-shape organ of non-convex curves and edges. Results obtained on Dice accuracy and Hausdorff distance are compared in reference to the regular U-Net baseline trained under the considered losses as well as the CoordConv(+1), which represents the addition of the CoordConv layer solely at the input stage as originally designed by [35].

5.1. Results on the Spleen dataset

Results on Dice accuracy and Hausdorff metric for the spleen as shown in Table 2 and Table 3 indicate the significance of CoordConv layers

Table 2: Dice accuracy results on the spleen dataset

	U-Net	CoordConv(+1)	CoordConv-Unet
Dice Loss	78.58±5.46	64.65±5.7	65.48±8.48
Size Loss	86.44±15.87	**94.86±1.72**	**94.96±1.59**
clDice loss	87.15±13.61	87.04±9.98	**94.54±1.06**

Table 3: Hausdorff distance results on the spleen dataset

	U-Net	CoordConv(+1)	CoordConv-Unet
Dice Loss	1.30 ± 0.24	1.89±0.26	1.71±0.32
Size Loss	1.02 ± 0.56	**0.76±0.13**	**0.76±0.12**
clDice loss	1.07 ± 0.53	1.11±0.31	**0.85±0.07**

when training via prior-based losses. Thus, CoordConv(+1) and CoordConv-Unet increase (case of size loss) or maintain (case of CoordConv(+1) with clDice loss) system performance relative to the regular U-Net baseline. CoordConv-Unet increases the Dice accuracy by over 8% from the regular U-Net baseline under the prior losses. This added value of CoordConv is further verified by the error computed on Hausdorff distance. Thus, the CoordConv-Unet results in a 28% decrease in the Hausdorff metric (from $1.02 to 0.76$) under size loss and a 20% decrease in Hausdorff distance (from $1.07 to 0.85$) under clDice loss relative to the U-Net and CoordConv-Unet respectively. This mainly indicates the ability of CoordConv-Unet to learn curvature features relative to the inter-distance position of the pixels relative to the spleen.

5.2. Results on the Atrium dataset

For the Atrium, we benchmark the results in Table 4 and Table 5. A closer look at the tables, one can realize the significance of the clDice loss generally against the size as well as the Dice baseline loss. We predict that, since the clDice is generally based on the skeleton concept, the clDice was hence able to distinguish between the different boundaries of the connected components relative to the heart. Moreover, training CoordConv-Unet under the clDice loss has increased model performance by about 3% on atrium dataset from baseline training. This indicates the proper role of CoordConv-Unet in increasing system performance under topological losses. This is also verified by the Hausdorff distance metric where CoordConv-Unet under clDice scored second best relative to the other frameworks.

Comparing relative to the size loss, CoordConv-Unet maintains system performance considerably over all its paradigms. From here, we can realize the dual role that the CoordConv plays in enabling learning of spatial dependent attributes when needed (case of the clDice) or mimicking typical convolutional functioning. Given the latter, one would anticipate that the behavior of the CoordConv would be typical to that of a regular convolution. However, a closer look at the evolution of the Dice accuracy over the number of training epochs given folds where CoordConv solution equates that of a regular U-Net from Figure 4, one would realize yet another significant advantage of using CoordConv-Unet in place of regular U-Net. Thus, CoordConv-Unet insures model stability and convergence by evading the undershoot evident when training with regular U-Net.

5.3. Analysis

Based on the results presented on both the spleen and atrium datasets, one can gather that CoordConv-Unet plays a dual role given prior constraint neural networks. Thus, CoordConv can either maintain segmentation performance while regularizing training and stabilizing evolution, or it increases system performance by evading the local solutions that prior losses suffer from.

Despite the significance of CoordConv-Unet as shown by the above study, however, limitations still persist with regards to CoordConv-Unet performance. Thus, when tested against the Dice loss baseline alone (first row of each table), the addition of the CoordConv components degrades the system performance considerably. One could cast the clarification of the phenomena to the complexity/regularizing trade-off relative to CoordConv-Unet. We have clarified earlier 2 main roles of CoordConv:1) the role of stabilize the undershoot evident when training against prior losses. 2) the role of enhancing segmentation performance by evading local solutions. Addressing the two attributes relative the Dice baseline training, we can gather the following: The undershoot revealed in the plotted diagrams are a result of the interchange between the main pixel-wise Dice loss on one hand and the anatomical prior loss on the

Table 4: Dice accuracy on the atrium dataset

	U-Net	CoordConv(+1)	CoordConv-Unet
Dice Loss	83.67 ± 3.66	82.10±2.51	82.52 ± 2.33
Size Loss	84.59±2.62	84.63±1.67	84.48±1.5
clDice loss	83.85 ± 2.56	85.35 ± 1.65	**86.15±1.39**

Table 5: Hausdorff distance results on the atrium dataset

	U-Net	CoordConv(+1)	CoordConv-Unet
Dice Loss	1.62 ± 0.16	1.64±0.11	1.64±0.11
Size Loss	1.59±0.17	1.57±0.08	**1.57±0.10**
clDice loss	1.64 ± 0.16	1.60±0.14	1.59±0.14

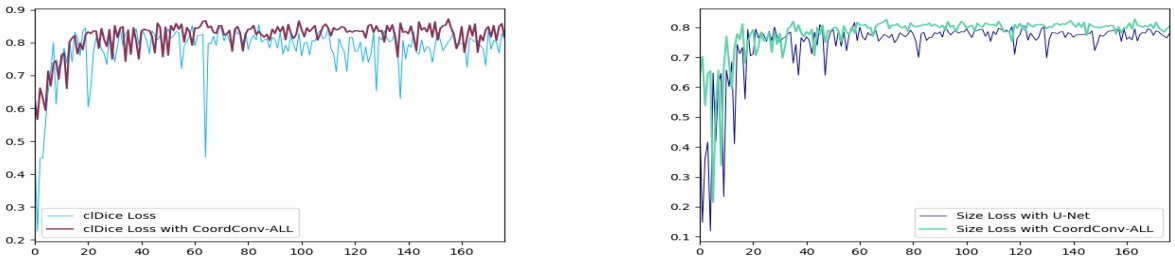


Figure 4: Evolution of the Dice accuracy in validation under Dice +clDice (LEFT), Dice +Size (RIGHT) for the atrium dataset.

other in prior-constrained training problem. In the Dice baseline, the undershoot is rather non-existent; hence CoordConv cannot play the role of the regularizer that evades training in stability and ends up decreasing model performance by adding complexity to the system. The second role of CoordConv is that it increases model performance under prior-based losses by evading local solutions. As we have previously said, prior-based losses are interestingly used because they integrate expert knowledge onto the automatic training. However, designing these loss functions is often tedious and subjected to various differentiability and stability challenges. One of the reasons for these issues is the discrete nature of these prior information vs the continuous soft probability output of the network. With CoordConv-Unet, the network can act as a regular U-Net if needed, i.e., if no stability problems persist or the network could integrate spatial knowledge thus fulfilling the true CoordConv concept. In regular Dice baseline training, prior does not exist, hence, addressing stability is not an issue. From all of the above, we can hence realize the insignificance of using CoordConv component if there is no prior constraint problem involved.

6. Conclusion

In this paper, we have proposed a new model, the CoordConv-Unet model as a proficient substitute to U-Net given prior-constraint tasks. We have exposed the dual role of CoordConv-Unet given these constrained tasks. Thus, CoordConv-Unet can either improve performance by evading stability problems, or can mimic a regular U-Net if the translation invariance attribute is required. In the later role, CoordConv-Unet still poses a further significance that can help stabilize the network performance under the prior-based losses.

Future work includes designing certain frameworks that can impose whether the weights connecting the coordinate layers with the convolutional ones to be trained or fixed so as to resolve the problem faced when training CoordConv-Unet under a regular unconstrained Dice loss. Moreover, efficiency of CoordConv-Unet given multi-organ and lesion segmentation is also to be explored.

Acknowledgments

The authors would like to acknowledge the National Council for Scientific Research of Lebanon (CNRS-L) and the Agence Française de la Francophonie (AUF) for granting a doctoral fellowship to Rosana El Jurdi, as well as the ANR (project APi,

grant ANR-18-CE23-0014). This work is part of the DAISI project, co-financed by the European Union with the European Regional Development Fund (ERDF) and by the Normandy Region.

References

[1] C. Xu, D. L. Pham, and J. L. Prince, "Image segmentation using deformable models," in Handbook of medical imaging. SPIE, 2000, vol. 2, no. 3, pp. 129–174.

[2] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 8, pp. 888–905, 2000.

[3] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient nd image segmentation," International journal of computer vision, vol. 70, no. 2, pp. 109–131, 2006.

[4] M. S. Nosrati and G. Hamarneh, "Incorporating prior knowledge in medical image segmentation: a survey," CoRR, vol. abs/1607.01092, 2016. [Online]. Available: http://arxiv.org/abs/1607.01092

[5] S. Vicente, V. Kolmogorov, and C. Rother, "Graph cut based image segmentation with connectivity priors," in IEEE Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[6] I. B. Ayed, S. Li, A. Islam, G. Garvin, and R. Chhem, "Area prior constrained level set evolution for medical image segmentation," in Medical Imaging 2008: Image Processing, vol. 6914. International Society for Optics and Photonics, 2008, p. 691402.

[7] A. Foulonneau, P. Charbonnier, and F. Heitz, "Multi-reference shape priors for active contours," International journal of computer vision, vol. 81, no. 1, p. 68, 2009.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in CVPR, June 2015, pp. 3431–3440.

[9] X. Zhou, T. Ito, R. Takayama, S. Wang, T. Hara, and H. Fujita, "Three-dimensional ct image segmentation by combining 2d fully convolutional network with 3d majority voting," in Deep Learning and Data Labeling for Medical Applications, G. Carneiro, D. Mateus, L. Peter, A. Bradley, J. M. R. S. Tavares, V. Belagiannis, J. P. Papa, J. C. Nascimento, M. Loog, Z. Lu, J. S. Cardoso, and J. Cornebise, Eds., 2016, pp. 111–120.

[10] G. Li and Y. Wan, "Adaptive seeded region growing for image segmentation based on edge detection, texture extraction and cloud model," in Information Computing and Applications, R. Zhu, Y. Zhang, B. Liu, and C. Liu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 285–292.

[11] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in 2016 Fourth International Conference on 3D Vision (3DV), Oct 2016, pp. 239–248.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.

[13] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[14] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in MICCAI, 2017, pp. 203–211.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, 2015, pp. 234–241.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.

[17] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 5, pp. 749–753, 2018.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in CVPR, 2015, pp. 1–9. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298594

[19] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. B. Ayed, "Constrained-cnn losses for weakly supervised segmentation," Medical Image Analysis, vol. 54, pp. 88 – 99, 2019.

[20] R. El Jurdi, C. Petitjean, P. Honeine, and F. Abdallah, "Bb-unet: U-net with bounding box prior," IEEE Journal of Selected Topics in Signal Processing, pp. 1–1, 2020.

[21] O. Oktay, J. Schlemper, L. L. Folgoc, M. C. H. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," in Medical Imaging with Deep Learning, 2018.

[22] F. Caliva, C. Iriondo, A. M. Martinez, S. Majumdar, and V. Pedoia, "Distance map loss penalty term for semantic segmentation," in International Conference on Medical Imaging with Deep Learning – Extended Abstract Track, London, United Kingdom, 08–10 Jul 2019. [Online]. Available: https://openreview.net/forum?id=B1eIcvS45V

[23] J. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. King, "A topological loss function for deep-learning based image segmentation using persistent homology," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2020.

[24] H. Oda, H. R. Roth, K. Chiba, J. Sokolić, T. Kitasaka, M. Oda, A. Hinoki, H. Uchida, J. A. Schnabel, and K. Mori, "Besnet: Boundary-enhanced segmentation of cells in histopathological images," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 228–236.

[25] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. de Marvao, T. Dawes, D. P. O'Regan, B. Kainz, B. Glocker, and D. Rueckert, "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation," IEEE Transactions on Medical Imaging, vol. 37, no. 2, pp. 384–395, Feb. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8051114/

[26] R. El Jurdi, C. Petitjean, P. Honeine, and F. Abdallah, "Towards semi-supervised segmentation of organs at risk using deep convolutional neural networks," in GDR-ISIS workshop: Apprentissage faiblement supervisé ou nonsupervisé pour l'analyse d'images et de video, Paris, France, May 2019.

[27] R. El Jurdi, T. Dargent, C. Petitjean, P. Honeine, and F. Abdallah, "Investigating coordconv for fully and weakly supervised medical image segmentation," in Tenth International Conference on Image Processing Theory, Tools and Applications, IPTA 2020, Paris, France, 2020Submitted.

[28] F. Caliva, C. Iriondo, A. M. Martinez, S. Majumdar, and V. Pedoia, "Distance map loss penalty term for semantic segmentation," in International Conference on Medical Imaging with Deep Learning, London, UK, 2019.

[29] R. Trullo, C. Petitjean, S. Ruan, B. Dubray, D. Nie,

and D. Shen, "Joint segmentation of multiple thoracic organs in CT images with two collaborative deep architectures," MICCAI'17 workshop Deep Learning in Medical Image Analysis, 2017.

[30] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. Ben Ayed, "Boundary loss for highly unbalanced segmentation," in Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning, ser. Proceedings of Machine Learning Research, M. J. Cardoso, A. Feragen, B. Glocker, E. Konukoglu, I. Oguz, G. Unal, and T. Vercauteren, Eds., vol. 102. London, United Kingdom: PMLR, 08–10 Jul 2019, pp. 285–296.

[31] A. Arif, S. M. M. Rahman, K. Knapp, and G. Slabaugh, "Shape-aware deep convolutional neural network for vertebrae segmentation," in Computational Methods and Clinical Applications in Musculoskeletal Imaging, 2018, pp. 12–24.

[32] J. Dolz, I. Ben Ayed, and C. Desrosiers, "Unbiased shape compactness for segmentation," in Medical Image Computing and Computer Assisted Intervention MICCAI 2017, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, pp. 755–763.

[33] S. Shit, J. C. Paetzold, A. Sekuboyina, A. Zhylka, I. Ezhov, A. Unger, J. P. W. Pluim, G. Tetteh, and B. H. Menze, "cldice - a topology-preserving loss function for tubular structure segmentation," ArXiv, vol. abs/2003.07311, 2020.

[34] Z. Mirikharaji and G. Hamarneh, "Star shape prior in fully convolutional networks for skin lesion segmentation," in MICCAI, ser. Lecture Notes in Computer Science, vol. 11073. Springer, 2018, pp. 737–745.

[35] R. Liu, J. Lehman, P. Molino, F. Petroski Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," in Advances in Neural Information Processing Systems 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 9605–9616.

[36] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end Training of Deep Visuomotor Policies," J. Mach. Learn. Res., vol. 17, no. 1, pp. 1334–1373, Jan. 2016. [Online]. Available: http://dl.acm.org/citation.cfm?id=2946645.2946684

[37] H. Qi, S. Collins, and J. A. Noble, "UPI-Net: Semantic Contour Detection in Placental Ultrasound," in Visual Recognition for Medical Images (VRMI), ICCV 2019 workshop, Sep. 2019, arXiv: 1909.00229.

[38] X. Yao, H. Yang, Y. Wu, P. Wu, B. Wang, X. Zhou, and S. Wang, "Land Use Classification of the Deep Convolutional Neural Network Method Reducing the Loss of Spatial Features," Sensors, vol. 19, no. 12, p. 2792, Jan. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/12/2792

[39] W. Wang, J. Chen, J. Zhao, Y. Chi, X. Xie, L. Zhang, and X. Hua, "Automated segmentation of pulmonary lobes using coordination-guided deep neural networks," in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 2019, pp. 1353–1357.

[40] S. Wang, K. Liang, C. Pan, C. Ye, X. Li, F. Liu, Y. Yu, and Y. Wang, "Segmentation-based method combined with dynamic programming for brain midline delineation," in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 772–776.

[41] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. Van Uden, C. Sanchez, G. Litjens, F.-E. Leeuw, B. Ginneken, E. Marchiori, and B. Platel, "Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities," Scientific Reports, vol. 7, 10 2016.

[42] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in CVPR. Los Alamitos, CA, USA: IEEE Computer Society, jul 2017, pp. 1665–1674.

[43] R. El Jurdi, C. Petitjean, P. Honeine, and F. Abdallah, "Organ segmentation in ct images with weak annotations: A preliminary study," in 27th GRETSI Symposium on Signal and Image Processing, Lille, France, Aug. 2019.

[44] A. Mosinska, P. Márquez-Neila, M. Kozinski, and P. Fua, "Beyond the pixel-wise loss for topology-aware delineation," in CVPR, 2018, pp. 3136–3145. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Mosinska_Beyond_the_Pixel-Wise_CVPR_2018_paper.html

[45] S. Yang, J. Kweon, and Y.-H. Kim, "Major vessel segmentation on x-ray coronary angiography using deep networks with a novel penalty loss function," in International Conference on Medical Imaging with Deep Learning – Extended Abstract Track, London, United Kingdom, 08–10 Jul 2019. [Online]. Available: https://openreview.net/forum?id=H1lTh8unKN

[46] J. Peng, H. Kervadec, J. Dolz, I. B. Ayed, M. Pedersoli, and C. Desrosiers, "Discretely-constrained deep network for weakly supervised segmentation," Neural networks : the official journal of the International Neural Network Society, vol. 130, pp. 297–308, 2020.

[47] F. Ségonne and B. Fischl, Integration of Topological Constraints in Medical Image Segmentation. Boston, MA: Springer US, 2015, pp. 245–262. [Online]. Available: https://doi.org/10.1007/978-0-387-09749-7_13

[48] X. Hu, F. Li, D. Samaras, and C. Chen, "Topology-preserving deep image segmentation," in Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 5657–5668. [Online]. Available: http://papers.nips.cc/paper/8803-topology-preserving-deep-image-segmentation.pdf

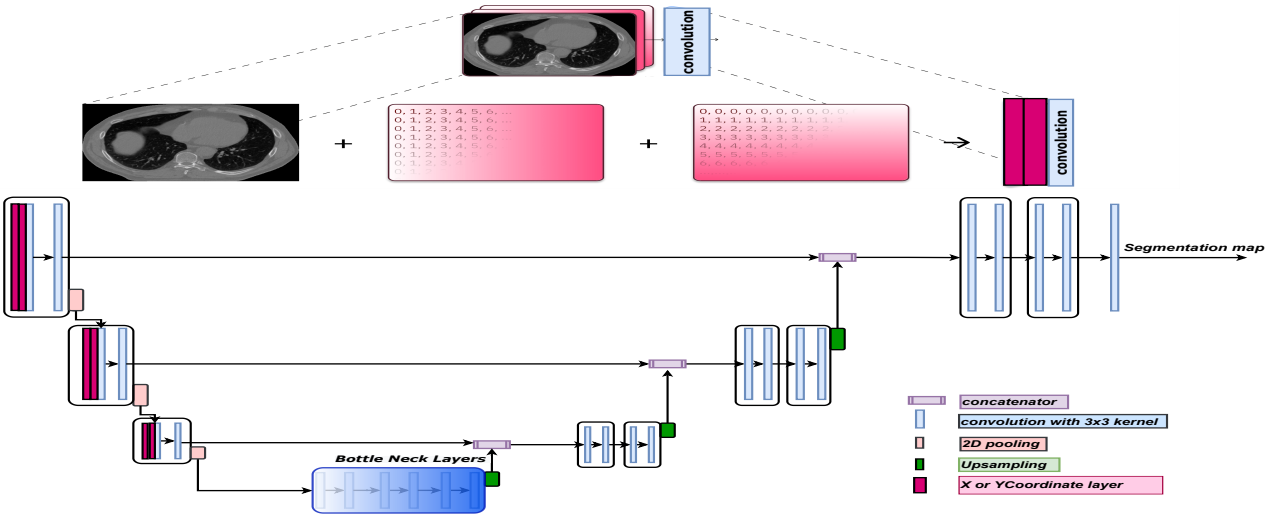
[49] P.-A. Ganaye, M. Sdika, B. Triggs, and H. Benoit-Cattin, "Removing segmentation inconsistencies with semi-supervised non-adjacency constraint," Medical Image Analysis, vol. 58, p. 101551, Dec. 2019.

[50] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. de Marvao, T. Dawes, D. P. O'Regan, B. Kainz, B. Glocker, and D. Rueckert, "Anatomically constrained neural networks : Application to cardiac image enhancement and segmentation," IEEE Transactions on Medical Imaging, vol. 37, no. 2, pp. 384–395, Feb 2018.

[51] C. Zotti, Z. Luo, O. Humbert, A. Lalande, and P. Jodoin, "Gridnet with automatic shape prior registration for automatic MRI cardiac segmentation," in Statistical Atlases and Computational Models of the Heart STACOM, Held in Conjunction with MICCAI, Quebec City, Canada, ser. LNCS, vol. 10663, 2017, pp. 73–81.

[52] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," Statistics Surveys, vol. 4, no. none, pp. 40 – 79, 2010. [Online]. Available: https://doi.org/10.1214/09-SS054

Abstract



Proposed CoordConv-Unet model. In the top panel, the CoordConv layer consists in concatenating the x-layer and y-layer to the convolutional layer. CoordConv-Unet consists of replacing the first convolutional layer of each stage with the CoordConv layer.