3D Spatial Priors for Semi-Supervised Organ Segmentation with Deep Convolutional Neural Networks

Nicolas Thome, Luc Soler, Olivier Petit

Noname manuscript No.
(will be inserted by the editor)

3D Spatial Priors for Semi-Supervised Organ Segmentation with Deep Convolutional Neural Networks

Olivier Petit · Nicolas Thome · Luc Soler

Received: date / Accepted: date

Abstract
Purpose: Fully Convolutional neural Networks (FCNs) are the most popular models for medical image segmentation. However, they do not explicitly integrate spatial organ positions, which can be crucial for proper labeling in challenging contexts.

Methods: In this work, we propose a method that combines a model representing prior probabilities of an organ position in 3D with visual FCN predictions by means of a generalized prior-driven prediction function. The prior is also used in a self-labeling process to handle low-data regimes, in order to improve the quality of the pseudo-label selection.

Results: Experiments carried out on CT scans from the public TCIA pancreas segmentation dataset reveal that the resulting STIPPLE model can significantly increase performances compared to the FCN baseline, especially with few training images. We also show that STIPPLE outperforms state-of-the-art semi-supervised segmentation methods by leveraging the spatial prior information.

Conclusion: STIPPLE provides a segmentation method effective with few labeled examples, which is crucial in the medical domain. It offers an intuitive way to incorporate absolute position information by mimicking expert annotators.

O. Petit
Visible Patient, Strasbourg, France
CEDRIC Lab, Conservatoire National des Arts et Metiers, Paris, France
E-mail: olivier.petit@visiblepatient.com

N. Thome
CEDRIC Lab, Conservatoire National des Arts et Metiers, Paris, France

L. Soler
Visible Patient, Strasbourg, France

Keywords Deep Learning · Medical Image Segmentation · 3D Spatial Prior · Semi-Supervised Learning · Pseudo-Labeling

1 Introduction

Organ segmentation in medical images is a challenging but important task for many clinical applications like computer-aided diagnosis. It is a powerful tool for intervention planning and other computer-assisted applications.

In the last few years, deep learning and Convolutional Neural Networks (ConvNets) [8] achieved a breakthrough in visual recognition. In semantic segmentation, Fully Convolutional Networks (FCNs) [10, 2, 17] achieve state-of-the-art performance, while being computationally efficient. In medical image segmentation, the most common architectures include an encoder-decoder network with the recovery of absolute position information, *e.g.* skip connections [17, 11, 3]. Despite the huge performance gain brought by deep learning and modern FCN, medical image segmentation remains a very challenging task, due to low contrasts between organs, and visual ambiguities. In many cases, the local visual context of an image is insufficient to perform a clear decision and external knowledge is required.

In this paper, we tackle the problem of including prior knowledge about the spatial position of organs to improve the quality of the segmentation. It is a particularly strong and relevant prior for medical images since there are some conventions on how the image should be, *e.g.* the position of a patient. Using prior knowledge is common for practitioners, which perform segmentation not only by using the visual appearance of medical images, but also leverage their strong knowledge on the position of organs or relative layout between them.

We introduce STIPPLE, a method that incorporates SpaTIal Priors and Pseudo LabEls. The spatial prior is a probability map of the organ presence at a given position. This map is merged with the visual information extracted by the FCN through a prior-driven prediction function. We also propose a semi-supervised extension of our model with an iterative self-labeling process. It forms a virtuous circle where the 3D prior is leveraged for selecting relevant pseudo-labels, leading to refined interactions between visual and prior predictions.

We perform experiments on a pancreas segmentation dataset and show that our method outperforms the performances of other state-of-the-art approaches for both semi-supervision and integration of position information.

The main contributions of this paper are as follow:

- We introduce STIPPLE, a 3D spatial prior that explicitly incorporates knowledge in a deep FCN for medical image segmentation. The prior is added in the final activation function via a prior-driven softmax.
- We show the relevance of such a prior in a fully-supervised setting and how it could be leveraged for semi-supervised within a pseudo-labeling scheme. For the latter, our prior helps to select new labels by limiting the

incorporation of wrong predictions, especially outliers that could ruin the training.
– Experiments show that our prior is particularly powerful when very few labels are available. Moreover, compared to other state-of-the art methods, STIPPLE shows better results for every proportion of missing labels.

2 Related Work

Including absolute position information to bias a FCN is not straightforward in semantic segmentation. FCNs are by design equivariant to small transformations and thus unable to directly encode spatial location information to bias predictions as shown in [9]. The authors show that FCNs are unable to model a coordinate transform task. Then, they show that adding the absolute coordinated of the pixels in a feature map could fix this issue. However the CoordConv layer is added in the first layer contrary to STIPPLE which explicitly integrates the absolute position information by biasing the visual prediction.

Locally Connected Networks (LCNs) are able to model absolute position information. LCNs learn prediction models specific to each spatial position, and have been successfully applied to face recognition, e.g. DeepFace [19]. However, LCNs significantly increase the number of parameters of the model (compared to their convolutional counterparts), and thus require huge labeled datasets to be robust to overfitting. LCNs are consequently not adapted for medical image segmentation where only few labeled data are available.

In the medical image analysis literature, cascaded networks [18, 7] include absolute position information by relying on the selection of a Region of Interest (RoI) by a first model, which is subsequently refined by a second one which performs a more accurate segmentation. Although these approaches are efficient, they are intrinsically limited by the quality of the first RoI selection step. Some works simply take cropped images of the expected RoI [5, 14] which is in fact a very strong prior about the organ position. However, it does not use the whole image and is very limited to the selected region. Thus, each class should be learned independently [5] which drastically increases the model complexity and computational burden.

Other methods try to incorporate spatial prior information by biasing the learning of internal deep representations in an implicit manner [20, 4, 13]. In the same way, attention mechanisms have gained popularity in the last few years. New parameters bias the intermediate representations to focus on a specific region of an image. For example, in [14] the method integrates an additive attention block in the decoder part of a U-Net model. The attention coefficients are learned during training and are completely implicit. Thus, we cannot assure that the model actually learns a prior on the spatial position. Moreover, despite the reasonable improvements shown by these methods in fully-supervised settings, they are intrinsically limited to 2D absolute position information, which may arguably be inaccurate for organ segmentation with a complex shape varying in 3D. In STIPPLE, we use a spatial prior that captures

the complete organ shape in 3D and explicitly bias the visual prediction to leverage the depth information.

Medical image analysis often faces the problem of limited amounts of labeled data. Semi-supervised methods allow training models on a large dataset of unlabeled images. There are three main categories of methods: using adversarial training, consistency and pseudo-labeling.

In adversarial training, a model in trained to fool a discriminator that is trained to distinguish true and generated examples. In [12], the authors use the strategy of [6] which consists in building a generator which produces a segmentation of an input image. Then, the discriminator takes the segmentation map and produces a confidence map which is used to select pixels that could be used in the segmentation loss. This work is further improved in [23] which uses a 3D deep atlas prior to weight the pixels in the loss function with a focal loss. This method is very different from ours, the prior is used to weight examples based on their difficulty through the focal loss and is not directly integrated to the network.

The consistency approaches [22], e.g. mean teacher, are purely designed for semi-supervision. The main idea is to train two similar models in parallel: a student network which is trained directly on the labeled data and a teacher model which is trained by using the moving average of the student weights. On top of that, a consistency loss leverages the fact that the same input under different transformations or noises should give the same result. This loss could be computed both with labeled data and unlabeled data.

Finally pseudo-labeling is a large category of methods which aims to assign labels to unlabeled examples before fine-tuning or training a new model. Those methods are state-of-the-art in semi-supervised learning. For example, in [1], all the unlabeled images are pseudo-labeled and added to the train set. However it could add too many wrong predictions.

STIPPLE follows state-of-the-art pseudo-labeling methods for semi-supervised segmentation, and leverages the proposed spatial prior to improve the automatic selection of pseudo-labels. We also use an iterative approach which sequentially adds more pseudo-labels and retrains the model from the augmented training set.

3 Organ segmentation with 3D spatial priors and pseudo-labeling

In this section, we introduce our STIPPLE model dedicated to leverage spatial priors and pseudo-labeling for semantic segmentation of medical images. The overall prediction model of STIPPLE is depicted in Figure 1.

A given input volume $\mathbf{V}$ is processed by the backbone FCN segmentation model which outputs a probability prediction volume $\mathbf{S} = \{s_k\}_{k\in\{1;K\}}$ where K in the number of classes. Our approach is agnostic to the choice of the FCN: in our experiments we use 2D U-Net [17] due to hardware limitations and for experiment efficiency, but it could easily extend to 3D models [3].

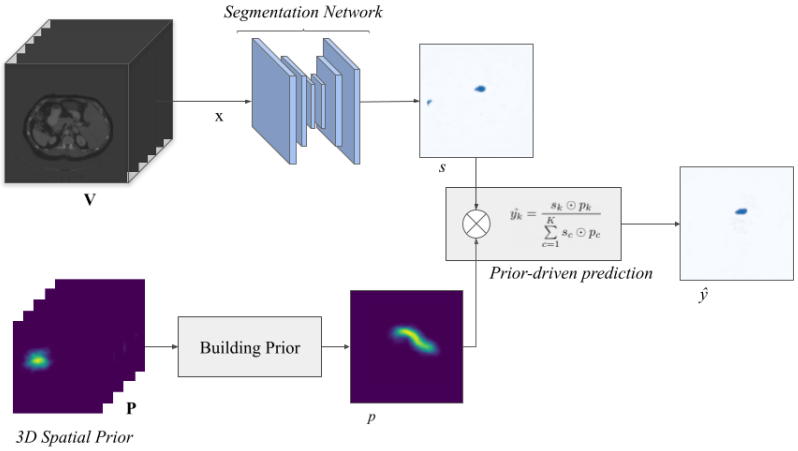


Fig. 1 The input volume $\mathbf{V}$ is sliced along the axial view. The segmentation network outputs a visual prediction $\mathbf{S}$. The 3D spatial prior $\mathbf{P}$ is aligned to the slice before being combined through a prior-driven prediction function. The result is the final prediction $\hat{\mathbf{Y}}$.

Formally, let us consider a volume $\mathbf{V} \in \mathbb{R}^{W \times H \times Z}$ composed of Z axial slices, *i.e.* $\mathbf{V} = \{x_z\}_{z \in \{1;Z\}}$, with $x_z \in \mathbb{R}^{W \times H}$. The semantic segmentation problem consists in predicting a label among K organ classes (including the background) for each voxel of the volume $\mathbf{V}(w,h,z)$[1]. The FCN segmentation network computes posterior probabilities, *e.g.* $s(w,h)_{z,k} = \mathbf{Pr}\left(\mathbf{Y}_{w,h,z} = k \mid N(x(w,h)_z), \mathbf{W}\right)$ for our case with a 2D model, where $\mathbf{W}$ represents the model parameters and $N(x(w,h)_z)$ is the voxel neighborhood in a given slice z, characterized by the FCN receptive field.

As previously mentioned, the computation of $s(w,h)_{z,k}$ doesn't incorporate any absolute position information. We propose to define a 3D spatial prior $\mathbf{P}$ which represents the probability of an organ presence given its 3D position. The final prediction of STIPPLE $\hat{\mathbf{Y}}$ consists in merging $\mathbf{P}$ and $\mathbf{S}$, as described in section 3.2.

3.1 3D spatial prior design and computation

To overcome the lack of absolute position information encoded in our FCN predictions $s(w,h)_{z,k} = \mathbf{Pr}\left(\mathbf{Y}_{w,h,z} = k \mid N(x(w,h)_z), \mathbf{W}\right)$, we propose to model the prior probabilities of the organ position, *i.e.* with $\mathbf{P} = \{p_k\}_{k \in \{1;K\}}$, $p(w,h)_{z,k} = \mathbf{Pr}\left(\mathbf{Y}_{w,h,z} = k \mid (w,h,z)\right)$, independently of the visual input $N(x(w,h)_z)$ and model parameters $\mathbf{W}$.

[1] Here we choose to designate the coordinates with (w,h,z) so it is a different notation than the model's output and input, x and y.

The construction of the proposed 3D spatial prior is based on the following assumptions: (1) the 3D volumes are given in the axial direction (z), with the patient lying on the back ; (2) In the axial (z) direction, there might be strong variations in the organ position, i.e. the $[z_{min}; z_{max}]$ interval where the organ is visible might significantly change. On the other hand, the variability in the (w,h) plane for a given z value is supposed to be much smaller, such that we can accumulate the organ positions in this plane across the dataset to obtain relevant statistics of organ position.

Note that these assumptions are valid in many clinical cases, since acquisitions in the axial direction are common. Moreover, it is also common for anatomical structures to be visible in variable $[z_{min}; z_{max}]$ values in the z direction because of differences in acquisition procedures.

Our prior $\mathbf{P}$ is estimated on a training dataset of labeled organs $\{\mathbf{Y}_i\}_{i\in\{1;N\}}$ where N is the number of examples, by computing statistics of the organ presence in a 3D rectangular volume of size $(W_p \times H_p \times \Delta_z)$ with W_p, H_p and Δ_z being respectively the width, the height and depth of the rectangular volume. This size is determined by taking the maximum width, height and depth of the considered organ in the training set such that every example fits into it. We observed that the position of the organs are relatively stable in the (w, h) coordinates, but may largely vary in the z direction. So we decide to discretize the prior over the z axis such that the prior $\mathbf{P}$ itself is of size $(W_p \times H_p \times B)$; where B bins aggregate the Δ_z slices, with $B < \Delta_z$ to gain invariance with respect to misalignment of organs in the z direction, but $B > 1$ to capture organ shape variations. Eventually, $p(w, h)_{z,k}$ is estimated from the full training dataset by a non-parametric estimation, *i.e.* histogram estimation:

$$p(w,h)_{z,k} = \mathbf{Pr}\left(\mathbf{Y}_{w,h,z} = k \mid (w,h,z)\right) = \frac{1}{Z_{tot}} \sum_{z=1}^{Z_{tot}} \mathbb{1}(\mathbf{Y}_{w,h,z} = k) \qquad (1)$$

where Z_{tot} is the total number of slices in a given bin b.

In practice, the training volumes are first aligned with the center of the organ segmentation masks and then a sub-volume of size $(W_p \times H_p \times \Delta_z)$ is cropped around this center.

The prior computation is illustrated in Algorithm 1. An example of a 3D prior map with $B = 3$ bins is shown in Figure 2. We can see that each bin results in an average of multiple neighboring slices from the input volume. The bin (1) corresponds to the top of the segmentation mask whereas the bin (3) is the bottom of the pancreas. For those two bins the corresponding probabilities are localised in very different regions.

3.2 Prior-driven prediction function

The prior probabilities are introduced through a prior-driven prediction function which explicitly integrates our 3D spatial prior in a late fusion manner.

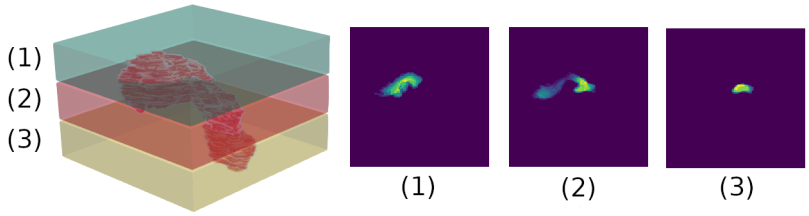


Fig. 2 Prior computation visualisation on one volume with $B = 3$ bins in the z axis

Algorithm 1: Prior construction for a given organ. y_i designates a volume label and N the total number of training volumes. Then B is the number of expected bins for the final prior and W_p, H_p are respectively the prior's Width and Height when Δ_z is the maximum depth observed for the organ in the training set.

Data: $\{(y_i)\}_N, B, W_p, H_p, \Delta_z$
Result: *Prior*
$N \leftarrow$ number of label maps y;
$Prior \leftarrow \text{zeros}(w, h, B)$;
for $i \leftarrow 1$ *to* N **do**
  $c_w, c_h, c_z \leftarrow get\_organ\_center(y_i)$;
  $w_{min} \leftarrow c_w - W_p/2$;
  $w_{max} \leftarrow c_w + W_p/2$;
  $h_{min} \leftarrow c_h - H_p/2$;
  $h_{max} \leftarrow c_h + H_p/2$;
  $z_{min} \leftarrow c_z - \Delta_z/2$;
  $z_{max} \leftarrow c_z + \Delta_z/2$;
  for $s = z_{min}$ *to* z_{max} **do**
    $idx\_in\_prior \leftarrow \frac{B \times (s - z_{min})}{z_{max} - z_{min}}$;
    $Prior[:, :, idx\_in\_prior] += y_i[x_{min} : x_{max}, y_{min} : y_{max}, s]$;
$Prior \leftarrow normalize\_bins(Prior) \leftarrow$ normalize the values between 0 and 1 by dividing by the number of slices added in a given bin;

For the sake of clarity we remove the notation of the dependency in (w,h,z). The main intuition which is presented in Figure 1 is to take the visual predictions of the FCN $\mathbf{S} \in \mathbb{R}^{W,H,Z,K}$ where K in the number of classes, so $\mathbf{S} = \{s_k\}_{k \in \{1;K\}}$ and apply a Hadamard product with the prior probabilities $\mathbf{P} = \{p_k\}_{k \in \{1;K\}}$. Then we normalize to rescale the values between 0 and 1.

When combining those operations, the final formulation (Equation 2) is denoted as a "prior-driven softmax", which outputs $\hat{\mathbf{Y}} = \{\hat{y_k}\}_{k \in \{1;K\}}$.

$$\hat{y_k} = \frac{s_k \odot p_k}{\sum_{c=1}^{K} s_c \odot p_c} = \frac{e^{\tilde{s}_k}\ p_k}{\sum_{c=1}^{K} e^{\tilde{s_c}}\ p_c} = \frac{e^{\tilde{s}_k + ln(p_k)}}{\sum_{c=1}^{K} e^{\tilde{s}_c + ln(p_c)}} \tag{2}$$

$\tilde{\mathbf{S}} = \{\tilde{s_k}\}_{k \in \{1;K\}}$ are the values before activation, usually denoted as "logits".

Interestingly, we can notice that our prediction function in Equation 2 is a consistent generalization of the standard softmax, since it reduces to it when the prior is uniformly distributed through the classes, *i.e.* when $p_k = p_c = \frac{1}{K}\ \forall k \in \{1..K\}$.

When the prior $\mathbf{P}$ is not uniform, it can be used to bias the prediction of a given class k based on its visual input $e^{\tilde{s}_k}$, depending on its spatial location. For example, if p_k is close to 1 (resp. 0), the prediction of class k is made close to 1 (resp. 0) whatever the $e^{\tilde{s}_k}$ value. Our prior-driven softmax prediction function in Equation 2 can thus be leveraged to overcome visual ambiguities between organs and the background.

This formulation is obviously applicable in binary segmentation using a sigmoid (σ) as shown in Equation 3. It becomes a "prior-driven sigmoid".

$$\hat{y_k} = \frac{s_k \odot p_k}{s_k \odot p_k + (1 - s_k) \odot (1 - p_k)} = \sigma(\tilde{s}_k - ln(1 - p_k) + ln(p_k)) \quad (3)$$

Positioning the prior in a volume During training, we can use the position of the organ label to position the prior in the image. However, for unlabeled volume and test volumes we need to find the position. We first take the output probabilities of a segmentation network on the target (unlabeled) volume, which gives a first but coarse position of the organ. Then, a reference volume is randomly selected among the labeled volumes in the training set. For that volume, we have a segmentation map and the true position of the considered organ. With that, we compute the KL divergence between the two with different small translations applied to the probabilities obtained on the target volume. We can finally keep the translation that gives the lowest KL divergence value and adjust the position of the organ for the target volume.

3.3 Integration in a semi-supervised context

We propose a semi-supervised extension of our model, dedicated to leverage unlabeled data. We use a self-training strategy based on pseudo-labeling, which recently showed very good performances for medical image segmentation [6, 15, 23]. Pseudo-labeling is a technique which consists in automatically labeling unlabeled examples. State-of-the-art segmentation methods in computer vision and medical imaging for semi-supervised learning use those kinds of methods in addition to other techniques. The selection of the examples is crucial and should be properly performed. In our case, we select the pseudo-labels by taking the most confident pixels. Concretely, we consider that a prediction with a high probability is more certain than another with a lower probability. Then, for a given volume, we select among the predictions of the organ the top-k most confident voxels that will be selected as pseudo-labels. Our STIPPLE method actually provides a "prior-driven uncertainty measure", in the sense that our 3D prior is leveraged to improve the selection of pseudo-labels by using

3D absolute position information. The pseudo-labeling schema is illustrated in Algorithm 2.

Algorithm 2: Iterative Pseudo-labeling strategy used in STIPPLE. (x_i, y_i) is a training example with x_i the image and y_i the ground truth (which could be partially-labeled). γ_t is the number of voxels to relabel at iteration t, T the number of iterations, m_t the model at iteration t.

Data: $\{(x_i, y_i)\}$, γ_{max}, T, m_0
Result: m_T
$y_{i,0} = y_i$;
for $t \leftarrow 1$ *to* T **do**
  $\gamma_t = \frac{t}{T}\gamma_{max}$;
  for i *in* $\{unlabeled\_image\_indices\}$ **do**
    $\hat{y}_i \leftarrow m_t(x_i)$ // Predict image x_i;
    $y^+_{i,t} \leftarrow select\_\gamma\_most\_confident\_predictions(\hat{y}_i, \gamma_t)$ // Select γ_t new target labels from the prediction $\hat{y}_i$ by taking the most confident;
    $y_{i,t} = y_{i,t-1} \cup y^+_{i,t}$ // Augment training set by adding the pseudo-labels;
  $m_t = train(\{(x_i, y_{i,t})\})$ // Re-train model with the augmented dataset

4 Experiments and Results

4.1 Experimental setup

Evaluation dataset. We evaluate our method on the publicly available dataset TCIA [16] for pancreas segmentation in CT-scans. It is composed of 82 CT-scans with manual labels of the pancreas. In all our experiments, we performed 5 fold cross-validation and reported the standard deviation between the folds. For each fold, a different spatial prior is computed.

Implementation Details We carried out experiments in a semi-supervised setting. Thus, we randomly removed labels (uniform sampling without replacement) at a patient level to reach proportions (α) like 70%, 50%, 30% and 10% of labeled volumes in the training set such that the test set remains the same across the experiments. We also report the results for a fully-supervised setting, *i.e.* a label proportion of 100%. In practice we use one step of relabeling for the low proportions from 50% to 10% and two steps at 70%.

The input volumes are preprocessed by clipping the Hounsfield Units (HU) values in the abdominal organ range [-160,300]. Then the values are normalized to have zero mean and unit variance. In all the experiments, we use a backbone 2D U-Net. The models are trained using the Adam optimizer with standard data augmentation techniques, *i.e.* random translations, random rotations.

The spatial prior is estimated with the available training examples only. We choose $B = 5$ for every proportion, and study its impact in section 4.5.

4.2 Pancreas segmentation results

The results on the TCIA pancreas dataset are given in Figure 3. STIPPLE is compared with a U-Net baseline for every proportion. In each case, our method shows significant gains which are validated with a paired t-test, see Table 1. At a label proportion of 100%, we see an improvement of +1.4pts, at 70%: +4.0pts, at 50%: +3.7pts, at 30%: +5.9pts and finally at 10%: +9.9pts. The gains are more pronounced when the proportion α is low. It is validated by the $p-$values shown in Table 1. The gains increase and the $p-$values decrease when α decreases.

Table 1 $p-$values given by a paired t-test between the baseline and STIPPLE

Proportion (α)	100%	70%	50%	30%	10%
p values	4.51%	3.00×10^{-4}%	5.53×10^{-2}%	6.40×10^{-6}%	2.60×10^{-7}%

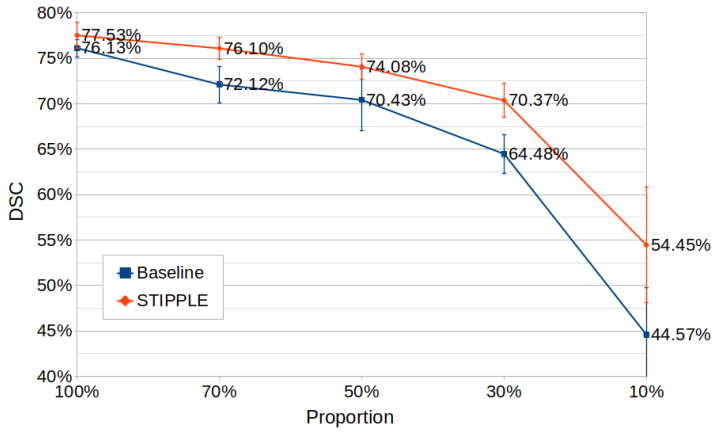


Fig. 3 Segmentation results for STIPPLE ($B = 5$) compared to the baseline. Values are Dice Scores (DSC) for every proportion of missing labels from 100% (every image is labeled) to 10% (only 10% of the images are labeled). Error bars show the standard deviations of the results between the folds

The images could be ambiguous due to the low contrast between the objects and because of the reduced size of the organ region. In medical image segmentation, it is common that the local visual content is insufficient, such that one needs external knowledge for proper segmentation. Moreover, the low balance of labeled pixels makes the model naturally under-segment, and this effect is exacerbated when very few labeled images are provided.

All this causes multiple kinds of errors which are addressed by the prior. Firstly, it reinforces the probabilities in the most probable region and allows to recover missed predictions. Secondly, it reduces false positives by cleaning out

errors far from the region of interest. Finally, the prior stabilizes the relabeling step by selecting only the pixels in the correct region which avoid potential errors that could cause drops in performances.

To illustrate how the spatial prior acts on the predictions, we show in Figure 4 two examples. The first row is a missed prediction which has been correctly recovered thanks to the prior. In that case the visual prediction has been reinforced by the spatial prior shown in the last column. The second row shows how the prior removes improbable segmentation and more generally false positives out of the organ region. We see that the wrong prediction of the baseline is out of the high prior probabilities in the last column. The visual prediction was not sufficient to correctly decide in this area but with STIPPLE the prior has removed the ambiguity and filtered out those errors. In this case, the prior combined with the visual prediction reduces the false positives and has a positive impact on the relabeling step by preventing adding errors.

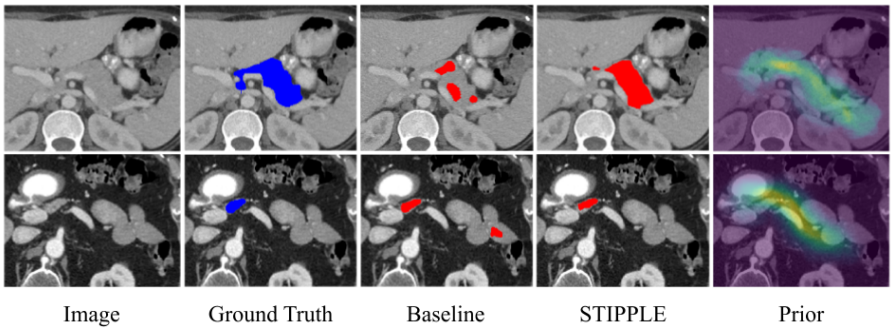


Fig. 4 Examples of two behaviours induced by the spatial prior. First row: recovery of a missed prediction. Second row: cleaning of a wrong prediction in an unexpected area. The last column represents the spatial prior on top of the input image to illustrate where the prior influences the prediction.

To show that our method is agnostic to the choice of the backbone, we carry out experiments using a patch-based 3D U-Net. We choose a fixed fold and add the prior using the same method as explained. At 50%, we observe an improvement for the baseline of +3pts from 68% in DSC to 71% for the 3D U-Net. Then, with the spatial prior we observe an improvement of +1pt validating the relevance of our method. At 10%, our spatial prior with a 3D U-Net gets a 58% DSC outperforming both the baseline (+6pts) and our prior (+3pts) with the 2D U-Net. Our method can easily be extended to other backbones and our 3D spatial prior still improves the final results even with a strong baseline, e.i. 3D U-Net.

4.3 Ablation study

To understand how the different parts of STIPPLE act on the final performance, we show in Table 2 an ablation study of the method. The results are given for the different stages: the 2D U-Net baseline which is also the backbone in our experiments; after adding the 3D prior but without relabeling; the complete method, including the prior and the relabeling step.

Table 2 Ablation study of STIPPLE. The reported values are Dice Similarity Scores (DSC,%).

Proportion (α)	100%	70%	50%	30%	10%
Baseline	76.13 (± 0.94)	72.12 (± 2.01)	70.43 (± 3.38)	64.48 (± 2.13)	44.57 (± 5.24)
STIPPLE w/o relab	77.53 (± 1.44)	75.02 (± 2.21)	71.74 (± 2.02)	65.99 (± 1.71)	47.41 (± 8.40)
Baseline w relab	-	75.12 (± 1.91)	73.71 (± 2.59)	69.00 (± 2.04)	51.91 (± 7.77)
STIPPLE	**77.53** (± 1.44)	**76.10** (± 1.23)	**74.08** (± 1.39)	**70.37** (± 1.88)	**54.45** (± 6.37)

Adding the prior alone outperforms the baseline for every proportion. The relative gains are +1.41pts at 100%, +2.90pts at 70%, +1.32pts at 50%, +1.50pts at 30%, and finally +2.84pts at 10%. The information brought by the spatial prior allows to increase the results consistently through the proportions. This shows the relevance of exploiting the absolute position for organ segmentation. Then, the relabeling step boosts the performances as we can see in the last row. This step is particularly interesting for the low proportions. As discussed in section 4.2, the gains are more and more important when the proportion α is decreasing.

Using a prior impacts positively the performances in the two contexts: with or without relabeling. We can also notice that the relabeling step boosts the results especially for the low αs.

4.4 State-of-the art comparison

We compare our method with other semi-supervised approaches in addition to a method that includes an attention mechanism. In [1], the unlabeled images are completely relabeled before training a new model. [12] propose an adversarial training to incorporate unlabeled images during training. Finally, [22], use a mean teacher method where the unlabeled images are used through the consistency loss. We also compare our method with an attention model from [14]. It uses an additive attention gate in the decoder part of the U-Net before the concatenation of the skip-connections.

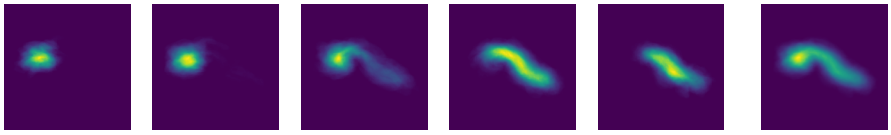
Table 3 shows the results of the comparison. For every row, we implement the method with the same backbone 2D U-Net. STIPPLE shows better results for every proportion with a more pronounced gain in the low αs, *e.g.* at 10%, STIPPLE is better by 2.4pts than the best method (the adversarial). The pseudo-labels method [1] is the closest to ours but we see that STIPPLE stays above for every proportion thanks to the spatial prior and the progressive adding of pseudo-labels.

Table 3 State-of-the-art comparison on TCIA

Proportion (α)	100%	70%	50%	30%	10%
Baseline	76.13 (± 0.94)	72.12 (± 2.01)	70.43 (± 3.38)	64.48 (± 2.13)	44.57 (± 5.24)
Pseudo-labels ([1])	-	75.12 (± 1.91)	73.71 (± 2.59)	69.00 (± 2.04)	51.91 (± 7.77)
Adversarial ([12])	-	75.41 (± 1.78)	73.91 (± 2.27)	67.60 (± 1.84)	52.09 (± 6.00)
Consistency ([22])	-	74.53 (± 2.10)	72.68 (± 3.05)	66.99 (± 1.38)	46.04 (± 3.70)
Attention U-Net [14]	76.38 (± 1.27)	74.18 (± 1.57)	71.37 (± 1.73)	64.25 (± 2.49)	41.28 (± 6.47)
STIPPLE (Ours)	**77.53** (± 1.44)	**76.10** (± 1.23)	**74.08** (± 1.39)	**70.37** (± 1.88)	**54.45** (± 6.37)

Concerning the attention model in [14], we can see that compared to the baseline, it helps consistently from $\alpha = 100\%$ to $\alpha = 50\%$. Then, the scores drop below the baseline. STIPPLE is better for every proportion and especially for the low αs. It could be explained by the fact that our prior exploits the three dimensions unlike the attention module which is 2D. Moreover it is built beforehand by following a specific method which is adapted to low label proportions.

4.5 Further Analysis



(a) $b = 0$ (b) $b = 1$ (c) $b = 2$ (d) $b = 3$ (e) $b = 4$ (f) 2D prior

Fig. 5 Visualization of a spatial prior with $B = 5$. We can see how it captures the depth information compared to (f) which is a 2D prior.

Impact of the prior size B The number of bins, B, of the prior impacts the final results and the best value may depends on the available data. As an example, Figure 5 shows a spatial prior with $B = 5$ and $B = 1$, *i.e.* 2D prior. At $B = 5$, we can see how the spatial position evolves through the 3D prior bins. As a contrary, the 2D prior ($B = 1$) doesn't encode the depth information and is thus less informative.

We evaluate STIPPLE without relabeling with different B values (1, 2, 5, 7, 10 and 90) at 10% and 70% of labeled images, see Figure 6. $B = 90$ means that there is no discretization in z, *i.e.* the spatial prior is complete.

We observe that the best value at 70% is 5 but for every B there is a significant improvement compared to the baseline. At 10%, the best results are given for 5, 7 and 10 with an optimal value at 7. In our experiments in section 4.3, we choose a standard value of $B = 5$. Though it is good in practice, it means that we could get better results by increasing B for lower proportions.

For both proportions, we can see that the prior has better results than the baseline. Using a 2D prior ($B = 1$) is effective but using more bins boosts the

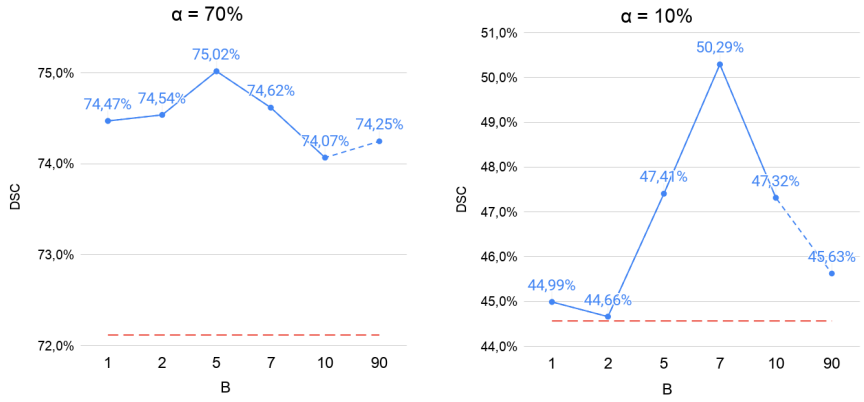


Fig. 6 Dice score versus the number of bins B at 70% and 10% of labeled images. In blue, STIPPLE without relabeling. In dotted red, the baseline.

performances. Then, with a complete prior, $B = 90$, the scores decrease which shows that discretizing the z axis is relevant.

Impact of the prior positioning As explained in section 3.2, the prior has to be positioned in the test volumes. We use the predicted position refined by an adjustment step. Table 4 shows the results with the naive method of using only the center given by the segmentation model. Then, with the adjustment step used in STIPPLE.

As we can see the naive approach is not sufficient and alters the final results. The adjustment step is necessary and allows to reach optimal results comparable to the one obtained by using the true organ position.

Table 4 Impact of the prior positioning on the final results

Proportion	100%	70%	50%	30%	10%
Naive	74.48 (± 2.53)	72.84 (± 3.15)	69.90 (± 1.85)	61.82 (± 3.49)	41.80 (± 9.94)
Ours	**77.53** (± 1.44)	**75.02** (± 2.21)	**71.74** (± 2.02)	**65.99** (± 1.71)	**47.41** (± 8.40)

5 Discussion and Limitations

STIPPLE relies on assumptions such that the position of an organ in (w,h) varies slightly compared to the variations in z. Thus, there could be an issue when strong rotations (e.g. of the patient) occur, or for data mixing various acquisition directions (axial/coronal/sagittal). In this case, our approach would require a (manual or automatic) method to register with respect to those variations.

A second problem could emerge for atypical cases. For examples, for patients with *situs inversus* where the major abdominal organs are reversed

from their normal positions. With STIPPLE we define a spatial prior which translates the observed average position of the organs. However, with certain conditions, it could not apply and a human professional is needed. We must point out that those conditions represent a fraction of the cases and most of the available segmentation datasets do not contain any atypical cases.

However, our method could be adapted to other imaging modalities by adapting the prior computation or the prior positioning depending on the problem. The main idea is the same when a segmentation dataset with dense labels is provided.

6 Conclusion and perspectives

This paper introduces STIPPLE, a method that integrates a 3D spatial prior and pseudo-labels for training FCNs in a semi-supervised context. STIPPLE shows very important gains especially when few images are available making it particularly relevant in the medical field where labeled data are limited and very expensive to obtain. Comparisons with state-of-the-art methods further highlight the relevance of our method compared to attention models and semi-supervision techniques. Future works could be to transfer a prior computed on a large external dataset to another dataset with less data. For example from a modality to another (*e.g.* CT to MRI). Another idea that could be explored is to integrate our spatial prior at different stages of the network. It could be done by combining the prior with a specifically designed attention module. For example a transformer [21].

Declarations

Funding: This study was funded by Visible Patient with le CNAM as an academic partner.
Conflicts of interest/Competing interests: The authors declare that they have no conflict of interest.
Availability of data and material: Data are publicly available.
Ethics approval: This article does not contain any studies with human participants or animals performed by any of the authors.
Informed consent: Not applicable.

References

1. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac mr image segmentation. In: MICCAI, pp. 253–260 (2017)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**, 834–848 (2018)

3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation. In: MICCAI, pp. 424–432 (2016)
4. Dalca, A.V., Guttag, J., Sabuncu, M.R.: Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In: IEEE CVPR, pp. 9290–9299 (2018)
5. Feng, X., Qing, K., Tustison, N.J., Meyer, C.H., Chen, Q.: Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3d images. Medical Physics **46**(5), 2169–2180 (2019)
6. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. In: BMVC (2018)
7. Kakeya, H., Okada, T., Oshiro, Y.: 3d u-japa-net: Mixture of convolutional networks for abdominal multi-organ ct segmentation. In: MICCAI, pp. 426–433 (2018)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS, pp. 1097–1105 (2012)
9. Liu, R., Lehman, J., Molino, P., Such, F.P., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the coordconv solution. In: NeurIPS (2018)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE CVPR, pp. 3431–3440 (2015)
11. Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016)
12. Nie, D., Gao, Y., Wang, L., Shen, D.: Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In: MICCAI, pp. 370–378 (2018)
13. Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M.P., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T., O'Regan, D.P., Kainz, B., Glocker, B., Rueckert, D.: Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation. IEEE Transactions on Medical Imaging **37**, 384–395 (2018)
14. Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: learning where to look for the pancreas. MIDL (2018)
15. Petit, O., Thome, N., Charnoz, A., Hostettler, A., Soler, L.: Handling missing annotations for semantic segmentation with deep convnets. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 20–28 (2018)
16. R. Roth, H., Farag, A., Turkbey, E.B., Lu, L., Liu, J., Summers, R.M.: Data from pancreas-ct. In: The Cancer Imaging Archive, (TCIA) (2016). DOI https://doi.org/10.7937/K9/TCIA.2016.tNB1kqBU
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI, pp. 234–241 (2015)
18. Roth, H.R., Lu, L., Lay, N., Harrison, A.P., Farag, A., Sohn, A., Summers, R.M.: Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Medical image analysis **45**, 94–107 (2018)
19. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: IEEE CVPR (2014)
20. Trullo, R., Petitjean, C., Dubray, B., Ruan, S.: Multiorgan segmentation using distance-aware adversarial networks. Journal of Medical Imaging **6**(1), 014001 (2019)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS, pp. 5998–6008 (2017)
22. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: MICCAI, pp. 605–613 (2019)
23. Zheng, H., Lin, L., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y.W., Tong, R., Wu, J.: Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. In: MICCAI, pp. 148–156 (2019)

A Details on the network used in the study

Table 5 Details of the network's blocks and layers used in STIPPLE. This architecture comes from U-Net [17]. Convolutions are given by conv(kernel_size, filters). The model has 32M parameters.

block name	output size	layer's parameters
input	$512 \times 512 \times 1$	
encoder_block_1	$256 \times 256 \times 64$	conv(3×3, 64) + relu conv(3×3, 64) + BN + relu $\rightarrow$ res_1 max_pool(2×2)
encoder_block_2	$128 \times 128 \times 128$	conv(3×3, 128) + relu conv(3×3, 128) + BN + relu $\rightarrow$ res_2 max_pool(2×2)
encoder_block_3	$64 \times 64 \times 256$	conv(3×3, 256) + relu conv(3×3, 256) + BN + relu $\rightarrow$ res_3 max_pool(2×2)
encoder_block_4	$32 \times 32 \times 512$	conv(3×3, 512) + relu conv(3×3, 512) + BN + relu $\rightarrow$ res_4 max_pool(2×2)
decoder_block_4	$64 \times 64 \times 1024$	conv(3×3, 1024) + relu conv(3×3, 1024) + BN + relu upsampling(2×2) conv(2×2, 512) + BN + relu concat(res_4)
decoder_block_3	$128 \times 128 \times 512$	conv(3×3, 512) + relu conv(3×3, 512) + BN + relu upsampling(2×2) conv(2×2, 256) + BN + relu concat(res_3)
decoder_block_2	$256 \times 256 \times 256$	conv(3×3, 256) + relu conv(3×3, 256) + BN + relu upsampling(2×2) conv(2×2, 128) + BN + relu concat(res_2)
decoder_block_1	$512 \times 512 \times 128$	conv(3×3, 128) + relu conv(3×3, 128) + BN + relu upsampling(2×2) conv(2×2, 64) + BN + relu concat(res_1)
final_prediction	$512 \times 512 \times 64$	conv(3×3, 64) + relu conv(3×3, 64) + relu
output_probabilities	$512 \times 512 \times nb\_classes$	conv(1×1, nb_classes) + {softmax;sigmoid}

B Additional training details

In this work we use a 2D U-Net as our main backbone FCN. It was trained with a batch size of 6. The learning rate was $1e-4$ with an inverse time decay scheduler and a decay rate set to get a learning rate of $1e-5$ at the end of training. We train the model for 25 epochs which corresponds to the observed convergence in all the experiments.

The data augmentation consists of small random translations (e.g. between -15 and +15), small rotations (e.g. -6 to +6 degrees) and zooms (e.g. 0.9 to 1.1). This augmentation is applied to the image and label but also to the prior. Moreover, we have another augmentation on the prior to simulate imperfect positioning by using an additional translation.

The code was developed with tensorflow and the training performed on Nvidia RTX 2080Ti GPU cards.