Iterative Confidence Relabeling with Deep ConvNets for Organ Segmentation with Partial Labels

Olivier Petit, Nicolas Thome, Luc Soler

Iterative Confidence Relabeling with Deep ConvNets for Organ Segmentation with Partial Labels

Olivier Petit[a,b], Nicolas Thome[a] and Luc Soler[b]

[a]CEDRIC, Conservatoire National des Arts et Metiers, 292 rue Saint-Martin, Paris, 75003, France
[b]Visible Patient, 8 rue Gustave Adolphe Hirn, Strasbourg, 67000, France

ARTICLE INFO

Keywords:
medical images
deep learning
convolutional neural networks
partial-labels
noisy labels
self-training
uncertainty estimation

ABSTRACT

Training deep ConvNets requires large labeled datasets. However, collecting pixel-level labels for medical image segmentation is very expensive and requires a high level of expertise. In addition, most existing segmentation masks provided by clinical experts focus on specific anatomical structures. In this paper, we propose a method dedicated to handle such partially labeled medical image datasets. We propose a strategy to identify pixels for which labels are correct, and to train Fully Convolutional Neural Networks with a multi-label loss adapted to this context. In addition, we introduce an iterative confidence self-training approach inspired by curriculum learning to relabel missing pixel labels, which relies on selecting the most confident prediction with a specifically designed confidence network that learns an uncertainty measure which is leveraged in our relabeling process. Our approach, INERRANT for Iterative coNfidencE Relabeling of paRtial ANnoTations, is thoroughly evaluated on two public datasets (TCAI and LITS), and one internal dataset with seven abdominal organ classes. We show that INERRANT robustly deals with partial labels, performing similarly to a model trained on all labels even for large missing label proportions. We also highlight the importance of our iterative learning scheme and the proposed confidence measure for optimal performance. Finally we show a practical use case where a limited number of completely labeled data are enriched by publicly available but partially labeled data.

1. Introduction

Abdominal organ segmentation is a major challenge in medical imaging and computed-aided diagnosis. Good localization and segmentation of internal structures are important for radiologists, which helps them to compare physical changes in response to a treatment. It also offers important tools for surgeons in planning treatments and interventions in addition to other computer-assisted applications, *e.g.* Augmented Reality.

Currently, state-of-the-art methods for visual recognition rely on deep learning. Convolutional Neural Networks (ConvNets) [25] and more precisely Fully Convolutional Neural Networks (FCNs) [31] have become standard solutions for semantic segmentation of generalist images. In the context of medical image segmentation, specific architectures such as U-Net and variants [39, 9, 32, 29] are standard choices showing optimal performances.

However, an important issue when training deep ConvNets is the need of having a large amount of labeled data. The problem is particularly pronounced for medical image segmentation, where the label process is extremely time-consuming and requires highly qualified professionals. As a consequence, large-scale and clean medical image datasets are rarely available. In abdominal organ segmentation, the manual label process often focuses on specific anatomical structures, *e.g.* the liver and its pathologies. Thus, large datasets containing partially labeled images are easier to obtain by aggregating smaller labeled datasets with different amounts of labels compared to a complete dataset containing all the abdominal organs.

In this paper, we address the problem of training deep ConvNets with partially labeled datasets. Our training context is illustrated in Figure 1: in this example, the input slice is partially labeled with 3 organ classes out of 7 for the unknown complete labeling. As we verify experimentally, naively applying state-of-the-art models such as U-Net to these partial labels leads to bad performances, since it includes wrongly labeled background pixels for missing organs.

To specifically handle the partial labeling problem, we introduce a method which encompasses two main contributions:

- Firstly, we propose a specific loss to train the segmentation network dedicated to include only correct labels, *i.e.* it selects pixels that could be learned and those that should be ignored during training (white *vs* black pixels in Figure 1). The general motivation is to eliminate all pixels that are wrongly labeled as background for missing organs.
- Secondly, we propose a self-supervised scheme to iteratively relabel the missing organs by introducing pseudo-labels into the training set, in order to estimate the unknown complete ground-truth labels. For that, we add a confidence network that helps to select the best pseudo-labels and thus reduce the introduction of wrong predictions.

Our overall approach is called INERRANT, Iterative coNfidencE Relabeling of paRtial ANnoTations.

Our proposed method strongly relies on the medical nature of the considered images. We leverage import priors

*Corresponding author
olivier.petit@visiblepatient.com (O. Petit)
ORCID(s): 0000-0002-7895-9217 (O. Petit)

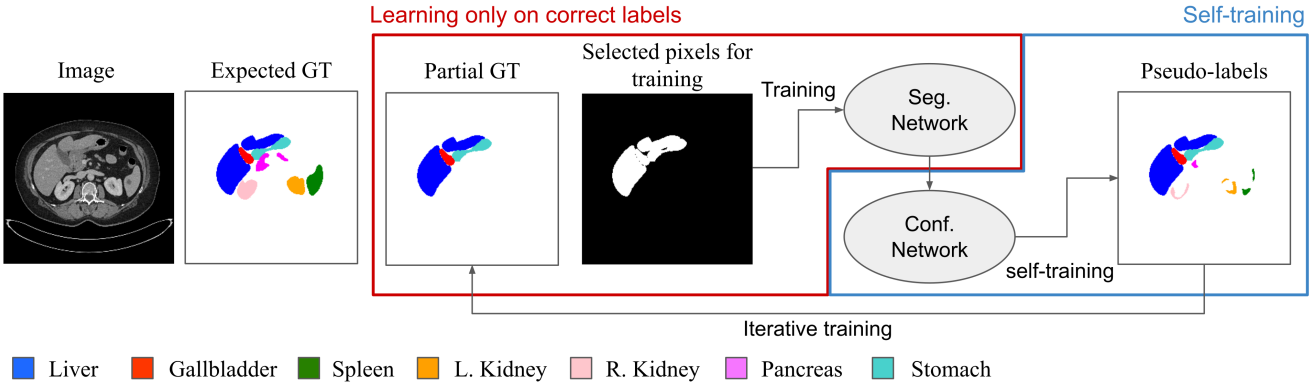


Figure 1: The 3D CT-scan is partially labeled: in this slice, only 3 out of 7 organs are labeled. Naively using such partial ground-truth (GT) labels is inappropriate since it includes wrong background labels for missing organs. INERRANT is based on identifying pixels for which labels are correct, and ignoring others. The segmentation network is trained on those data and a confidence network outputs confidence scores for each pixel to incrementally add pseudo-labels to the training set and recover the unknown complete ground-truth labels.

about the organ labels such that every organ is present, even if non-labeled, in the input volume at only one place (a class is assigned to only one object in the volume). It allows us to deduce the unlabeled organs for each volume and thus, which classes should be ignored during training.

INERRANT builds upon the preliminary work proposed in [38]. We extend [38] by:

1. Using a better confidence estimation for generating pseudo-labels. The confidence is learned via a dedicated network best suited for distinguishing errors from correct predictions, which enables to maximise the number of correct labels introduced during pseudo-labeling.
2. Providing a comparison with recent state-of-the-art semi-supervised methods for learning with partial labels. In addition, we include ablation studies highlighting the importance of the new confidence measure and the iterative pseudo-labeling process.
3. Providing a much more thorough evaluation by reporting performances on two public datasets (TCIA and LiTS) and one internal dataset containing seven organ classes (*vs* 3 in [38]). Moreover, we show the genericity of our approach by using a U-Net as our backbone model whereas [38] uses a simple FCN model.
4. Showing a practical use case where we combine a multi-organ dataset with a single-organ public dataset (TCIA). This shows that we can exploit large amounts of labeled images by gathering heterogeneous data.

2. Related Work

2.1. Abdominal CT Organ Segmentation

Automatic organ segmentation has been widely studied. Early works used for instance atlas-based segmentation [24, 20, 41, 45, 21, 44] or statistical shape models [33, 40, 35, 17].

Over the last decade, deep learning has made dramatic breakthroughs in machine learning. Since their historical success for image classification [25], deep ConvNets have been used in every visual recognition problem including semantic segmentation [31, 8, 39, 1].

In the medical imaging field, deep learning has been adapted to answer various types of problems. For organ segmentation, U-Net [39] is the most popular for 2D images. Then equivalents have been proposed for 3D segmentation [9, 32, 13] based on the same encoder-decoder, skip connections architecture. Other networks aim to compensate the memory issue of the 3D networks by composing the 2D and 3D segmentation into a single model [29].

State-of-the-art approaches segment a specific organ or a limited number of structures. Using them directly on a partially-labeled dataset implies learning with noisy labels due to the default background label assigned for the missing organs. Thus, we need a specific method for training deep learning models with partially labeled data.

2.2. Semi-supervised learning and self-training

In the context of partially labeled data, some image pixels are incorrectly labeled, *i.e.* there are incorrect "background" labels when an organ is missing. As we explain in section 3, our approach is based on distinguishing pixels for which labels are certain from those for which labels are ambiguous. Ambiguous pixels are first ignored (Section 3.1) and then regarded as unlabeled (Section 3.2). Therefore, our approach is cast as a Semi-Supervised Learning (SSL) problem [7] and we discuss here the approaches most related to ours. Basically, SSL approaches in medical image segmentation can be classified into generative models, teacher-student networks and pseudo-labeling methods [7].

Generative models can be leveraged to incorporate training signals on unlabeled data for medical image segmentation. For example, [42] uses a variational autoencoder (VAE) to learn representations on all images, and then train a decoder only on labeled data. In the same idea, [11] applies

a generative model based on a VAE, where the encoder is trained to reconstruct input images, and the decoder to reconstruct unpaired segmentation masks. Adversarial training [14] is another appealing direction for semi-supervised semantic segmentation. The overall idea initially applied to generalist images in [19], is to consider the segmentation network as a conditional generator given input images, which output distribution should be similar to the ground truth distribution of segmentation masks. The appealing feature in SSL is that this adversarial loss can be applied on unlabeled data to improve segmentation performances. Recently, the approach has also been successfully applied for medical image segmentation [34, 48]. In those methods, when an image is unlabeled the output of the discriminator is used as a confidence map to compute a segmentation loss between the encoder prediction and its binarized counterpart for the most confident pixels.

Teacher-student networks have also been used in SSL to enforce desirable behaviours on the segmentation models, where the teacher is learned only on the labeled data and the student is subsequently trained on all data. Some methods introduce an auxiliary task that does not need the segmentation ground truth. In [23] and [49], the authors proposed to regress the region size and use a consistency term that penalizes non-realistic sizes. In the same way, the fact that the same image under different transformations should get the same output is used to create a consistency term. For example in [43, 4, 6, 46] this idea is applied by defining two losses, the first is the classic segmentation loss and the second the consistency loss which does not need ground truth labels. In [46], the authors proposed an advanced method by introducing a confidence estimation based on monte carlo dropout to select the most certain predictions in the consistency term for the unlabeled images.

Although these SSL methods show good results, the incorporation of the unlabeled data in the final results is implicit. Pseudo-labeling [15, 28] consists in using the model's predictions as ground truth training signals on unlabeled data. In our context of partial labels, the goal of these approaches is to automatically relabel unlabeled data from a model trained on a labeled set. Recently, this strategy has been extensively applied for semi-supervised semantic segmentation, [30, 53, 52], leading to state-of-the-art performances. This strategy has also recently been applied for medical image segmentation [2, 50, 47]: the idea is to first learn a model on the labeled data. Then, enlarge the training set with the union of the labeled data and the model's predictions for the unlabeled data. Finally, either the same model or a new model is trained on the new training set.

Our approach is based on pseudo-labeling. In contrast to previous works [30, 53, 52, 2, 50, 47], which perform a complete pseudo-labeling of unlabeled data in a single relabeling iteration, we use a smoother and more progressive way of introducing new labels by selecting more confident labels first to control the rate of miss-predictions added to the training set. This approach could be seen as a curriculum learning strategy [3] or self-paced learning [26], where the easy examples have the most confident predictions and the hard examples have the least confident ones.

2.3. Confidence estimation in Deep Learning

Confidence estimation in deep learning is a crucial yet complex problem. The most naive confidence estimation for deep neural networks consists in using the probability of the predicted class, *i.e.* the Maximum Class Probability (MCP) [18]. Although this baseline is widely used in practice, it also suffers from fundamental drawbacks, *e.g.* the probabilities are known to be non-calibrated [16]. In the last few years, there has been an extensive revival of Bayesian deep learning, especially by the connections drawn between variational inference and stochastic regularization in deep learning, *e.g.* Monte-Carlo Dropout [12]. However, this confidence measure is computationally demanding since it requires several forward passes, and does not yield accurate uncertainty measures when aleatoric uncertainty is crucial. In contrast, misclassification approaches design confidence estimates targeted to properly separate correct predictions from errors, *e.g.* trust score [22] or ConfidNet [10].

In pseudo-labeling, the chosen confidence measure should prevent incorporating wrong labels to improve the final prediction. It is worth mentioning that most recent approaches for semantic segmentation rely on MCP for selecting target labeled pixels [30, 53, 52, 2, 50, 47], although MCP by design assigns overestimated confidence values to prediction errors. In this paper, we train an auxiliary network to design a relevant confidence measure, which is based on misclassification detection and explicitly assigns low confidence values to prediction errors. We verify experimentally that this confidence measure leads to better final segmentation performances than MCP.

3. Training from partial labels with INERRANT

In this section we detail INERRANT for training deep ConvNets on partially labeled data.

Firstly, we introduce in section 3.1 a learning scheme that only leverages correct labels. More precisely, INERRANT is trained not only with the true positives (TPs), *i.e.* the positive labels which are actually positive in the complete ground truth, but also with the true negatives (TNs), *i.e.* the background labels which are actually background in the ground truth. As mentioned in section 1, a naive method that learns directly with partial labels incorporate false negatives (FNs), which negatively impact performances. We also provide a statistical analysis of the ratio of correct labels used by our method *vs* the naive baseline.

Then, we introduce in section 3.2 a self-supervised scheme which iteratively adds pseudo-labels for the missing organs, in order to recover the missing ground-truth labels. Since the pseudo-labeling is automatic, the challenge is to maximise the number of correct label predictions, denoted as true positives (TPs), while minimizing the number of wrong predictions denoted as false positives (FPs). Ultimately, we aim at maximizing the relabel precision $TP/(TP + FP)$.

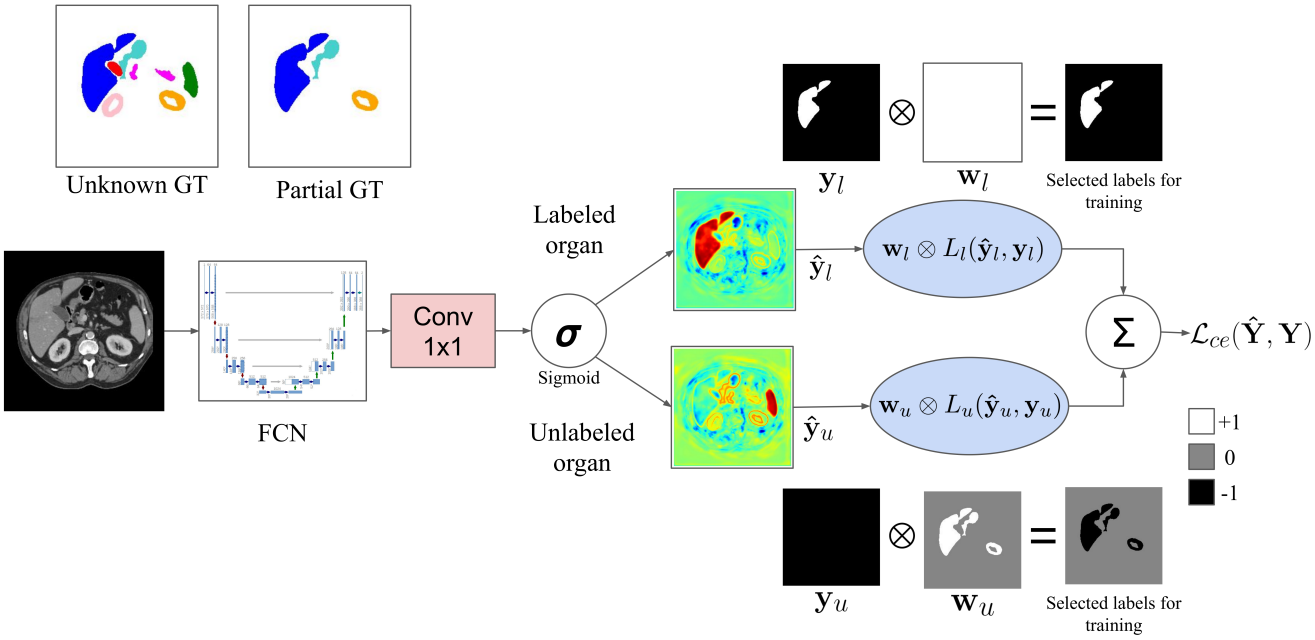


Figure 2: Training INERRANT on a partially labeled dataset. Each organ is predicted by a common FCN. Depending on the missing organs deduced by the available labels, an ambiguity map $\mathbf{w}_k$ is created to ignore potential wrong labels in the loss. It acts as a weighting in the final loss function.

Since our method is iterative, INERRANT^0 is the first step which consists in learning on the partially labeled data without relabeling, and INERRANT corresponds to the method after training the model on the incorporated pseudo-labels.

3.1. Learning on a partially labeled dataset

We address the issue of learning on partially labeled data by a simple yet effective method, which is shown in Figure 2. The first step consists in extracting the maximum of information from the partially labeled data, by deducing from the labeled organs where there are ambiguities that should be handled.

3.1.1. Training exclusively with correct labels

We know by construction that if an organ is unlabeled, then it is the case for the entire volume, *i.e.* no intermediate slice contains this label. Thus, we can deduce beforehand the missing classes for every patient. However, we do not know where they are located and thus where the wrong labels are.

However, if we want to exclusively use correct labels, we cannot use a classic softmax activation function and a multi-class loss. Indeed, in that configuration when only one organ is missing no background label can be used. To address this problem, we transform the $(K+1)$ multiclass classification problem into K binary classification problems where each organ is learned independently. The rationale behind this is to control the classes that are labeled and can be learned and those that are unlabeled and have to be ignored. By doing that we can learn features from the labeled classes for both the positives (the organs) and the negatives (the background) whereas for the unlabeled classes, both positives and negatives are ignored.

In practice, we replace the final softmax by a sigmoid activation function in the last 1×1 convolution layer. However, we still want to keep the exclusive aspect of the softmax, *i.e.* only one class is predicted for a given voxel. Thus, our class prediction is computed by taking, for each voxel, the class with the highest probability among all K classes - and the background label is assigned if all probabilities are lower than 0.5.

Training K binary classifiers requires adjustments, especially on the loss function. Actually, we have K losses, one for each class. We choose the binary cross entropy to train our model defined in Equation 1 for each voxel i and class k:

$$l_{i,k}(\hat{y}_{i,k}, y_{i,k}) = -(y_{i,k}\, log(\hat{y}_{i,k}) + (1-y_{i,k})\, log(1-\hat{y}_{i,k})) \quad (1)$$

Let us denote as $\hat{\mathbf{Y}} \in \mathbb{R}^{H,W,K}$ the dense prediction of our model and $\mathbf{Y} \in \mathbb{R}^{H,W,K}$ as the ground truth. Then the K losses are aggregated to obtain one final loss in Equation 2:

$$\mathcal{L}_{ce}(\hat{\mathbf{Y}}, \mathbf{Y}) = \sum_{k=1}^{K} \sum_{i=1}^{N} w_{i,k}\; l_{i,k}(\hat{y}_{i,k}, y_{i,k}) \quad (2)$$

where $\mathbf{W} \in \mathbb{R}^{H,W,K}$ composed of K maps $\mathbf{w_k} \in \mathbb{R}^{H,W}$, is a binary matrix which selects or discards the voxels that should be learned for class k, for which back-propagation is applied.

$\mathbf{W}$ is an ambiguity map since it represents the pixels' location where we cannot decide if the label is correct or not.

Table 1
TP/ FP training label analysis

(a) Naive

GT \ Used	Pos	Neg
Pos	$(1-\alpha)\cdot\beta_k$	$\alpha\cdot\beta_k$
Neg	0	$1-\beta_k$

(b) INERRANT^0

GT \ Used	Pos	Neg
Pos	$(1-\alpha)\cdot\beta_k$	0
Neg	0	$(1-\alpha)\cdot(1-\beta_k)+\epsilon$

$\mathbf{W}$ is built beforehand based on the missing organs of each patient. As shown in Figure 2, if an organ is labeled, we fill $\mathbf{w}_k$ with ones to learn the associated model. On the other hand, when an organ is missing $\mathbf{w}_k$ is set to zeros to ignore this organ during training. However, we can still use extra information from other organs, which are assigned as negative labels.

In the example of Figure 2, three organs are labeled. However, when learning a missing organ like the spleen (bottom branch), we use an ambiguity map containing zeros everywhere except where the other organs. In that case the label of the organ is used to fill the ambiguity map of the spleen with ones.

3.1.2. Statistical analysis of the training labels

To quantify the quality of the labels used during training, let us consider the binary classification problem for the k^{th} organ class. We denote as β_k the number of pixels for this organ and α the ratio of missing organs on the whole dataset. Table 1 shows confusion matrices for two different methods: *naive* consists in learning directly with the partial labels, and our method *INERRANT*^{0}. We can see that the naive method has $\alpha\cdot\beta_k$ FNs. Meanwhile, INERRANT^0 completely discards FNs but also reduces the number of TNs.

The naive approach learns with $(1-\beta_k)$ TNs whereas INERRANT^0 learns with $(1-\alpha)(1-\beta_k)+\epsilon$ TNs, where $\epsilon=\sum_{k'\neq k}\beta_{k'}$ corresponds to the other organ labels. In medical image segmentation, organs represent usually a small proportion of the total volume of labels, which induces a high class imbalance between positives and negatives, such that $\beta << 1$, *e.g.* $\beta = 0.05$. As a consequence, we still have largely enough information to properly learn the background class with *INERRANT*^{0}.

3.2. Self-supervision and pseudo-labeling

The number of TPs linearly decreases with the ratio of missing organs α. To recover missing labels in training images, we propose to iteratively add new positive labels $y_{i,t}=1$ in an image with missing labels $\mathbf{x}_i$ for each class k[1], using a curriculum strategy [3].

[1] We drop the dependence of class in $y_{i,t}$ for clarity.

3.2.1. Iterative relabeling

Initially, the model is trained on all correct labels that can be regarded as "easy positive samples". Let us denote as $\hat{y}_i^+$, the pixels predicted as positive for a given unlabeled image $\mathbf{x}_i$. The idea of INERRANT is to recover positive labels, $y_{i,t}^+$ by selecting the top scoring pixels among $\hat{y}_i^+$. Then, the model is retrained with the new labels added to the training set.

This procedure is iteratively performed T times, by selecting a ratio $\gamma_t=\frac{t}{T}\gamma_{max}$ of top scoring pixels among the positives. The pseudo-labels incorporated at each step are the "hard examples" since they come from a pseudo-labeling scheme that could introduce errors.

Algorithm 1: Training INERRANT for class k

Data: $\{(x_i, y_i)\}$, γ_{max}, T, m_0
Result: m_T
$N_u \leftarrow$ number of unlabeled images;
$y_{i,0} = y_i$;
for $t \leftarrow 1$ *to* T **do**
  $\gamma_t = \frac{t}{T}\gamma_{max}$;
  for $i = 1$ *to* N_u **do**
    $\hat{y}_i^+ \leftarrow m_t(x_i)$ // Take predicted $\oplus$;
    $y_{i,t}^+ \leftarrow s(\hat{y}_i^+, \gamma_t)$ // Assign new $\oplus$ target labels;
    $y_{i,t} = y_{i,t-1} \cup y_{i,t}^+$ // Augment training set;
  $m_t = train(\{(x_i, y_{i,t})\})$ // Re-train model

3.2.2. Uncertainty estimate for collecting pseudo-labels

Our pseudo-labeling approach in Algorithm 1 is based on selecting the most confident pixels of the segmentation model. We therefore seek an accurate confidence criterion for our deep FCN in semantic segmentation.

Measuring model uncertainty in deep learning is an open and difficult problem, as detailed in Section 2. Although the Maximum Class Probability (MCP [18]) gives decent performances in practice, it also suffers from important conceptual limitations (see Section 2). Especially, misclassified pixels (failures) receive an unjustified high confidence. In our pseudo-labeling approach, this presents the risk of including wrong labels and negatively impacting performances.

Therefore, we propose to use a more relevant uncertainty measure. Our target confidence criterion is the True Class Probability (TCP [10]), from which guarantees can be derived for discriminating correct from incorrect predictions (TCP is able to assign small confidence values to misclassifications). Since TCP requires the knowledge of the ground truth class for each pixel, which is not accessible at test time, we need an auxiliary network specifically dedicated to predict the TCP value computed by our segmentation model, e.g. U-Net. For each pixel $i \in \{1, ..., N\}$ and each class $k \in \{1, ..., K\}$, we want the predicted confidence $\hat{c}_{i,k}$ to match $TCP_{i,k} = c_{i,k}$: learning the confidence network is

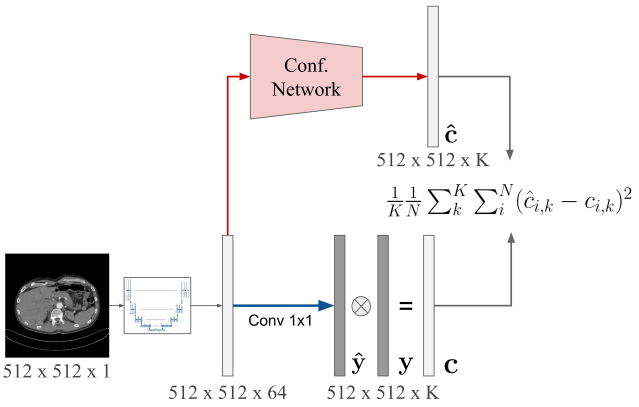


Figure 3: The confidence network part is included at the end of the segmentation network by taking the features before the final 1×1 convolutional layer

a regression task where we use the following L2 loss:

$$\mathcal{L}_{conf} = \frac{1}{K}\frac{1}{N}\sum_{k}^{K}\sum_{i}^{N}(\hat{c}_{i,k} - c_{i,k})^2 \tag{3}$$

The confidence network is illustrated in Figure 3. It is attached to the segmentation model in order to leverage latent representations learned for the segmentation task. In practice, we connect it to the antepenultimate layer, i.e. before the final 1×1 convolutional layer.

The confidence network is thus initialized with parameters from the segmentation model. During training, we can freeze these parameters or fine-tune them, which we find superior in practice. If the entire model is fine-tuned, a duplicate of the original FCN allows to keep the same segmentation predictions.

The confidence network is trained before relabeling and after training the segmentation network. Algorithm 2 shows the different steps of training our model by using pseudo-labels generated iteratively.

Algorithm 2: Relabeling the missing organs with the confidence network

```
Train the FCN on partially labeled data;
for t ← 1 to T do
    Train the confidence network;
    Relabel the (t/T)γ_max pixels with the highest
      confidence score;
    Fine-tune the initial FCN with the new labels;
```

4. Experiments and Results

4.1. Experimental setup

4.1.1. Datasets

We use three datasets for abdominal organ segmentation: Liver and Tumour Segmentation challenge[2] (LiTS), TCIA pancreas segmentation dataset[3] and a private multi-organ dataset.

LiTS dataset contains 131 CT-scans with the segmentation of livers and tumors. We focus on the task of liver segmentation and discard the tumors. Each CT-scan is composed of $74 \sim 987$ slices of 512×512 pixels and a voxel spatial resolution of $([0.56 \sim 1.0] \times [0.56 \sim 1.0] \times [0.70 \sim 5.0])\text{mm}^3$.

The TCIA dataset contains 82 CT-scans with the pancreas completely labeled in each image. Each CT-scan is composed of $181 \sim 466$ slices of 512×512 pixels and a voxel spatial resolution of $([0.66 \sim 0.98] \times [0.66 \sim 0.98] \times [0.5 \sim 1.0])\text{mm}^3$.

The private multi-organ dataset is composed of 90 CT-scans where the liver, gallbladder, pancreas, spleen, right and left kidneys and stomach are completely labeled. Each CT-scan is composed of $57 \sim 500$ slices of 512×512 pixels and a voxel spatial resolution of $([0.42 \sim 0.98] \times [0.42 \sim 0.98] \times [0.63 \sim 4.00])\text{mm}^3$.

4.1.2. Simulating partially labeled datasets

Large datasets for organ segmentation are tedious and expensive to obtain. Depending on the medical center and the patient's pathology only some organs are labeled for a given case. Consequently, it is easier to gather data with heterogeneous labels but the resulting dataset will be partially labeled. To reproduce this context and analyse how to model performed under different amounts of missing labels, we start from fully labeled datasets and randomly remove the labels at a volume level. Thus, we reproduce real clinical conditions and keep control over the exact quantity of available information. Moreover we can evaluate the method on a completely labeled test set. The proportion of labeled organs in each volume is denoted as α. When $\alpha = 100\%$, all the organs are labeled and when $\alpha = 0\%$ no label is available in each volume.

For the multi-organ dataset, the label proportion α is applied to every organ, independently. It means that $\alpha\%$ of the cases have a labeled liver, $\alpha\%$ a labeled spleen, etc. Thus, a case could have between 0 and 7 labeled organs. Moreover, we paid attention to incrementally remove labels. The same labeled organs are found through the different proportions, i.e. with $\alpha = 70\%$, the dataset contains all the labels of a dataset with $\alpha = 50\%$ but with more of them. In the labels point of view, we can say that $D(10\%) \subset D(30\%) \subset D(50\%) \subset D(70\%)$. This allows fair comparisons between the different proportions as they are trained with the same labeled images.

4.1.3. Implementation details

We use a U-Net as our main FCN which is well-known for 2D medical image segmentation. This model is still extensively used as it gives competitive results though it requires reasonable memory cost and can be trained on standard GPUs.

The standard U-Net used in our experiments is around

[2] https://competitions.codalab.org/competitions/17094

[3] https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT

Table 2
Quantitative results for the TCIA pancreas dataset. The scores are the mean DSC (± std) for every missing label proportion (α). In bold the highest results that pass a t-test with p-value < 0.05 compared to the other methods.

Proportion (α)	100%	70%	50%	30%	10%
Naive	76.13 (± 0.94)	49.75 (± 5.58)	28.99 (± 6.07)	10.75 (± 5.71)	1.16 (± 0.77)
INERRANT^0	-	72.12 (± 2.01)	70.43 (± 3.38)	64.48 (± 2.13)	44.57 (± 5.24)
INERRANT	-	**75.52** (± 1.74)	**74.23** (± 2.50)	**71.10** (± 1.52)	**56.19** (± 6.22)
INERRANT^0 3D	78.76 (± 1.91)	77.22 (± 2.41)	75.59 (± 1.69)	71.73 (± 1.93)	52.98 (± 8.83)
INERRANT 3D	-	**77.35** (± 1.67)	**76.02** (± 0.88)	**73.41** (± 1.00)	**57.77** (± 7.53)

Table 3
Quantitative results for the LiTS dataset. The scores are the mean DSC (± std) for every missing label's proportions (α). In bold the highest results that pass a t-test with p-value < 0.05 compared to the other methods.

Proportion (α)	100%	30%	10%	5%	1%
Naive	94.72 (± 1.22)	14.10 (± 6.28)	0.41 (± 0.18)	1.14 (± 2.47)	0.31 (± 0.53)
INERRANT^0	-	93.12 (± 1.41)	89.70 (± 2.51)	88.22 (± 2.87)	51.08 (± 13.80)
INERRANT	-	**93.51** (± 1.15)	**90.05** (± 1.41)	**88.88** (± 2.48)	**58.76** (± 10.94)

31M parameters. The confidence network only adds 0.8M parameters but this network is only used for the relabeling step and is discarded for the final prediction network which is simply the U-Net, thus our method does not add any computational nor memory overhead compared to the baseline in test (See Table 11 for a detailed overview of the network used in the study including the layers' parameters). The models are trained with the Adam optimizer and an initial learning rate of 10^{-4} which exponentially decreases to 10^{-5} at the end of the training. Standard data augmentation techniques are used including random translations, random rotations and random scales. The models are implemented with the Tensorflow library and the training is performed on RTX 2080Ti GPUs. We perform 5 fold cross-validation for every dataset and proportion. The results shown in section 4.2 give the mean Dice Similarity Coefficient (DSC) and standard deviation across the folds.

The overall quantitative evaluation carried out in section 4.2 gives the results for the naive baseline, *i.e.* when the model is trained directly on the data. Then with the proposed INERRANT^0. And finally with INERRANT that iteratively adds pseudo-labels using the previously introduced confidence network. Then, a finer analysis of the impact of curriculum iterations and confidence measures is provided in section 4.3.

4.2. Quantitative results

To highlight the problem of training on partially labeled data, we evaluate the naive approach which consists in learning on the partially labeled data with the background label assigned to missing pixel labels. Then, we show the results using our method, first with the ambiguity map only (INERRANT^0) and then using pseudo-labels (INERRANT).

4.2.1. TCIA pancreas

Results for the TCIA pancreas dataset are given in Table 2. As we can see, the naive approach quickly deteriorates when the number of missing labels increases, i.e. α decreases. For example, with $\alpha = 70\%$, we already observe a drop of about 26.4pts in DSC. By assigning the background label to missing organ labels, this naive baseline makes the model trained with many wrong labels of an already over-represented class. So, it naturally tends to predict "background" for the entire image.

INERRANT^0 gives better results as the model is trained only on correct labels. We can see that even with $\alpha = 30\%$, which is less than the third of the labels, we lose 11.6pts when the naive baseline is at less than 11% in DSC.

Next, INERRANT which introduces pseudo-labels helps to improve the mean DSC for every proportion. We can even see that the gain increases when α decreases. At $\alpha = 10\%$ INERRANT has improved the results by 10pts. The gains are significant and shows the relevance of the proposed

Table 4
Quantitative results on multi-organ dataset. The scores are the mean DSC (± std) for every missing label proportion (α). In bold the highest results that pass a t-test with p-value < 0.05 compared to the other methods.

Proportion (α)	100%	70%	50%	30%	10%
Naive	86.03 (± 2.16)	66.85 (± 4.89)	45.32 (± 2.67)	19.51 (± 2.39)	2.82 (± 1.30)
INERRANT^0	-	84.19 (± 2.85)	81.25 (± 5.51)	76.58 (± 7.15)	67.69 (± 5.34)
INERRANT	-	**85.36** (± 2.70)	**84.43** (± 3.56)	**82.60** (± 3.40)	**73.49** (± 3.08)

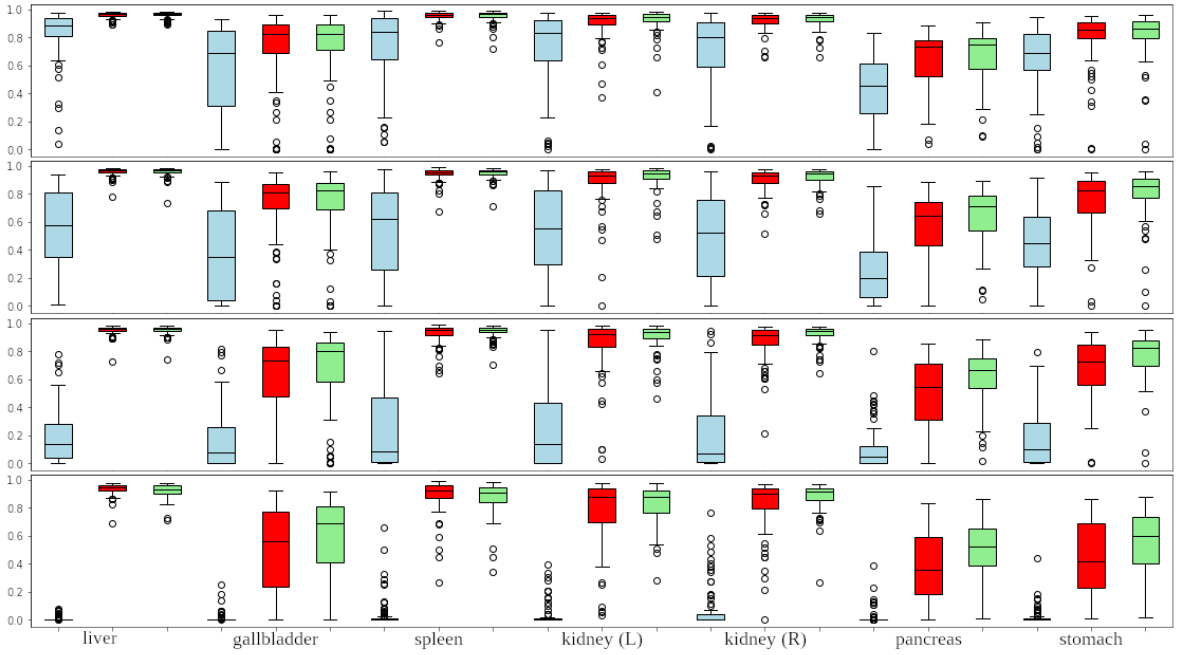


Figure 4: Per patient DSC scores analysis for the multi-organ dataset. First row with $\alpha = 70\%$, second $\alpha = 50\%$, third $\alpha = 30\%$ and fourth $\alpha = 10\%$. In blue the naive method, red INERRANT^0 and green INERRANT with pseudo-labeling.

method and how using pseudo-labels can improve the final scores.

Finally, Table 2 also reports results with a 3D backbone. In INERRANT^0 3D and INERRANT 3D we replace the 2D U-Net with its 3D counterpart to show that our method is agnostic to the chosen backbone FCN. In this setup, we have an input patch size of $144 \times 144 \times 96$ which is cropped in the center of the image. This could also explain the performance boost compared to the 2D U-Net, however this method could not be applied to the multi-organ setup where one should perform predictions with, for example, a sliding window. Nevertheless, the same trends are observable: the relabeling step INERRANT 3D outperformed INERRANT^0 3D for every proportion and the highest gain is at $\alpha = 10\%$ with +4.79pts.

4.2.2. LiTS

Contrary to the pancreas, the liver is easier to segment, since it is one of the largest organs in the abdomen, leading to more pixel labels. In addition, its boundaries are less ambiguous.

Table 3 shows the results on the LiTS dataset. We observe that the performance of the baseline U-Net for $\alpha = 100\%$ is high, i.e. more than 94% DSC. It is worth mentioning that for $\alpha = 30\%$, the naive baseline already gives terrible results.

The interesting point here is the fact that INERRANT^0 gives very high results even with very few examples. As we can see the result with $\alpha = 5\%$ loses only 6.5pts compared to the model trained on 100% of data. In this dataset, $\alpha = 5\%$ correspond to only 5 labeled cases which correspond to a reduction in labels by a factor of 20.

Moreover, the relabeling step helps to consistently improve the results. The most important gain is again with the lower α (*i.e.* $\alpha = 1\%$) with a difference of 7.7pts.

With this dataset, the overall conclusion is similar to TCIA, but the regime is different. As described above, the scores of our approach without relabeling are very high even with few labels. However, introducing pseudo-labels still improves the model, with the largest gain at $\alpha = 1\%$.

4.2.3. multi-organ dataset

Figure 4 shows the results on this dataset detailed per organ and Table 4 the average DSC for all proportions and the different methods (scores per organ are detailed in Table 9). As we can see, all the methods give better results compared to LiTS and TCIA. This can be explained by two important points. Firstly, the background class is less represented because we have multiple organs. Secondly, considering INERRANT, for one particular case only 1 or 2 organs could be unlabeled especially for high proportions like 70%. It implies that a lot of background labels could be correctly learned even without the organ label thanks to the other organs. This shows that our method is actually strengthened in the case of multi-organ with missing labels. We can see in Figure 2 an example of an ambiguity map for a missing organ (bottom branch) in a case where some labels are available. We can notice that a wide part of the image can be used to learn a negative label where all the other organs are located.

Considering the naive baseline, as for the two previous datasets, the scores quickly fall until reaching a very low

Table 5
State-of-the-art comparison on the TCIA pancreas dataset

Proportion (α)	70%	50%	30%	10%
Naive	49.75 (± 5.58)	28.99 (± 6.07)	10.75 (± 5.71)	1.16 (± 0.77)
INERRANT^0	72.12 (± 2.01)	70.43 (± 3.38)	64.48 (± 2.13)	44.57 (± 5.24)
Pseudo-labels ([2])	**75.12** (± 1.91)	**73.71** (± 2.59)	69.00 (± 2.04)	51.91 (± 7.77)
Adversarial ([34])	**75.41** (± 1.78)	**73.91** (± 2.27)	67.60 (± 1.84)	52.09 (± 6.00)
Consistency ([46])	**74.53** (± 2.10)	**72.68** (± 3.05)	66.99 (± 1.38)	46.04 (± 3.70)
INERRANT (Ours)	**75.52** (± 1.74)	**74.23** (± 2.50)	**71.10** (± 1.52)	**56.19** (± 6.22)

value of 2.8% when $\alpha = 10\%$. For our method, however, it gives good performances even with few labels. But depending on the organ, the behavior is different. The liver, spleen and kidneys, stay with high scores even with few labels with an impressive result for the liver that only loses 3pts between 100% and 10%.

On the other hand, the gallbladder, the pancreas and the stomach fall more quickly than the other organs. Those organs are the smallest and in general more difficult to segment. For instance with the gallbladder, the segmentation model tends to segment it as the liver because it is located close to it in addition to being very small. The pancreas is also difficult to segment due to its complex boundaries and pixel intensities which are very close to the connected structures. Finally, the stomach is difficult to segment because of its shape, size and position variability in addition to the presence of air which makes holes in the structure that add randomness about the organ visibility.

INERRANT^0 gives 48.96% for the gallbladder, 37.04% for the pancreas and 44.05% for the stomach at $\alpha = 10\%$. But after adding the pseudo-labels the most important gains are with those 3 organs. Respectively, 57.93% (+9pts), 50.25% (+13.2pts) and 56.03% (+12pts).

The curriculum learning approach combined with the learned confidence boosts the results for every organ and every proportion. The most impressive gains are for the most difficult organs which are the gallbladder, pancreas and stomach. It could be explained by the fact that those organs need more labels due to their complexity, and we show that our pseudo-labeling approach greatly helps to comply with this requirement.

4.2.4. State-of-the-art comparison

We compare INERRANT with three other semi-supervised methods representing three different types of approaches. Firstly, [2] which consists in using all the predictions as pseudo-labels. Then, [34] which is an adversarial training where the output of the discriminator allows to select pseudo-labels on the fly by adding them to the segmentation loss during training. And finally [46] which is a mean teacher model based on [43] that uses unlabeled data through a consistency loss.

We implemented the above mentioned methods with the same backbone FCN (*i.e.* 2D U-Net) to segment the pancreas from the TCIA segmentation dataset. Each experiment is evaluated with a 5-fold cross-validation. The models are trained with the same procedure, *i.e.* with the same dataset, the same folds, the same missing labels and with the appropriate hyper-parameters. Table 5 shows the results for every approach compared to the baselines that didn't use unlabeled images.

With [2] all the predictions for the missing organs are used as pseudo-labels. It gives better results than INERRANT^0 as it injects more information with the correct pseudo-labels. However, though it performs well with high α values, it tends to add a lot of wrong labels with low α values which reduces the gains. We can see that using a better pseudo-label selection scheme, we can prevent this effect while preserving the performances with high α values.

Concerning the adversarial training [34], the method gives comparable results to [2]. We can see that the results are better than INERRANT^0 but INERRANT still outperformed it for every proportion. This model can leverage a meaningful loss applicable to unlabeled data, but is hard to train due to instabilities in the adversarial approach.

The consistency method based on mean teacher [46] still improves over INERRANT^0, but is not the best performing strategy for handling unlabeled data in our context. For $\alpha = 10\%$, the performance drop is significant compared to the other approaches. It can be explained by the fact that the loss function does not explicitly exploit predicted segmentation masks on unlabeled data.

In all cases, we can see that INERRANT performs better that the other methods, with a gain being more pronounced for low α proportions.

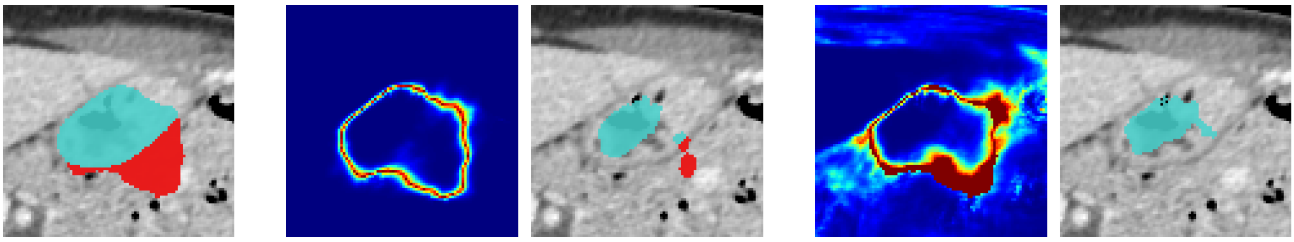
4.3. Model analysis

This section aims to provide an analysis of the relabeling. First, we discuss the differences between the uncertainty evaluation methods and how they impact the relabeling and final score. Then, we show the impact of the curriculum learning and how it behaves depending on the number of performed relabeling steps.

4.3.1. Uncertainty methods evaluation

We evaluate the performances of the proposed confidence network as described in section 3.2.2, and compared it with MCP, which corresponds to the previous work of SMILE [38], on the TCIA pancreas dataset.

The confidence network can be trained with two different configurations, by transfer learning: the U-Net is frozen during the confidence training, or by fine-tuning: the U-Net and the confidence network are both trained. For the last configuration, it is necessary to duplicate the U-Net part of the model which adds complexity to the final model. However,



(a) Prediction (b) MCP (c) Learned Conf.

Figure 5: Confidence maps for MCP and our confidence network for the stomach. The prediction in (a) gives the TPs in cyan and FPs in red. For both MCP (b) and the learned confidence (c), a confidence map is given with values between 0.5 (red) and 1.0 (blue) and the selected pseudo-labels with the TPs in cyan and the FPs in red. In (b), MCP gives low confidence only at the boundaries. As a contrary in (c) the confidence network gives low confidence values to the model errors and thus prevents relabeling wrong predictions.

we found in practice that fine-tuning gives better results, thus the following results are obtained with this method.

A detailed analysis of the impact of the two uncertainty estimation methods is provided in Table 6 on the TCIA pancreas dataset (Additional results for the multi-organ dataset could be find in Table 10). We evaluate how the confidence score ranks the pixels considered for relabel (We relabel only the positives and never the background). Three metrics are shown: the AUC (area under the ROC curve), the Average Precision of the success (AP_success), and the AP of the errors (AP_error). The first metric gives a measure of the overall ranking of the predictions. The second, measures the method's capacity of assigning high values to the correct predictions. Finally, the AP error gives a measure of the method's capacity of assigning low values to the wrong predictions.

Table 6 shows significant improvements for all the metrics and for the different proportions. At 10%, the relative gain is the more important. We observe an improvement of 1.53pts in AUC, 0.75pt in AP_success and 2.65pts in AP_error. It means that we have a better ranking of the candidates in addition to a better error detection which translates into an improvement of the final DSC after training the model on the pseudo-labels of 1.5pts. At this proportion, the absolute gain is equivalent to the one at 70% but the relative gain is higher in the way that it will impact much more the final results.

Qualitatively, Figure 5 shows uncertainty maps for both methods. We can notice that the learned confidence has a more detailed result than MCP. In fact, MCP concentrates the low confidence values at the border whereas the confidence network assigns lower confidence values to the model errors. In this example a part of the segmentation, at the bottom right, is wrong and we can see that the confidence network has assigned lower values at this place than MCP. This illustrates how our confidence network helps to prevent the relabeling of wrong predictions and thus the incorporation of errors in the training set.

Table 6
Analysis of ranking metrics for uncertainty estimation with MCP, equivalent to SMILE [38], and the learned confidence method. The metrics are computed only on the pixels that are considered for relabeling, *i.e.* predicted as positive and not already relabeled. The values are percentages.

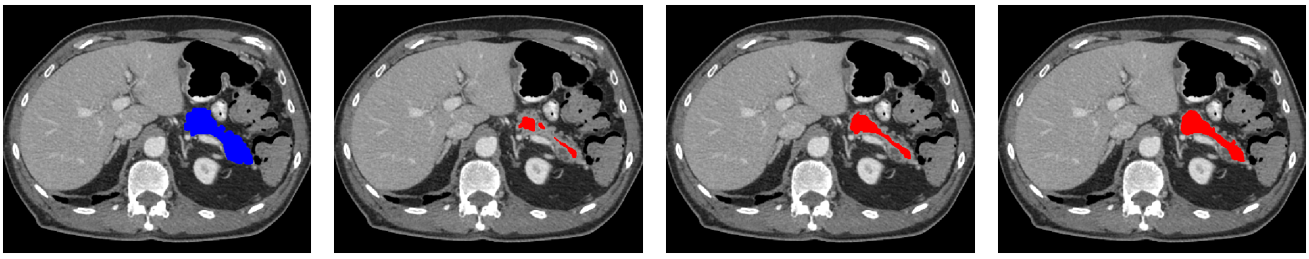
Method	AUC	AP_success	AP_error	Final DSC
70%				
MCP	73.86 (± 1.02)	92.00 (± 0.71)	34.99 (± 2.34)	73.97 (± 1.28)
L. conf.	**75.50** (± 1.77)	**92.66** (± 0.83)	**38.17** (± 3.86)	**75.52** (± 1.74)
50%				
MCP	72.67 (± 1.05)	90.51 (± 1.97)	36.50 (± 3.06)	73.82 (± 2.15)
L. conf.	**73.94** (± 0.94)	**91.06** (± 1.60)	**38.69** (± 4.55)	**74.23** (± 2.50)
30%				
MCP	71.55 (± 1.95)	90.58 (± 1.43)	34.29 (± 2.68)	69.72 (± 1.75)
L. conf.	**73.06** (± 2.00)	**91.25** (± 1.24)	**36.80** (± 4.11)	**71.10** (± 1.52)
10%				
MCP	68.68 (± 2.28)	84.97 (± 3.91)	41.11 (± 4.85)	54.66 (± 6.53)
L. conf.	**70.21** (± 3.46)	**85.72** (± 4.09)	**43.76** (± 7.70)	**56.19** (± 6.22)

4.3.2. Curriculum learning analysis

Curriculum learning consists in introducing easy examples before adding more complex ones. In our application, the easy examples are the available labels and the more complex, the pseudo-labels which contain wrong labels. The pseudo-labels are introduced incrementally by first taking the most confident predictions and ending by the less confident that would by definition contain more wrong labels.

As we can see in Figure 6, using an iterative approach allows us to relabel progressively the missing organ from the center to the border. In fact, we noticed that the most certain predictions were located in the center and that the confidence decreases as we move closer to the border (see Figure 5).

Table 7 presents a quantitative evaluation of the iterative relabeling, and shows the relabeling precision and recall for each step of the curriculum learning method with the final DSC after fine-tuning the model on them. Overall, using $T = 2$ relabeling iterations is the best strategy, although differences can be observed for different levels of missing labels α. For $\alpha > 50\%$, the relabeling precision is higher and thus the best results are with the last step. On the other hand, with $\alpha < 30\%$, the best results are for an intermediate step



(a) Ground truth (b) $t = 0$ (c) $t = 1$ (d) $t = 2$

Figure 6: Complete relabeling of a pancreas with INERRANT, $T = 3$ iterations, $\gamma_{max} = 1.0$ and $\alpha = 50\%$.

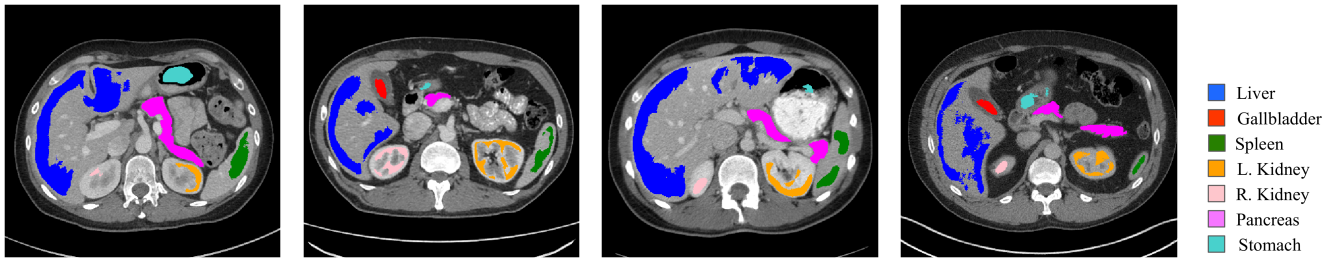


Figure 7: Relabeling of TCIA images with a model trained with only 9 completely labeled images from the private multi-organ dataset.

Table 7
Complete organ relabel detailed for 3 steps on the multi-organ dataset. Information given are the percentage of added pixels, the relabeling precision and recall and the final DSC after training on the updated dataset. Values are percentages.

α/step	Added pixels	Relal. P	Relab. R	Final DSC
50%				
0	0%	-	-	79.81
1	33%	98.14	19.99	84.93
2	66%	95.11	25.89	85.15
3	100%	89.04	25.87	**85.53**
30%				
0	0%	-	-	72.27
1	33%	96.62	18.37	82.17
2	66%	93.62	25.58	**83.79**
3	100%	85.20	25.81	83.30
10%				
0	0%	-	-	58.98
1	33%	94.44	15.57	72.89
2	66%	81.80	23.70	**74.26**
3	100%	65.09	23.96	72.01

because performing the last iterations adds too many wrong predictions and thus deteriorates the model performances. However, it is worth noting that for every proportion the relabeling improves the final score.

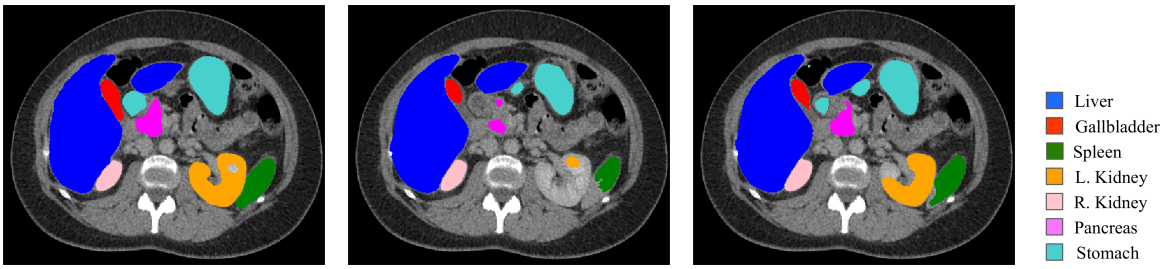
4.3.3. Qualitative results

To illustrate the previous results, Figure 8 shows an example of a segmentation result for the multi-organ dataset for INERRANT^0 and after training on the pseudo-labels, INERRANT. We can notice that INERRANT helps by segmenting more pixels and thus fill organs that have been missed by INERRANT^0.

Figure 6 is an example of a complete relabel of a missing pancreas. It illustrates how the method progressively adds more pixels from the most certain (in the center) to the least certain (at the border).

4.4. Fusion of heterogeneous data from multiple datasets

Completely labeled data for abdominal organ segmentation are expensive and tedious to obtain. In this experiment, we show that with INERRANT we can build a good segmentation model by starting with few completely labeled examples and leveraging public datasets with few labeled organs. Thus, we use 9 cases from the multi-organ dataset with the 7 organs completely labeled and add the 82 cases from the TCIA datasets which are partially-labeled compared to the multi-organ cases (only the pancreas is available). Then we evaluate on the remaining 81 multi-organ examples. In Table 8, we evaluate a model trained only on the completely labeled 9 cases. Then we add the 82 cases from TCIA and follow the INERRANT method. We can see a large improvement for every organ, especially for the small ones, *i.e.* the gallbladder, +9.5pts, the stomach, +10.5pts and obviously the pancreas, +25.6pts. It is worth noting that even if both datasets are abdominal CT-scans, there is a slight domain shift. In fact, we have two different sources that acquire data under different parameters, then the quality of the annotations could be very different. For instance in TCIA we can



(a) Ground Truth (b) INERRANT^0 (c) INERRANT

Figure 8: Segmentation results for INERRANT^0 and INERRANT, $\alpha = 30\%$.

Method	Multi-organ	TCIA	Liver	Gallbladder	Spleen	Kidney (L)	Kidney (R)	Pancreas	Stomach	Avg.
INERRANT^0	9	0	89.43	48.09	84.72	78.39	80.78	32.55	48.13	66.02
INERRANT	9	82	89.85	57.63	87.46	85.22	85.33	58.15	58.85	74.64

Table 8
Results in DSC (%) when combining 9 completely labeled examples from the multi-organ dataset with the 82 partially labeled examples (only the pancreas) of the TCIA dataset with INERRANT. The models are evaluated on the remaining 81 multi-organ examples.

assume that the pancreas' annotations are more precise as they focus on this very organ. A qualitative evaluation is provided in Figure 7. It shows how we relabel the TCIA examples based on a model trained only on 9 completely labeled images.

This experiment points out that even with a little domain shift we can build a better model by enriching a small dataset with external sources of images.

Discussion

The disposal of large-scale labeled and publicly available datasets has increased recently, *e.g.* CT-ORG [5]. Having access to such large-scale public datasets is very valuable and can help to provide more powerful prediction models. However, collecting large-scale datasets that are "universal" and could be useful for any medical image segmentation task arguably remains elusive. For example, the multi-organ dataset used in our paper contains 90 CT-scans with 7 abdominal organs, while CT-ORG is larger in terms of cases (140 CT-scans) but with fewer labeled organs: 6 organ classes, and only 3 in common with ours (liver, gallbladder and kidney). This illustrates the challenge addressed in this paper: despite the existence of massively annotated datasets, it is very difficult to compile a complete, exhaustive and homogeneous dataset for any medical problem. Heterogeneity in medical imaging can have various sources. Firstly, granularity between studies might substantially differ: datasets on the entire body will focus on large structures (*e.g.* bones, lungs, liver), a study focusing precisely on the abdomen will try to get finer structures (*e.g.* . pancreas, spleen, stomach), while finer tasks could even include the vascularisation with vein/artery networks. Secondly, there are commonly strong variabilities in the acquisition process between studies: images are acquired with different devices and different protocols: images depend on the injection time of the contrast media which is chosen depending on the targeted structure [51, 27, 37].

Our method is complementary with the access of large datasets by leveraging various types of labels and granularities to build a more exhaustive dataset and thus a more robust segmentation model. Moreover, it opens up the possibility to add a new organ class which is less represented in public and private datasets to enrich an existing one.

5. Conclusion

This paper introduces INERRANT, a method dedicated to address the challenging problem of learning with partial labels. The approach is based on a specifically designed loss for ignoring ambiguous labels coupled with an iterative pseudo-labeling scheme. We introduce a confidence network that learns an uncertainty criterion leveraged by the relabeling process which iteratively adds new labels to the training set. In our experiments we show very good results on three abdominal organ segmentation datasets. Moreover, we observe that our method is even more relevant and efficient with low label proportions.

Our approach is agnostic to the prediction model, and we generalize the results in [38] that uses a simple FCN with a U-Net model. We show the good performances obtained by INERRANT compared to state-of-the-art semi-supervised methods. Last but not least, we provide a showcase illustrating INERRANT's capacity to combine real datasets with different labeling and how it improves segmentation performances.

For future work, an interesting perspective is to explore how to leverage different ways of using unlabeled data into our training method, *e.g.* combining pseudo-labeling with other semi-supervised approaches in section 2. We could

also include prior knowledge about the organs to improve the relabeling, *e.g.* using attention mechanisms [36].

References

[1] Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 2481–2495.

[2] Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D., 2017. Semi-supervised learning for network-based cardiac mr image segmentation, in: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017, Springer International Publishing, Cham. pp. 253–260.

[3] Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, NY, USA. pp. 41–48.

[4] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A., 2019. Mixmatch: A holistic approach to semi-supervised learning, in: Advances in Neural Information Processing Systems, pp. 5049–5059.

[5] Blaine Rister, Kaushik Shivakumar, T.N., Rubin, D.L., 2019. Ct-org: Ct volumes with multiple organ segmentations [dataset] doi:https://doi.org/10.7937/tcia.2019.tt7f4v7o.

[6] Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., de Bruijne, M., 2019. Semi-supervised medical image segmentation via learning consistency under transformations, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 810–818.

[7] Chapelle, O., Schölkopf, B., Zien, A., 2006. Semi-Supervised Learning (Adaptive Computation and Machine Learning). The MIT Press.

[8] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A., 2018. 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.

[9] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: Learning dense volumetric segmentation from sparse annotation, in: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer International Publishing, Cham. pp. 424–432.

[10] Corbière, C., Thome, N., Bar-Hen, A., Cord, M., Pérez, P., 2019. Addressing failure prediction by learning model confidence, in: Advances in Neural Information Processing Systems, pp. 2902–2913.

[11] Dalca, A.V., Guttag, J., Sabuncu, M.R., 2018. Anatomical priors in convolutional networks for unsupervised biomedical segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9290–9299.

[12] Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, pp. 1050–1059.

[13] Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018. Automatic multi-organ segmentation on abdominal ct with dense v-networks. IEEE Transactions on Medical Imaging 37, 1822–1834. doi:10.1109/TMI.2018.2806309.

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 2672–2680.

[15] Grandvalet, Y., Bengio, Y., 2005. Semi-supervised learning by entropy minimization, in: Advances in neural information processing systems, pp. 529–536.

[16] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks, PMLR, International Convention Centre, Sydney, Australia. pp. 1321–1330.

[17] Hammon, M., Cavallaro, A., Erdt, M., Dankerl, P., Kirschner, M., Drechsler, K., Wesarg, S., Uder, M., Janka, R., 2013. Model-based pancreas segmentation in portal venous phase contrast-enhanced ct images. Journal of digital imaging 26, 1082–1090.

[18] Hendrycks, D., Gimpel, K., 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. Proceedings of International Conference on Learning Representations .

[19] Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H., 2018. Adversarial learning for semi-supervised semantic segmentation, in: Proceedings of the British Machine Vision Conference (BMVC).

[20] Hyunjin Park, Bland, P.H., Meyer, C.R., 2003. Construction of an abdominal probabilistic atlas and its application in segmentation. IEEE Transactions on Medical Imaging 22, 483–492.

[21] Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: A survey. Medical Image Analysis 24, 205 – 219.

[22] Jiang, H., Kim, B., Guan, M., Gupta, M., 2018. To trust or not to trust a classifier, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31. Curran Associates, Inc., pp. 5541–5552.

[23] Kervadec, H., Dolz, J., Granger, É., Ayed, I.B., 2019. Curriculum semi-supervised segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 568–576.

[24] Klein, A., Mensh, B., Ghosh, S., Tourville, J., Hirsch, J., 2005. Mindboggle: automated brain labeling with multiple atlases. BMC medical imaging 5, 7.

[25] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105.

[26] Kumar, M.P., Packer, B., Koller, D., 2010. Self-paced learning for latent variable models, in: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 23. Curran Associates, Inc., pp. 1189–1197.

[27] Küstner, T., Müller, S., Fischer, M., Weiß, J., Nikolaou, K., Bamberg, F., Yang, B., Schick, F., Gatidis, S., 2018. Semantic organ segmentation in 3d whole-body mr images, in: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 3498–3502. doi:10.1109/ICIP.2018.8451205.

[28] Lee, D.H., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on challenges in representation learning, ICML.

[29] Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A., 2018. H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging 37, 2663–2674.

[30] Li, Y., Yuan, L., Vasconcelos, N., 2019. Bidirectional learning for domain adaptation of semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6936–6945.

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

[32] Milletari, F., Navab, N., Ahmadi, S., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571.

[33] Neumann, A., Lorenz, C., 1998. Statistical shape model based segmentation of medical images. Computerized Medical Imaging and Graphics 22, 133 – 143.

[34] Nie, D., Gao, Y., Wang, L., Shen, D., 2018. Asdnet: Attention based semi-supervised deep networks for medical image segmentation, in: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer

Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham. pp. 370–378.

[35] Okada, T., Linguraru, M.G., Hori, M., Summers, R.M., Tomiyama, N., Sato, Y., 2015. Abdominal multi-organ segmentation from ct images using conditional shape–location and unsupervised intensity priors. Medical Image Analysis 26, 1–18. doi:https://doi.org/10.1016/j.media.2015.06.009.

[36] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: learning where to look for the pancreas. MIDL .

[37] Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., Rueckert, D., 2019. Data efficient unsupervised domain adaptation for cross-modality image segmentation, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pp. 669–677.

[38] Petit, O., Thome, N., Charnoz, A., Hostettler, A., Soler, L., 2018. Handling missing annotations for semantic segmentation with deep convnets, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA workshop MICCAI). Springer, pp. 20–28.

[39] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham. pp. 234–241.

[40] Saito, A., Nawano, S., Shimizu, A., 2016. Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs. Medical Image Analysis 28, 46 – 65.

[41] Schreibmann, E., Marcus, D., Fox, T., 2014. Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search. Journal of applied clinical medical physics / American College of Medical Physics 15, 4468. doi:10.1120/jacmp.v15i4.4468.

[42] Sedai, S., Mahapatra, D., Hewavitharanage, S., Maetschke, S., Garnavi, R., 2017. Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 75–82.

[43] Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Advances in neural information processing systems, pp. 1195–1204.

[44] Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A., 2012. Multi-atlas segmentation with joint label fusion. IEEE transactions on pattern analysis and machine intelligence 35, 611–623.

[45] Wolz, R., Chengwen, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D., 2013. Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE transactions on medical imaging 32, 1723–1730. doi:10.1109/TMI.2013.2265805.

[46] Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 605–613.

[47] Zhao, Y.X., Zhang, Y.M., Song, M., Liu, C.L., 2019. Multi-view semi-supervised 3d whole brain segmentation with a self-ensemble network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 256–265.

[48] Zheng, H., Lin, L., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y.W., Tong, R., Wu, J., 2019. Semi-supervised segmentation of liver using adversarial learning with deep atlas prior, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 148–156.

[49] Zhou, Y., Li, Z., Bai, S., Chen, X., Han, M., Wang, C., Fishman, E., Yuille, A., 2019. Prior-aware neural network for partially-supervised multi-organ segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10671–10680.

[50] Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E., Yuille, A., 2019. Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 121–140.

[51] Zhu, X., Cheng, Z., Wang, S., Chen, X., Lu, G., 2021. Coronary angiography image segmentation based on pspnet. Computer Methods and Programs in Biomedicine 200, 105897. doi:https://doi.org/10.1016/j.cmpb.2020.105897.

[52] Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J., 2019. Confidence regularized self-training, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 5982–5991.

[53] Zou, Y., Yu, Z., Vijaya Kumar, B., Wang, J., 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training, in: Proceedings of the European conference on computer vision (ECCV), pp. 289–305.

A. Complementary results

Table 9
Results on the multi-organ dataset detailed per organ

Method	Liver	Gallbladder	Spleen	Kidney (L)	Kidney (R)	Pancreas	Stomach
70%							
Naive	83.37 (± 5.64)	58.36 (± 5.22)	74.60 (± 12.05)	72.64 (± 13.93)	71.11 (± 12.60)	43.26 (± 4.76)	64.63 (± 6.45)
INERRANT^0	96.14 (± 0.45)	72.25 (± 8.90)	95.31 (± 0.70)	90.33 (± 2.97)	91.83 (± 2.07)	64.06 (± 6.16)	79.42 (± 8.30)
INERRANT	**96.22** (± 0.50)	**72.95** (± 9.91)	**95.37** (± 1.12)	**92.51** (± 1.95)	**92.69** (± 1.49)	**67.25** (± 4.32)	**80.57** (± 8.23)
50%							
Naive	57.23 (± 9.00)	35.87 (± 10.13)	54.90 (± 12.90)	52.61 (± 13.05)	48.66 (± 6.89)	24.36 (± 6.27)	43.65 (± 5.36)
INERRANT^0	95.81 (± 0.62)	70.09 (± 9.77)	94.27 (± 0.84)	87.76 (± 5.83)	90.16 (± 3.33)	55.59 (± 16.37)	75.05 (± 9.56)
INERRANT	**95.93** (± 0.79)	**72.75** (± 9.54)	**94.99** (± 1.17)	**91.59** (± 3.26)	**92.14** (± 1.75)	**64.15** (± 8.23)	**79.49** (± 8.80)
30%							
Naive	19.07 (± 4.66)	15.51 (± 3.53)	27.48 (± 9.63)	24.95 (± 9.08)	21.36 (± 10.75)	10.26 (± 4.35)	17.95 (± 2.75)
INERRANT^0	95.34 (± 0.79)	60.75 (± 15.37)	92.56 (± 1.91)	84.53 (± 7.69)	86.81 (± 5.66)	48.78 (± 13.83)	67.26 (± 9.01)
INERRANT	**95.38** (± 0.81)	**67.23** (± 11.34)	**94.57** (± 0.97)	**90.69** (± 1.89)	**92.09** (± 1.58)	**61.99** (± 7.10)	**76.25** (± 5.67)
10%							
Naive	0.56 (± 0.58)	1.12 (± 0.64)	4.03 (± 3.28)	3.41 (± 1.49)	7.03 (± 9.12)	1.62 (± 1.37)	1.99 (± 1.25)
INERRANT^0	**93.56** (± 1.07)	48.96 (± 10.03)	**89.41** (± 2.82)	78.00 (± 14.13)	82.84 (± 9.87)	37.04 (± 5.87)	44.05 (± 11.67)
INERRANT	92.45 (± 1.35)	**57.93** (± 11.59)	87.20 (± 3.87)	**82.12** (± 7.06)	**88.46** (± 3.18)	**50.25** (± 3.63)	**56.03** (± 9.57)

Table 10
Analysis of ranking metrics for uncertainty estimation with MCP and the learned confidence method. Results are given per organ for the multi-organ dataset in average across the folds.

Method		Liver	Gallbladder	Spleen	Kidney (L)	Kidney (R)	Pancreas	Stomach
70%								
	AUC	86.09	79.37	91.89	83.69	86.46	75.55	75.70
MCP	AP_success	99.15	97.26	99.48	98.03	98.81	94.87	95.25
	AP_error	27.21	23.63	**29.17**	30.80	29.96	28.13	25.32
	AUC	**89.99**	**81.83**	**92.78**	**87.41**	**88.18**	**77.97**	**81.82**
Learned conf.	AP_success	**99.43**	**97.72**	**99.67**	**98.35**	**98.91**	**95.53**	**96.22**
	AP_error	**33.89**	**29.22**	26.69	**33.31**	**32.58**	**32.26**	**37.65**
50%								
	AUC	87.55	82.93	93.61	78.26	83.84	72.66	71.82
MCP	AP_success	99.24	97.91	99.81	97.36	98.59	93.84	94.69
	AP_error	29.20	27.70	23.64	24.80	24.69	27.35	22.33
	AUC	**91.11**	**85.76**	**94.38**	**83.96**	**85.59**	**77.38**	**81.24**
Learned conf.	AP_success	**99.49**	**98.37**	**99.87**	**98.03**	**98.73**	**95.03**	**96.71**
	AP_error	**39.60**	**39.65**	**27.17**	**37.12**	**31.08**	**37.67**	**44.28**
30%								
	AUC	87.04	**83.04**	**91.27**	79.91	83.21	69.06	**75.07**
MCP	AP_success	99.04	97.27	99.69	97.33	98.80	90.27	96.11
	AP_error	31.50	**36.41**	**20.92**	25.66	20.62	30.94	20.50
	AUC	**90.89**	82.57	90.92	**84.30**	**87.49**	**72.21**	74.51
Learned conf.	AP_success	**99.39**	**97.31**	**99.72**	**98.10**	**99.12**	**91.55**	**96.20**
	AP_error	**39.29**	34.34	18.15	**31.70**	**28.82**	**36.08**	**21.71**
10%								
	AUC	85.41	**76.35**	87.34	**82.75**	85.58	68.59	**71.49**
MCP	AP_success	98.41	**96.42**	98.79	96.39	98.42	83.85	**90.31**
	AP_error	31.40	24.69	27.49	**39.79**	30.63	42.57	32.98
	AUC	**88.50**	75.61	**89.00**	82.34	**86.05**	**69.48**	70.81
Learned conf.	AP_success	**98.88**	96.41	**99.09**	**96.83**	**98.56**	**84.39**	90.19
	AP_error	**34.21**	**27.39**	**32.15**	39.07	**34.16**	**44.77**	**32.32**

Table 11
Details of the network's blocks and layers used in the study. This architecture comes from U-Net [39]. Convolutions are given by conv(kernel_size, filters). The final two blocks: output_probabilities and confidence_network, are connected to the last block of the network, *i.e.* final_prediction. The overall number of parameters reaches 32M parameters including the confidence network which is around 0.8M 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 → 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 → 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 → 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 → 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}
confidence_network	$512 \times 512 \times nb\_classes$	conv(3×3, 400) + relu conv(3×3, 120) + relu conv(3×3, 64) + relu conv(3×3, 64) + relu conv(1×1, nb_classes) + sigmoid