Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Deep learning with mixed supervision for brain tumor segmentation

Pawel Mlynarski
Hervé Delingette
Antonio Criminisi
Nicholas Ayache

SPIE.

Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, Nicholas Ayache, "Deep learning with mixed supervision for brain tumor segmentation," *J. Med. Imag.* **6**(3), 034002 (2019), doi: 10.1117/1.JMI.6.3.034002.

Journal of Medical Imaging 6(3), 034002 (Jul–Sep 2019)

Deep learning with mixed supervision for brain tumor segmentation

Pawel Mlynarski,[a,*] Hervé Delingette,[a] Antonio Criminisi,[b] and Nicholas Ayache[a]
[a]Université Côte d'Azur, Inria, Epione Research Team, Sophia Antipolis, France
[b]Microsoft Research, Cambridge, United Kingdom

Abstract. Most of the current state-of-the-art methods for tumor segmentation are based on machine learning models trained manually on segmented images. This type of training data is particularly costly, as manual delineation of tumors is not only time-consuming but also requires medical expertise. On the other hand, images with a provided global label (indicating presence or absence of a tumor) are less informative but can be obtained at a substantially lower cost. We propose to use both types of training data (fully annotated and weakly annotated) to train a deep learning model for segmentation. The idea of our approach is to extend segmentation networks with an additional branch performing image-level classification. The model is jointly trained for segmentation and classification tasks to exploit the information contained in weakly annotated images while preventing the network from learning features that are irrelevant for the segmentation task. We evaluate our method on the challenging task of brain tumor segmentation in magnetic resonance images from the Brain Tumor Segmentation 2018 Challenge. We show that the proposed approach provides a significant improvement in segmentation performance compared to the standard supervised learning. The observed improvement is proportional to the ratio between weakly annotated and fully annotated images available for training. [DOI: 10.1117/1.JMI.6.3.034002]

Keywords: semisupervised learning; convolutional neural networks; segmentation; tumor; magnetic resonance imaging.

Paper 18265RR received Dec. 7, 2018; accepted for publication Jul. 16, 2019; published online Aug. 10, 2019.

1 Introduction

Today, cancer is the third highest cause of mortality worldwide. In this paper, we focus on segmentation of gliomas, which are the most frequent primary brain cancers.[1] Gliomas are particularly malignant tumors and can be broadly classified according to their grade into low grade gliomas (grades I and II defined by World Health Organization) and high grades gliomas (grades III and IV). Glioblastoma multiforme is the most malignant form of glioma and is associated with a very poor prognosis: the average survival time under therapy is between 12 and 14 months.

Medical images play a key role in diagnosis, therapy planning, and monitoring of cancers. Treatment protocols often include evaluation of tumor volumes and locations. In particular, for radiotherapy planning, clinicians have to manually delineate target volumes, which is a difficult and time-consuming task. Magnetic resonance (MR) images[2] are particularly suitable for brain cancer imaging. Different MR sequences (T2, T2-FLAIR, T1, T1 + gadolinium) highlight different tumor subcomponents, such as edema, necrosis, or contrast-enhancing core.

In recent years, machine learning methods have achieved impressive performance in a large variety of image recognition tasks. Most of the recent state-of-the-art segmentation methods are based on convolutional neural networks (CNNs).[3,4] CNNs have the considerable advantage of automatically learning relevant image features. This ability is particularly important for the tumor segmentation task. CNN-based methods[5–8] have obtained the best performances on the four last editions of the Multimodal Brain Tumor Segmentation Challenge (BRATS).[9,10]

Most of the segmentation methods based on machine learning rely uniquely on manually segmented images. The cost of this annotation is particularly high in medical imaging, where manual segmentation is not only time-consuming but also requires high medical competences. Image intensity of cancerous tissues in MRI or CT scans is often similar to one of the surrounding healthy or pathological tissues, making the exact tumor delineation difficult and subjective. In the case of brain tumors, according to Ref. 9, the inter-rater overlap of expert segmentations is between 0.74 and 0.85 in terms of Dice coefficient. For these reasons, high-quality manual tumor segmentations are generally available in very limited numbers. Segmentation approaches able to exploit images with weaker forms of annotations are therefore of particular interest.

In this paper, we assume that the training dataset contains two types of images: fully annotated (with provided ground truth segmentation) and weakly annotated, with an image-level label indicating the presence or absence of a tumor tissue within the image (Fig. 1). We refer to this setting as "mixed supervision." The latter type of annotations can be obtained at a substantially lower cost as it is less time-consuming, potentially requires less medical expertise, and can be obtained without the use of a dedicated software.

We introduce a CNN-based segmentation model, which can be trained using weakly annotated images in addition to fully annotated images. We propose to extend segmentation networks, such as U-Net,[11] with an additional branch, performing image-level classification. The model is trained jointly for both tasks on fully annotated and weakly annotated images. The goal is to exploit the representation learning ability of CNNs to learn from weakly annotated images while supervising the training using fully annotated images to learn features relevant for the segmentation task. Our approach differs from the standard

*Address all correspondence to Pawel Mlynarski, E-mail: pawel.mlynarski@inria.fr

2329-4302/2019/$28.00 © 2019 SPIE

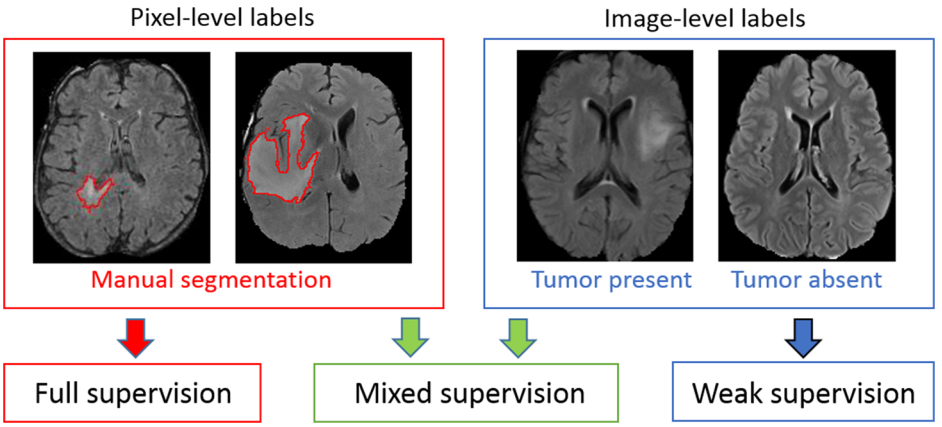

Fig. 1 Different levels of supervision for training of segmentation models. Standard models are trained on fully annotated images only, with pixel-level labels. Weakly supervised approaches aim to train models using only weakly annotated images, e.g., with image-level labels. Our model is trained with a mixed supervision, exploiting both types of training images.

semisupervised learning as we consider weakly annotated data instead of totally unlabeled data. To the best of our knowledge, we are the first to combine pixel-level and image-level labels for training of models for tumor segmentation.

We perform a series of cross-validated tests on the challenging task of segmentation of gliomas in MR images from the BRATS 2018 Challenge. We evaluate our model both for binary and multiclass segmentation using a variable number of ground truth segmentations available for training. Since all three-dimensional (3-D) images from the BRATS 2018 contain brain tumors, we focus on the two-dimensional (2-D) problem of tumor segmentation in axial slices of an MRI and we assume slice-level labels for weakly annotated images. Using approximately 220 MRI with slice-level labels and a varying number of fully annotated MRI, we show that our approach significantly improves the segmentation accuracy when the number of fully annotated cases is limited.

2 Related Work

In the literature, there are several works related to weakly supervised and semisupervised learning for object segmentation or detection. Most of the related works were applied to natural images.

The first group of weakly supervised methods aims to localize objects using only weakly annotated images for training. When only image-level labels are available, one approach is to design a neural network, which outputs two feature maps per class (interpreted as "heat maps" of the class), which are then pooled to obtain an image-level classification score penalized during the training.[12–16] At test time, these heat maps are used for detection (determining a bounding box of the object) or segmentation. To guide the training process, some works use self-generated spatial priors[13–15] or inconsistency measures[16] in the loss function. To obtain an image-level score, in Refs. 12 and 15, global maximum pooling is used. Application of the maximum function on large feature maps may cause optimization problems as training of neural networks is based on the computation of gradients.[17] LogSumExp approximation of the maximum[18] is therefore used in the works[13,14] to partially limit this problem. Average pooling on small feature maps was used by Wang et al.[16] for the problem of detection of lung nodules.

Dubost et al.[19] propose to extend a network similar to 3-D U-Net with a subnetwork performing image-level regression of the number of present lesions. The model is trained using only image-level labels (lesion counts) and the weights learned by the regression subnetwork are used during the test phase to construct heat maps of lesions. Detection of lesions is obtained by thresholding of the heat maps. The common point with our model is the extension of U-Net with a subnetwork performing an image-level task. One of the key differences is that our model is trained using both image-level and pixel-level labels and has a dedicated segmentation layer (trained with pixelwise labels and producing the final segmentation).

Another type of weakly supervised methods aims to detect objects in natural images based on the classification of image subregions[20,21] using pretrained classification networks such as VGG-Net[22] or AlexNet.[23] In fact, one particularity of natural images is their recursive aspect: one image can correspond to a subpart of another image (e.g., two images of the same object taken from different distances). A classification network trained on a large dataset may, therefore, be used on a subregion of an image to determine if it contains an object of interest.

Pretrained classification networks were also used to detect objects by determining image subregions, whose modification influences the global classification score of a class. Simonyan et al.[24] propose to compute the gradient of the classification score with respect to the intensities of pixels and to threshold it in order to localize the object of interest. However, these partial derivatives represent very weak information for tumor segmentation, which requires a complex analysis of the spatial context. The method proposed in Ref. 25 is based on replacing image subregions by the mean value to measure the drop of the classification score.

Overall, the reported segmentation performances of weakly supervised methods are considerably lower than the ones obtained by semisupervised and supervised approaches. In absence of pixel-level labels, a model may learn irrelevant features due, for example, to co-occurrences of objects or image acquisition differences in the case of multicenter medical data. Despite the cost of manual segmentation, at least few fully annotated images can still be obtained in many cases.

In standard semisupervised learning[26] for classification, the training data is composed of both labeled samples and unlabeled

samples. Unlabeled samples can be used to enforce the model to satisfy some properties on relations between labels and the feature space. Common properties include smoothness (points close in the feature space should be close in the target space), clustering (labels form clusters in the feature space), and low-density separation (decision boundaries should be in low-density regions of the feature space). Semisupervised learning based on these properties can be performed by graph-based methods such as the recent work of Kamnitsas et al.[27] The main idea of such methods is to propagate labels in a fully connected graph, whose nodes are samples (labeled and unlabeled) and whose edges are weighted by similarities between samples. The use of graph-based semisupervised methods is difficult for segmentation, in particular, because it implies computation of similarity metrics between samples, whereas every single image is generally composed of millions of samples (pixels or voxels).

Relatively, few works were proposed for semisupervised learning for image segmentation. Some semisupervised approaches are based on self-training, i.e., training of a machine learning model on self-generated labels. Iterative algorithms similar to expectation–maximization[28] were proposed for natural images[29] and medical images.[30] Recently, Hung et al.[31] proposed a method based on generative adversarial networks,[32] where the generator network performs image segmentation and the discriminator network tries to determine if a segmentation corresponds to the ground truth or the segmentation produced by the generator. The discriminator network is used to produce confidence maps for self-training. The approaches based on self-training have the drawback of learning on uncertain labels (produced by the model itself) and training of such models is difficult.

Other approaches assume mixed levels of supervision similar to our approach. Hong et al.[33,34] proposed decoupled classification and segmentation, an approach for segmentation of objects in natural images based on a two-step training with a varying level of supervision. This architecture is composed of two separate networks trained sequentially, one performing image-level classification and used as encoder, and the other one taking as input small feature maps extracted from the encoder and performing segmentation. An important drawback of such design, in the case of tumor segmentation, is that the segmentation network does not take as input the original image and can, therefore, miss important details of the image (e.g., small tumors).

Our approach is related to multitask learning.[35] In our case, the goal of training for two tasks (segmentation and classification) is to exploit all the available labels and to guide the training process to learn relevant features. The approach closest to ours is the one of Shah et al.[36] In this work, the authors consider three types of annotations: segmentations, bounding boxes, and seed points at the borders of objects. A neural network is trained using these three types of training data. In our work, we exploit the use of a significantly weaker form of annotations: image-level labels.

3 Joint Classification and Segmentation with Convolutional Neural Networks

3.1 *Deep Learning Model for Binary Segmentation*

We designed a deep learning model, which aims to take advantage of all available voxelwise and image-level annotations. We propose to extend a segmentation CNN with an additional subnetwork performing image-level classification and to train the model for the two tasks jointly. Most of the layers are shared between the classification and segmentation subnetworks to transfer the information between the two subnetworks. In this paper, we present the 2-D version of our model, which can be used on different types of medical images, such as slices of a CT scan or a multisequence MRI.

The proposed network takes as input a multimodal image of dimensions 300×300 and extends U-Net,[11] which is currently one of the most used architectures for segmentation tasks in medical imaging. The different image modalities (e.g., sequences of an MRI) correspond to channels of the data layer and are the input of the first convolutional layer of the network (as in most of the currently used CNNs for image segmentation). U-Net is composed of an encoder part and a decoder part, which are connected by concatenations between layers at the same scale, to combine low-level and local features with high-level and global features. This design is well suited for the tumor segmentation task since the classification of a voxel as tumor requires comparison of its value with its close neighborhood but also taking into account a large spatial context. The last convolutional layer of U-Net produces pixelwise classification scores, which are normalized by softmax function during the training phase. We apply batch normalization[37] in all convolutional layers except the final layer.

We propose to add an additional branch to the network, performing image-level classification (Fig. 2) to exploit the information contained in weakly annotated images during the training. This classification branch takes as input the second to last convolutional layer of U-Net (representing rich information extracted from a local and a long-range spatial context) and is composed of one mean-pooling, one convolutional layer, and seven fully connected layers.

The goals of taking a layer from the final part of U-Net as input of the classification branch are to guide the image-level classification task and to force the major part of the segmentation network to take into account weakly annotated images. This also helps the optimization process by taking advantage of the connectivity of layers in U-Net, helping the flow of gradients of the loss function during the training (in particular, note the connection between the first part and the last part of U-Net).

The second to last layer of the segmentation network outputs 64 feature maps of size 101×101 from which the classification branch has to output two global (image-level) classification scores (tumor absent/tumor present). We first reduce the size of these feature maps by applying a mean-pooling with kernels of size 8×8 and the stride of 8×8. We use the mean pooling rather than max-pooling to avoid information loss and optimization problems. One convolutional layer, with ReLU activation and kernels of size 3×3, is then added to reduce the number of feature maps from 64 to 32. The resulting 32 feature maps of size 11×11 are the input of the first fully connected layer of the classification branch.

According to our experiments, a relatively deep architecture of the classification branch with a limited number of parameters and a skip-connection between layers yields the best performance. This observation is in agreement with current common designs of neural networks. Deep networks have the capacity to learn more complex features due to applied nonlinearities. The connectivity between layers at different depths helps the optimization process (e.g., Res-Net[38]). In our case, we use seven fully connected layers with ReLU activations (except the final layer) and we concatenate the outputs of the first and the fifth fully

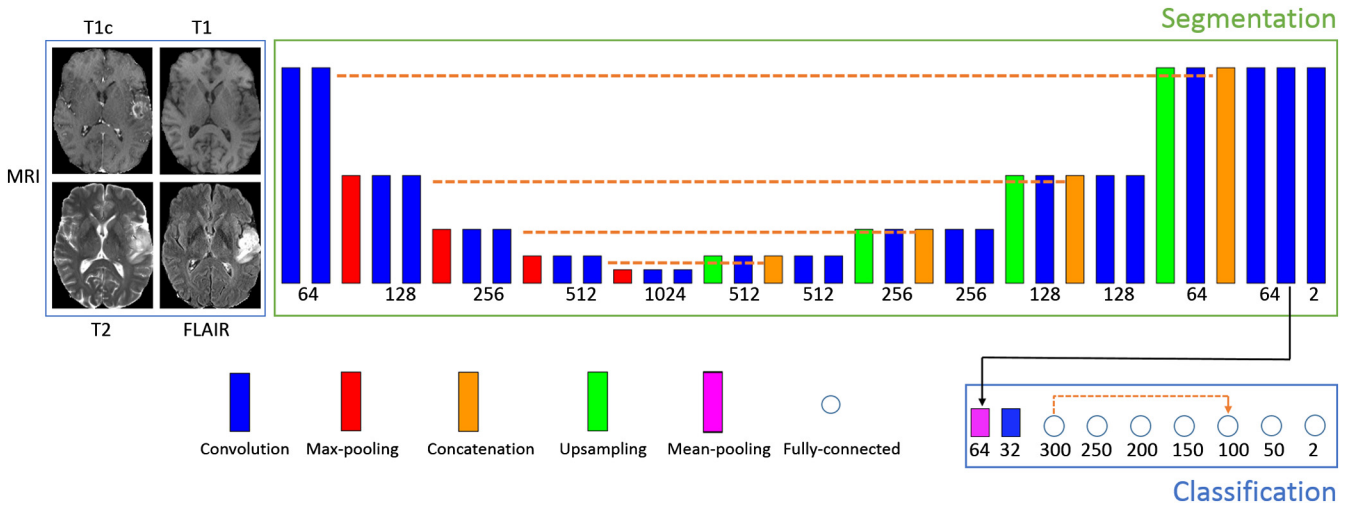

Fig. 2 Architecture of our model for binary segmentation. The numbers of outputs are specified below boxes representing layers. The height of rectangles represents the scale (increasing with pooling operations). The dashed lines represent concatenation operations. The proposed architecture is an extended version of U-Net, with a subnetwork performing image-level classification. Training of the model corresponds to a joint minimization of two loss functions related, respectively, to segmentation and image-level classification tasks.

connected layer. The role of this concatenation is similar to the one connecting the first and the last sequence of convolutional layers in U-Net. The concatenation is used before the second to last layer in order to have one layer to process the mixed information (concatenation of two layers) before the final decision in the seventh fully connected layer. We use only one concatenation as the subnetwork is composed of only a few layers while concatenations increase the number of parameters in the network. The last fully connected layer outputs image-level classification scores (tumor tissue absent or present).

The model is trained on both fully annotated and weakly annotated images for the two tasks jointly (segmentation and classification). We can distinguish between three types of training images. First, images containing a tumor and with provided ground truth segmentation are the most costly ones. The second type corresponds to images that do not contain tumor, which implies that none of their pixels corresponds to a tumor. In this case, the ground truth segmentation is simply the zero matrix. The only problematic case is the third one, when the image is labeled as containing a tumor but without provided segmentation.

To train our model, we propose to form training batches containing the three mentioned types of images: k positive cases (containing a tumor) with provided segmentation, m negative cases, and n positive cases without provided segmentation.

Given a training batch b and the network parameters θ, we use a weighted pixelwise cross-entropy loss on images of types 1 and 2: $\text{Loss}_s^b(\theta) = -\sum_{i=1}^{k+m}\sum_{(x,y)} w_{(x,y)}^i \log[p_{i,(x,y)}^l(\theta)]$, where $p_{i,(x,y)}^l$ is the classification score given by the network to the ground truth label for pixel (x,y) of the i'th image of the batch and $w_{(x,y)}^i$ is the weight given to this pixel. The weights are used to limit the effect of class imbalance since tumor pixels represent a small portion of the image. Weights of pixels are set automatically according to the composition of the training batch (number of pixels of each class) so that pixels associated with healthy tissues have a total weight of t_0 in the loss function and the pixels of the tumor class have a total weight of t_1, where t_0 and t_1 are target weights fixed manually. It means that if the training batch contains N_t pixels labeled as tumor, then each tumor pixel has a weight of t_1/N_1 (the pixelwise weight is high when the number of tumor pixels is low). This type of loss function was used in our previous work.[39]

The classification loss is a standard cross-entropy loss on all images of the training batch: $\text{Loss}_c = -\frac{1}{k+m+n}\sum_{i=1}^{k+m+n}\log[p_i^l(\theta)]$, where p_i^l is the global classification score given by the network to the ground truth global label for the i'th image of the batch. In particular, fully annotated images are also used for training of the classification branch in order to transfer the knowledge from the segmentation task to the image-level classification. We do not apply weights on the classification loss as image-level labels are balanced through the sampling of training batches (having a fixed number of nontumor images).

Since both segmentation and classification losses are normalized, we define the total loss as a convex combination of the classification and segmentation losses: $\text{Loss} = a * \text{Loss}_s + (1-a) * \text{Loss}_c$.

We train our model with a variant of stochastic gradient descent (SGD) with momentum,[40] used also in our previous work.[39] The main differences with the standard SGD are to divide the gradient by its norm and to compute gradients on several training batches in each iteration to take into account many training examples while bypassing GPU memory constraints.

3.2 *Extension to the Multiclass Problem*

We extend our model to the multiclass case, where each pixel has to be labeled with one of the K classes, such as the four ones considered in the BRATS Challenge (nontumor, contrast-enhancing core, edema, nonenhancing core). We now assume that image-level labels are provided for each class (absent/present in the image).

Extension of the segmentation subnetwork to the multiclass problem is straightforward by changing the number of final feature maps to match the number of classes. However, image-level labels are not exclusive, i.e., an image may contain several tumor subclasses. For this reason, we propose to consider one image-level classification output per tumor subclass, indicating the absence or presence of the given subclass.

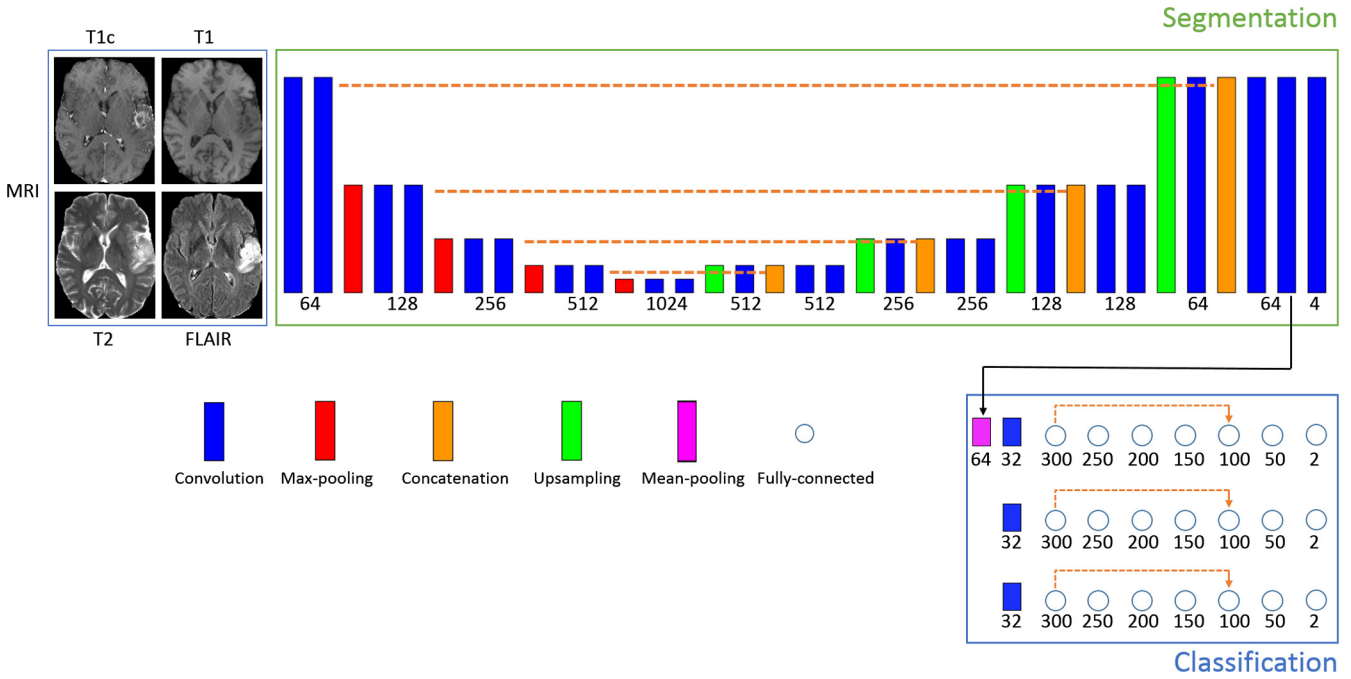

Fig. 3 Extension of our model to the multiclass problem. The number of final feature maps of the segmentation subnetwork is equal to the number of classes (four in our case). As image-level labels (class present/absent) are not exclusive, we consider one classification branch per tumor subclass.

According to our experiments, better performances are obtained when each subclass has its dedicated entire classification branch (Fig. 3). A possible reason is that the image-level classification of tumor subclasses is a challenging task requiring a sufficient number of dedicated parameters.

Training batches are sampled similarly to the binary case, however, each tumor subclass has to be present at least once in each training batch. In our implementation, we store lists of paths of images containing tumor subclasses to sample from these lists during the training of the model.

In the segmentation loss, we empirically fix the following target weights for the four classes (nontumor, nonenhancing tumor core, edema, enhancing-core): $t_0 = 0.7$, $t_1 = 0.1$, $t_2 = 0.1$, and $t_3 = 0.1$ (all tumor subclasses have equal weight in the loss function). The loss associated with each classification branch is the same as in the binary case and the total classification loss is the average across all classification branches.

4 Experiments

4.1 *Data*

We evaluate our method on the challenging task of brain tumor segmentation in multisequence MR scans, using the "training" dataset of the BRATS 2018 Challenge. It contains 285 multisequence MRI of patients diagnosed with low-grade gliomas or high-grade gliomas. For each patient, manual ground truth segmentation is provided. In each case, four MR sequences are available (Fig. 4): T1, T1 + gadolinium, T2, and fluid-attenuated inversion recovery (FLAIR). Preprocessing performed by the organizers includes skull-stripping, resampling to 1 mm^3 resolution, and registration of images to a common brain atlas. The resulting volumes are of size $240 \times 240 \times 155$. The images were acquired in 19 different imaging centers. In order to normalize image intensities, each image is divided by the median of nonzero voxels (which is supposed to be less affected by the tumor zone than the mean) and multiplied the image by a fixed constant.

Each voxel is labeled with one of the following classes: nontumor (class 0), contrast-enhancing core (class 3), nonenhancing core (class 1), and edema (class 2). The benchmark of the challenge groups classes in three regions: "whole tumor" (formed by all tumor subclasses), "tumor core" (classes 1 and 3, corresponding to the visible tumor mass), and "enhancing core" (class 3).

Given that all 3-D images of the database contain tumors (no negative cases to train a 3-D classification network), we consider the 2-D problem of tumor segmentation in axial slices of the brain.

4.2 *Test Setting*

The goal of our experiments is to compare our approach with the standard supervised learning. In each of the performed tests, our model is trained on fully annotated and weakly annotated images and is compared with the standard U-Net trained on fully annotated images only. The goal is to compare our model with a commonly used segmentation model on a publicly available database.

We consider three different training scenarios, with a varying number of patients for which we assume a provided manual tumor segmentation. In each scenario, we perform a fivefold cross-validation. In each fold, 57 patients are used for the test and 228 patients are used for training. Among the 228 training images, few cases are assumed to be fully annotated and the remaining ones are considered to be weakly annotated with slice-level labels. The fully annotated images are different in each fold. If the 3-D volumes are numbered from 0 to 284, then in k'th fold, the test images correspond to the interval $[(k-1) \times 57, k \times 57 - 1]$, the next few images correspond to fully annotated images and the remaining ones are considered as weakly annotated (the folds are generated in a circular way). In the following, FA denotes the number of fully annotated cases

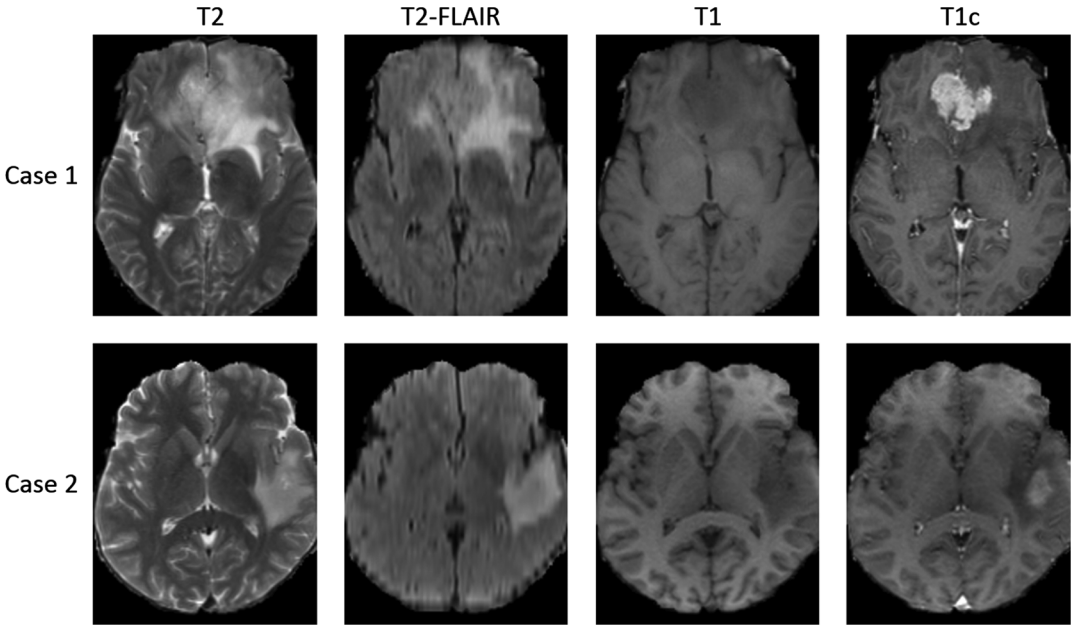

Fig. 4 Examples of multisequence MRI from the BRATS 2018 database. While T2 and T2-FLAIR highlight the edema induced by the tumor, T1 is suitable for determining the tumor core. In particular, T1 acquired after injection of a contrast product (T1c) highlights the tumor angiogenesis, indicating the presence of highly proliferative cancer cells.

and WA denotes the number of weakly annotated cases (with slice-level labels). In particular, note that the split training/test is on 3-D MRIs, i.e., the different slices of the same patient are always in the same set (training or test).

In the first training scenario, five patients are assumed to be provided with manual segmentation and 223 patients have slice-level labels. In the second and third scenarios, the numbers of fully annotated cases are, respectively, 15 and 30 and the numbers of weakly annotated images are, therefore, respectively, 213 and 198. The three training scenarios are independent, i.e., folds are regenerated randomly (the list of all images is permuted randomly and the folds are generated). In fact, results are likely to depend not only on the number of fully annotated images but also on qualitative factors (for example, the few fully annotated images may correspond to atypical cases), and the goal is to test the method in various settings. Overall, our approach is compared to the standard supervised learning on 60 tests (fivefold cross-validation, three independent training scenarios, three binary problems, and one multiclass problem).

We evaluate our method on both binary segmentation problems (separately for each of three tumor regions considered in the challenge) and the end-to-end multiclass segmentation problem. In each binary case, the model is trained for segmentation and classification of one tumor region (whole tumor, tumor core, or enhancing core).

Segmentation performance is expressed in terms of Dice score quantifying the overlap between the ground truth (Y) and the output of a model ($\tilde{Y}$):

$$\mathrm{DSC}(\tilde{Y}, Y) = \frac{2|\tilde{Y} \cap Y|}{|\tilde{Y}| + |Y|}. \quad (1)$$

In order to measure the statistical significance of the obtained results, we perform two-tailed and paired t-tests. Pairs of observations correspond to segmentation scores obtained with the standard supervised learning (U-Net trained on fully annotated images) and with our approach. Dice scores for all patients from fivefolds are concatenated to form a set of 285 pairs of observations. The statistical test is performed for each training scenario and for each segmentation task (three binary problems and one multiclass problem). We consider the significance level of 5%.

4.3 *Model Hyperparameters*

4.3.1 *Loss function and training of the model*

The main introduced training hyperparameter is the parameter a corresponding to the trade-off between classification and segmentation losses. We report mean Dice scores obtained with a varying value of the parameter a on a validation set of 57 patients (20% of the database used for testing and 80% used for training) in the case with 5 fully annotated cases and 223 weakly annotated cases. Segmentation accuracy obtained for the whole tumor in the binary case is reported in Fig. 5. The peak of performance is observed for $a = 0.7$ (improvement of approximately 12 points of Dice over the standard supervised learning on this validation set), i.e., for the configuration, where the segmentation loss accounts for 70% of the total loss. With high values of a, the improvement over the standard supervised learning is limited: around 2.5 points of Dice for $a = 0.9$. In fact, setting a high value of a corresponds to giving less importance to the image-level classification task and therefore ignoring weakly annotated images. For too low values of a, segmentation accuracy decreases too, probably because the model focuses on the secondary task of image-level classification. In the end-to-end multiclass case (Fig. 6), lower values of a seem more suitable, possibly because of an increased complexity of the image-level classification task. In all subsequent tests, we fix $a = 0.7$ for binary segmentations problems and $a = 0.3$ for the end-to-end multiclass segmentation.

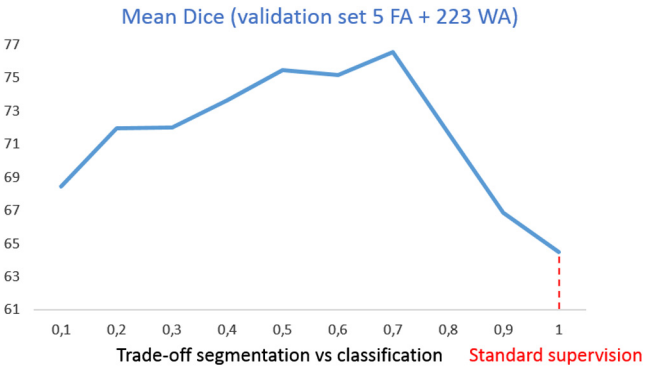

Fig. 5 Mean Dice scores for the "whole tumor" region obtained with a varying value of the parameter "*a*," corresponding to the trade-off between segmentation and image-level classification losses. Segmentation scores are evaluated on a validation set of 57 MRI in the training scenario, where 5 fully annotated MRI and 223 weakly annotated MRI are available for training. The case $a = 1.0$ corresponds to ignoring the classification loss and therefore ignoring weakly annotated images.

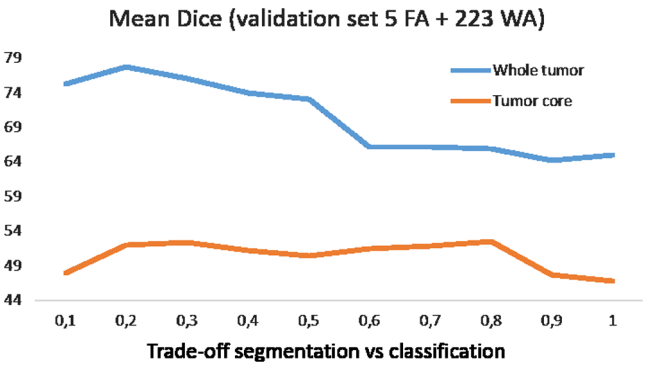

Fig. 6 Mean Dice scores for whole tumor and "tumor core" regions obtained with a varying value of the parameter *a* in the multiclass case. Segmentation scores are evaluated on a validation set of 57 MRI in the training scenario, where 5 fully annotated MRI and 223 weakly annotated MRI are available for training. The case $a = 1.0$ corresponds to ignoring the classification loss and weakly annotated images.

Training batches in our experiments contain 10 images, including 8 images with tumors (4 images with provided tumor segmentation and 4 without provided segmentation) and 2 images without tumors. The number of images was fixed to fit in the memory of the used GPUs (Nvidia GeForce GTX 1080 Ti), i.e., to form training batches for which backpropagation can be performed using the memory of the GPU. In each training batch, there are only two images without tumors because most of the pixels of tumor images correspond to nontumor zones.

The parameters t_c, corresponding to target weights of classes in the segmentation loss, were fixed manually. In both binary and multiclass cases, we chose $t_0 = 0.7$, which corresponds to giving a target weight of 70% to nontumor voxels. In fact, tumor pixels represent approximately 1% of pixels of the training batch and, therefore, nontumor pixels account approximately for 99% of nonweighted cross-entropy segmentation loss. With $t_0 = 0.7$, the relative weight of nontumor pixels is therefore decreased compared to the standard, nonweighted cross-entropy while still giving the nontumor class a high weight to avoid oversegmentation. In the multiclass setting, we fixed the same target weight to all three tumor subclasses, i.e., $t_1 = 0.1$, $t_2 = 0.1$, and $t_3 = 0.1$. As a good convergence of the training was obtained in terms of Dice scores of tumor subclasses, we did not further need to optimize these hyperparameters. Moreover, U-Net trained with these weights and using 228 fully annotated images obtained a mean Dice score of almost 0.87 for whole tumor (last row of Table 1), which is a satisfactory performance for a model independently processing axial slices without any postprocessing.

4.3.2 *Model architecture*

One of the most important attributes of our method is the architecture of classification branches extending segmentation networks. We perform experiments to compare our model with alternative types of architectures of classification subnetworks. We report the segmentation accuracy obtained on the previously defined validation set of 57 patients.

In the binary case, we consider two alternative architectures of classification subnetworks. The first one is composed of four fully connected layers having, respectively, 2000, 500, 100, and 2 neurons. It corresponds, therefore, to a shallow variant of the classification subnetwork with a relatively high number of parameters. We name this architecture "shallow" model. The second variant has the same architecture as our model (seven fully connected layers) but with removed concatenation between the first and the fifth fully connected layer. We name this architecture "deep-sequential." The comparison of segmentation accuracy for the whole tumor obtained by these two variants and by our model is reported in Fig. 7. All three models using a mixed level of supervision obtain a better segmentation accuracy than the standard U-Net using five fully annotated images (64.48). Among the three architectures, the shallow variant yields the lowest accuracy (72.29). Our model obtains the highest accuracy (76.56) and performs slightly better than its counterpart with removed concatenation, deep-sequential model (75.78). The improvements over the standard model and the shallow model were found statistically significant (two-tailed and paired *t*-tests).

We also report results obtained with an alternative architecture of the multiclass model. In our model, we considered

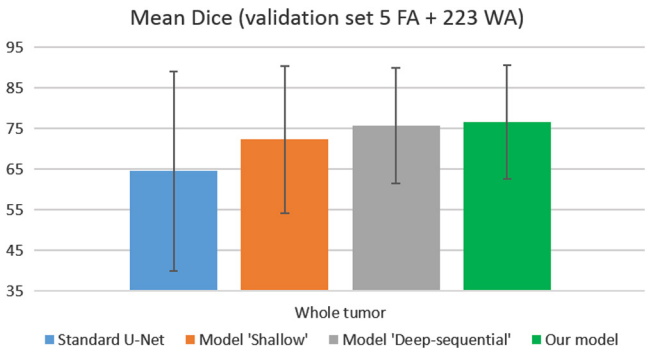

Fig. 7 Mean Dice scores for the whole tumor region obtained by the standard U-Net and by different models using a mixed level of supervision. Standard deviations are represented by error bars. The segmentation scores are evaluated on a validation set of 57 MRI in the training scenario, where 5 fully annotated MRI and 223 weakly annotated MRI are available for training. Our model corresponds to U-Net extended with a classification branch composed of seven fully connected layers and containing one skip-connection.

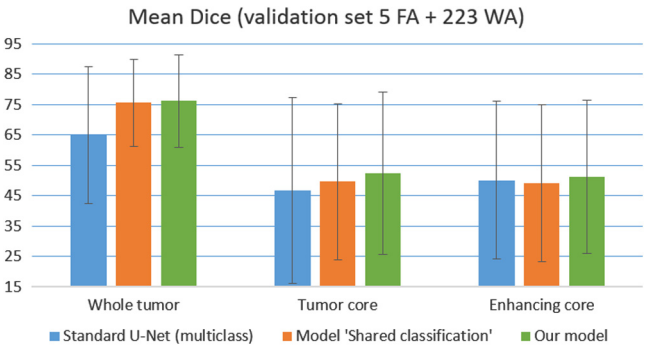

Fig. 8 Mean Dice scores for the whole tumor region obtained by the standard multiclass U-Net and by different multiclass models using mixed level of supervision. The error bars represent standard deviations. The segmentation scores are evaluated on a validation set of 57 MRI in the training scenario, where 5 fully annotated MRI and 223 weakly annotated MRI are available for training. Our model is multiclass U-Net extended with three separate classification branches (for each tumor subclass), each branch having the same architecture as in the binary segmentation/classification problem.

separate classification branches for all tumor subclasses. We consider an alternative architecture, having only one classification branch (with the same architecture as our model for binary segmentation and classification) shared between the three final fully connected layers performing image-level classification. In this configuration, the classification layer of each tumor subclass takes as input the sixth fully connected layer of the shared classification branch. We name this architecture "shared classification." The comparison with our multiclass model (separate classification branches for all tumor subclasses) on the same validation set as previously is reported on Fig. 8. Our model obtains the highest accuracy for the three tumor subregions while the alternative model (shared classification) obtains higher accuracy than the standard multiclass U-Net for whole tumor and tumor core. The improvements of our model over the standard model were found statistically significant for the whole tumor and tumor core regions. The improvements over the alternative model with mixed supervision (shared classification) were not found statistically significant (p-values >0.05).

4.4 *Results*

The main observation is that our model with mixed supervision provides a significant improvement over the standard supervised approach (U-Net trained on fully annotated images) when the number of fully annotated images is limited. In the two first training scenarios (5 FA and 15 FA), our model outperformed the supervised approach on the three binary segmentation problems (Table 1) and in the multiclass setting (Table 2). The largest improvements are in the first scenario (5 FA) for the whole tumor region, where the improvement is of eight points of the mean Dice score in the binary setting and of nine points of Dice in the multiclass setting. Results on different folds of the second scenario (intermediate case, 15 FA) are displayed in Table 3 for the binary problems and in Table 4 for the multiclass problem. Our approach provided an improvement in all folds of the second scenario and for all tumor regions, except one fold for enhancing core in the binary setting. In the third scenario (30 FA + 198 WA), our approach and the standard supervised approach obtained similar performances. Furthermore, we observe that standard deviations are consistently lower with our approach in all

Table 1 Mean Dice scores (fivefold cross-validation, 57 test cases in each fold) in the three binary segmentation problems obtained by the standard supervised approach and by our model trained with mixed supervision. The numbers in brackets denote standard deviations computed on the distribution of Dice scores for all patients of the fivefolds.

	Whole tumor	Tumor core	Enhancing core
Standard supervision 5 FA	70.39 (21.78)	48.14 (28.31)	55.74 (26.73)
Mixed supervision 5 FA + 223 WA	**78.34*** (13.01)	**50.11*** (25.95)	**60.06*** (22.72)
Standard supervision 15 FA	77.91 (16.77)	58.33 (29.00)	62.88 (25.80)
Mixed supervision 15 FA + 213 WA	**80.92*** (11.17)	**63.23*** (26.40)	**66.61*** (23.12)
Standard supervision 30 FA	**83.95** (11.84)	66.17 (25.61)	**69.15** (23.51)
Mixed supervision 30 FA + 198 WA	83.84 (9.68)	**68.30*** (23.73)	67.18 (21.69)
Standard supervision 228 FA	86.80 (8.47)	77.09 (18.58)	72.20 (19.11)

Note: The bold values highlight the higher accuracy in a given test (comparison of the two models).
*Statistically significant improvements (p-value <0.05) provided by our method compared to the standard supervised learning.

Table 2 Mean Dice scores (fivefold cross-validation, 57 test cases in each fold) obtained by the standard supervised approach and by our model in the multiclass setting. The numbers in brackets denote standard deviations computed on the distribution of Dice scores for all patients of the fivefolds.

	Whole tumor	Tumor core	Enhancing core
Standard supervision 5 FA	67.61 (22.24)	51.12 (26.98)	58.15 (24.65)
Mixed supervision 5 FA + 223 WA	**76.64*** (14.14)	**56.30*** (22.65)	**58.19** (23.05)
Standard supervision 15 FA	74.46 (18.04)	59.87 (25.97)	61.85 (24.86)
Mixed supervision 15 FA + 213 WA	**79.39*** (12.99)	**63.91*** (24.72)	**65.71*** (23.07)
Standard supervision 30 FA	81.10 (14.29)	**67.48** (24.78)	**68.67** (22.79)
Mixed supervision 30 FA + 198 WA	**81.23** (10.90)	66.33 (24.12)	67.69 (21.87)
Standard supervision 228 FA	85.67 (9.66)	78.78 (18.31)	74.14 (19.62)

Note: The bold values highlight the higher accuracy in a given test (comparison of the two models).
*Statistically significant improvements (p-value <0.05) provided by our method compared to the standard supervised learning.

Table 3 Results obtained for the three binary problems (whole tumor, tumor core, enhancing core) on different folds in the case with 15 fully annotated images and 213 weakly annotated images. The numbers in brackets denote standard deviations computed on the distribution of Dice scores for all patients.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Total
Standard supervision, whole tumor	76.23 (14.68)	78.15 (19.24)	78.13 (16.88)	77.67 (18.46)	79.35 (13.76)	77.91 (16.77)
Mixed supervision, whole tumor	**82.36** (9.28)	**81.03** (10.21)	**78.96** (12.47)	**79.88** (13.60)	**82.35** (9.16)	**80.92** (11.17)
Standard supervision, tumor core	61.46 (28.94)	61.17 (26.55)	56.68 (27.90)	56.42 (28.63)	55.94 (32.17)	58.33 (29.00)
Mixed supervision, tumor core	**63.15** (25.92)	**66.82** (21.74)	**63.45** (26.73)	**60.83** (27.22)	**61.91** (29.40)	**63.23** (26.40)
Standard supervision, enhancing core	66.33 (24.51)	61.08 (26.49)	57.86 (25.85)	**68.09** (22.40)	61.02 (27.82)	62.88 (25.80)
Mixed supervision, enhancing core	**68.72** (23.66)	**70.65** (17.91)	**60.34** (25.84)	67.55 (20.49)	**65.80** (25.46)	**66.61** (23.12)

Note: The bold values highlight the higher accuracy in a given test (comparison of the two models).

Table 4 Results obtained in the multiclass setting on different folds in the case with 15 fully annotated images and 213 weakly annotated images. The numbers in brackets denote standard deviations computed on the distribution of Dice scores for all patients.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Total
Standard supervision, whole tumor	74.31 (13.78)	78.91 (15.41)	67.57 (23.14)	75.55 (17.59)	75.96 (16.85)	74.46 (18.04)
Mixed supervision, whole tumor	**77.53** (12.81)	**82.20** (9.39)	**73.72** (16.37)	**80.96** (13.40)	**82.55** (9.38)	**79.39** (12.99)
Standard supervision, tumor core	61.17 (23.64)	63.89 (22.79)	55.72 (26.34)	55.36 (28.33)	63.18 (27.06)	59.87 (25.97)
Mixed supervision, tumor core	**62.83** (24.65)	**65.26** (22.63)	**62.23** (25.82)	**61.99** (27.87)	**67.23** (21.74)	**63.91** (24.72)
Standard supervision, enhancing core	66.15 (24.58)	64.83 (23.14)	53.83 (25.52)	61.68 (24.38)	62.77 (24.77)	61.85 (24.86)
Mixed supervision, enhancing core	**68.33** (21.70)	**68.39** (18.55)	**59.51** (26.07)	**68.63** (21.76)	**63.70** (25.14)	**65.71** (23.07)

Note: The bold values highlight the higher accuracy in a given test (comparison of the two models).

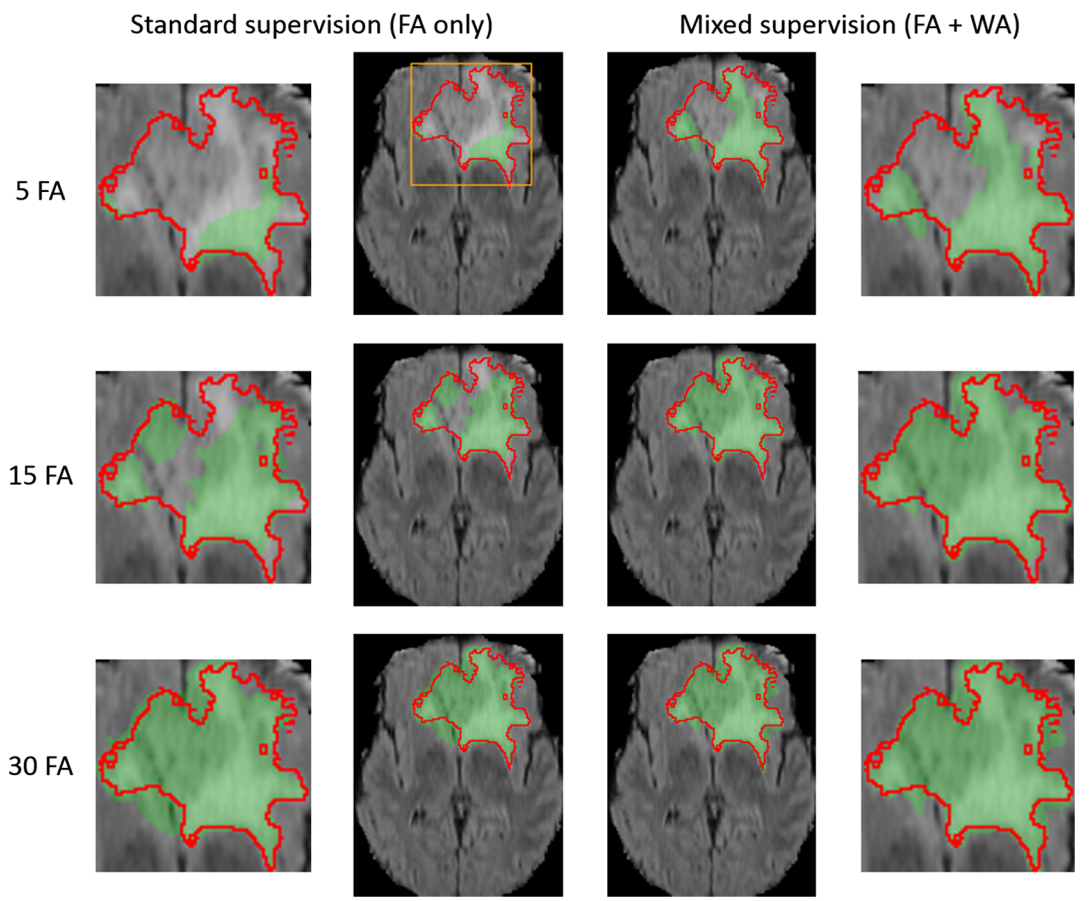

Fig. 9 Comparison of our approach with the standard supervised learning for binary segmentation of the whole tumor region. Each row represents the same test example (first image of Fig. 4) from a different training scenario (5, 15, or 30 fully annotated scans available for training). FA and WA refer, respectively, to the number of fully annotated MRI and weakly annotated MRI (with slice-level labels). The results are displayed on MRI T2-FLAIR sequence. The performance of both models improves with the number of manual segmentations available for training.

training scenarios and for all tumor subregions. The results obtained with mixed supervision are therefore more stable than the ones obtained with the standard supervised learning.

All improvements were found statistically significant for binary segmentation problems. In the multiclass case, all improvements were found statistically significant except for enhancing core in the first training scenario and for whole tumor in the third training scenario.

Qualitative results are displayed in Figs. 9–11. Each figure shows segmentations of one tumor region (whole tumor, tumor core, and enhancing core) produced by models trained with a varying number of fully annotated and weakly annotated images available for training.

Segmentation performance increases quickly with the first fully annotated cases, both for the standard supervised learning and the learning with mixed supervision. For instance, mean Dice score obtained by the supervised approach for whole tumor increases from 70.39, in the case with 5 fully annotated images, to 77.9 in the case with 15 fully annotated images. Our approach using 5 fully annotated images and 223 weakly annotated images obtained a slightly better performance (78.3) than the supervised approach using 15 fully annotated cases (77.9). This result is represented on Fig. 12.

From Fig. 13, we report cross-validated results obtained with a varying number of weakly annotated while images keeping a fixed number of fully annotated images. This complementary experiment is performed for segmentation of whole tumor in the first training scenario (five fully annotated images). We observe that the improvement slows down with the number of added weakly annotated scans. Inclusion of the first 100 weakly annotated MRIs yields an improvement of approximately five points of the cross-validated mean Dice score (from 70.39 to 75.28), whereas the addition of the remaining 123 weakly annotated images improves this score by three points (from 75.28 to 78.34).

Note that each fully annotated case corresponds to a large 3-D volume with voxelwise annotations. Each manually segmented

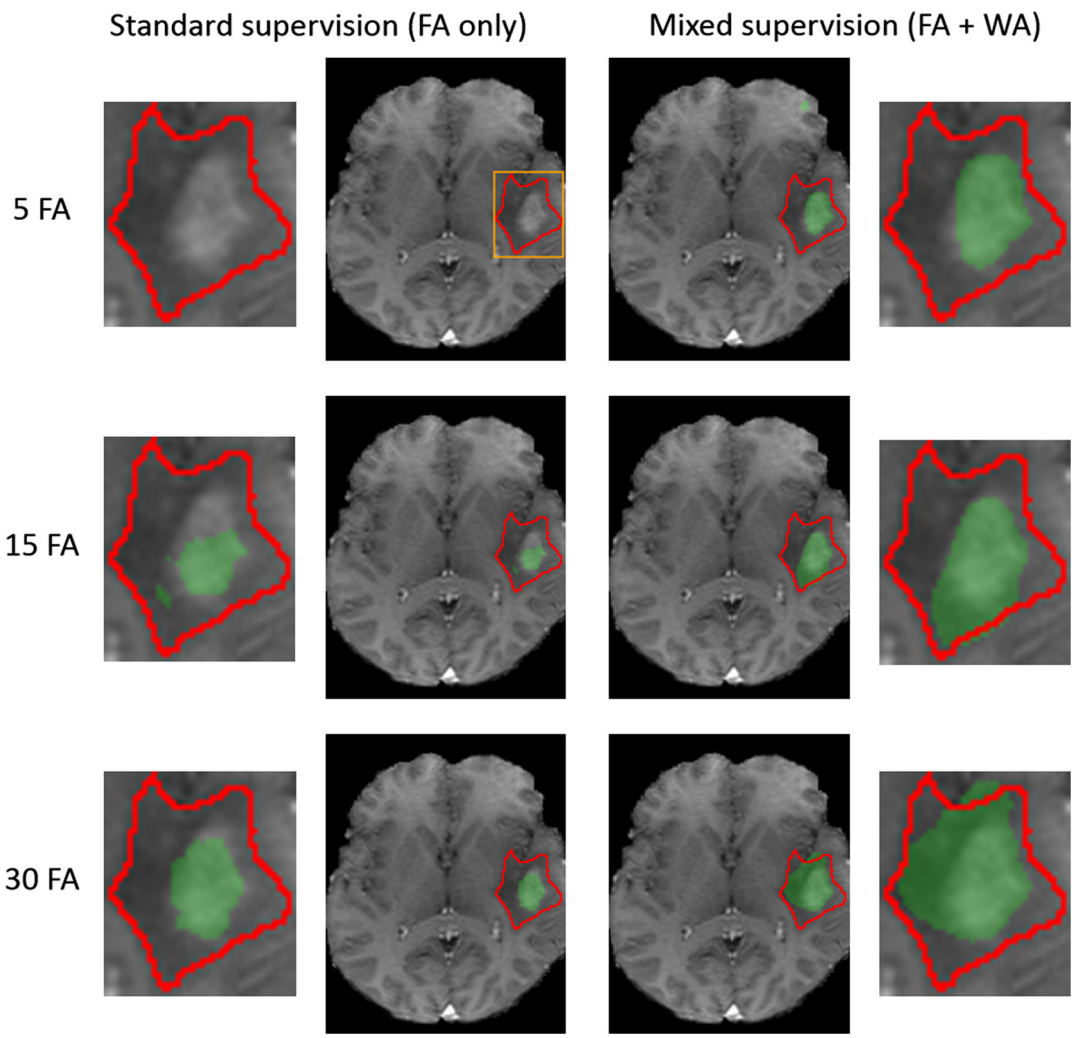

Fig. 10 Comparison of our approach with the standard supervised learning for binary segmentation of the tumor core region (test example corresponding to the bottom image of Fig. 4). Each row corresponds to a different training scenario (5, 15, or 30 fully annotated scans available for training). FA and WA refer to the numbers of fully annotated and weakly annotated scans. The results are displayed on MRI T1 +gadolinium. The observations are similar to the problem of binary segmentation of the whole tumor region. In particular, in the first training scenario, the standard supervised approach does not detect the tumor core zone in contrast to our method.

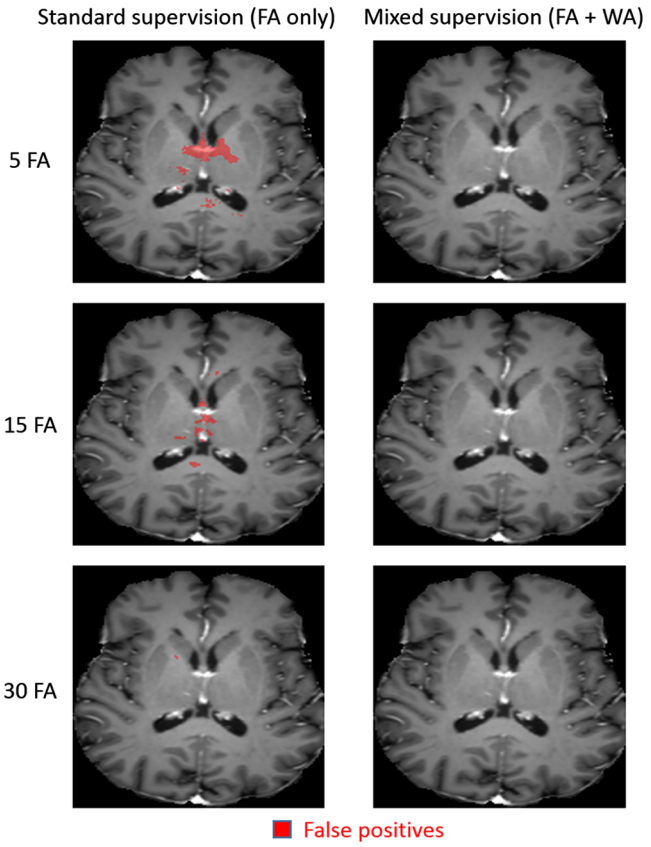

Fig. 11 Comparison of our approach with the standard supervised learning for binary segmentation of the "enhancing core" region. Each row corresponds to a different training scenario (5, 15, or 30 fully annotated scans available for training). FA and WA refer to the numbers of fully annotated and weakly annotated scans. The results are displayed on MRI T1+gadolinium. The example shows false positives obtained by the model trained with standard supervision. The number of false positives decreases with the number of fully annotated images available for training. No false positives are observed for our model trained with mixed supervision, in any of the three training scenarios.

axial slice of size 240×240 corresponds to 57,600 labels, which represents indeed a huge amount of information compared to one global label simply indicating the presence or absence of a tumor tissue within the slice.

In terms of the annotation cost, manual delineation of tumor tissues in one MRI may take about 45 min for an experienced oncologist using a dedicated segmentation tool. Determining the range of axial slices containing tumor tissues may take 1 to 2 min but can be done without specialized software. More importantly, determining global labels may require less medical expertise than performing an exact tumor delineation and can, therefore, be performed by a larger community.

5 Conclusion and Future Work

In this paper, we proposed a deep learning approach for tumor segmentation, which takes advantage of weakly annotated medical images during the training of neural networks, in addition to a small number of manually segmented images. In our approach, we propose to use neural networks producing both voxelwise and image-level outputs. The classification and segmentation subnetworks share most of their layers and are trained jointly using both fully annotated and weakly annotated data. We performed a large number of cross-validated experiments to test our method in both binary and multiclass settings. Our experiments showed that the use of weakly annotated data improves the segmentation performance significantly when the number of manually segmented images is limited. Our model is end-to-end and straightforward to implement with common deep learning libraries, such as Theano[41] or TensorFlow.[42] To encourage other researchers to continue the research in the field, the code of our method will be made publicly available on https://github.com/PawelMlynarski/segmentation_mixed_supervision.

In our paper, we focused on the 2-D segmentation problem, in particular, because all 3-D images from the BRATS 2018 database contain tumors, whereas we also need nontumor images to train the classification part of our model. A practical difficulty of collecting databases containing both tumor and nontumor 3-D scans is the heterogeneity of available imaging modalities. For example, MRI + gadolinium, commonly used for tumor imaging, is generally available for patients with suspected tumors or vascular problems (requiring imaging of blood vessels using a contrast product). In this paper, we chose to focus only on the problem of available ground truth annotations, assuming the availability of the same imaging modalities for all patients, for both supervised learning and learning with mixed supervision. Dealing with the variability of available modalities is a very important problem of medical imaging and is beyond the scope of this paper.

Extension of our model to an end-to-end segmentation of entire 3-D scans could be difficult with the current GPUs because of computational costs of CNNs. One advantage of

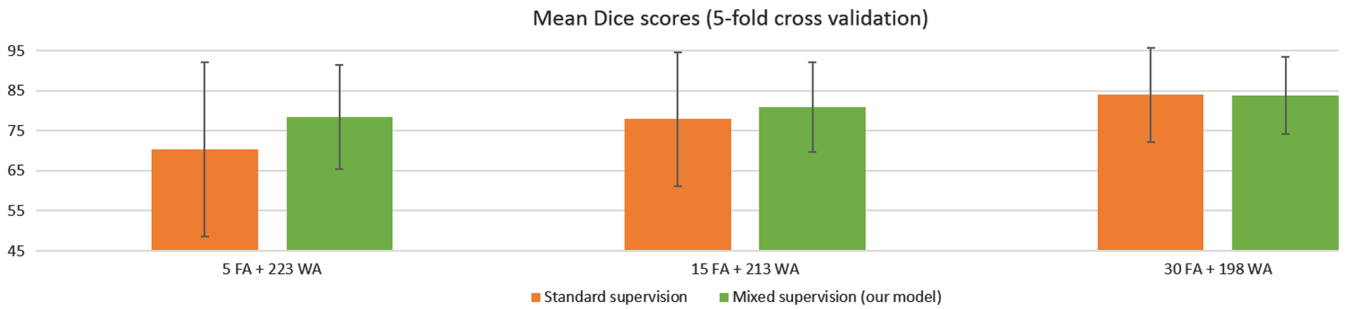

Fig. 12 Illustration of the improvement provided by the mixed supervision for binary segmentation of the whole tumor region (mean Dice scores and their standard deviations). Mixed supervision using 5 fully annotated MRI and 223 weakly annotated MRI obtains a slightly better performance than the standard supervised approach using 15 fully annotated MRI. The improvement provided by the weakly annotated images decreases with the number of available ground truth segmentations.

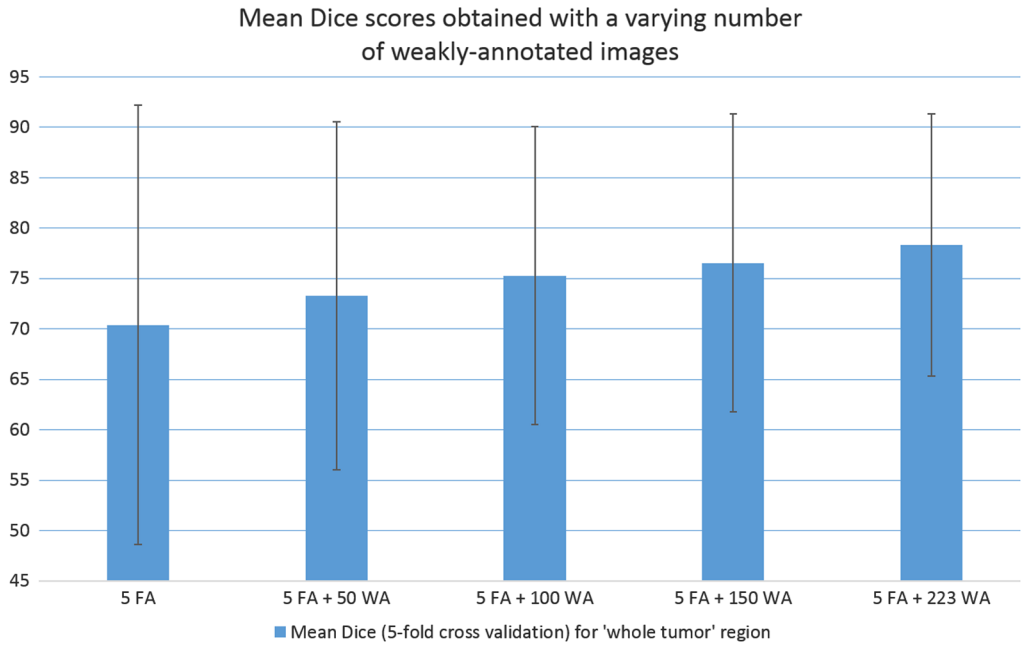

Fig. 13 Mean Dice scores (fivefold cross-validation, 57 test cases in each fold) obtained for binary segmentation of the whole tumor with training on five fully annotated scans and a varying number of weakly annotated scans. The error bars represent standard deviations.

a 3-D model would be to take into account a richer spatial context in the case of MRI or CT scans. Furthermore, volume-level labels require less effort than slice-level labels and would, therefore, be easier to obtain, even if these labels are also less informative. However, 2-D CNNs still perform reasonably well on 3-D scans. As reported in the last row of Table 1, U-Net processing independently axial slices obtains a mean Dice of almost 0.87 for the whole tumor region and of 0.77 for the tumor core region, using 228 fully annotated images (80% of the database of BRATS), without any postprocessing.

In our tests, we used approximately 220 weakly annotated MRI, which is a relatively limited number. An important future step would be to test our method on a database containing a considerably larger number of weakly annotated images (thousands, millions).

Disclosures

The authors have no conflicts of interest to disclose.

Acknowledgments

P.M. is funded by the Microsoft Research-INRIA Joint Center, France. This work was supported by Inria Sophia Antipolis—Méditerranée, "NEF" computation cluster.

References

1. M. L. Goodenberger and R. B. Jenkins, "Genetics of adult glioma," *Cancer Genet.* **205**(12), 613–621 (2012).
2. S. Bauer et al., "A survey of MRI-based medical image analysis for brain tumor studies," *Phys. Med. Biol.* **58**(13), R97 (2013).
3. Y. LeCun et al., "Convolutional networks for images, speech, and time series," *Handb. Brain Theory Neural Networks* **3361**(10), 1995 (1995).
4. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3431–3440 (2015).
5. S. Pereira et al., "Deep convolutional neural networks for the segmentation of gliomas in multi-sequence MRI," *Lect. Notes Comput. Sci.* **9556**, 131–143 (2015).
6. K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.* **36**, 61–78 (2017).
7. K. Kamnitsas et al., "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *Int. MICCAI Brainlesion Workshop*, Springer, pp. 450–462 (2017).
8. G. Wang et al., "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *Int. MICCAI Brainlesion Workshop*, Springer, pp. 178–190 (2017).
9. B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imaging* **34**(10), 1993–2024 (2015).
10. S. Bakas et al., "Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data* **4**, 170117 (2017).
11. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
12. D. Pathak et al., "Fully convolutional multi-class multiple instance learning," arXiv:1412.7144 (2014).
13. P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1713–1721 (2015).
14. F. Saleh et al., "Built-in foreground/background prior for weakly-supervised semantic segmentation," *Lect. Notes Comput. Sci.* **9912**, 413–432 (2016).
15. A. Bearman et al., "What's the point: semantic segmentation with point supervision," *Lect. Notes Comput. Sci.* **9911**, 549–565 (2016).
16. Z. Wang et al., "Automated detection of clinically significant prostate cancer in MP-MRI images based on an end-to-end deep neural network," *IEEE Trans. Med. Imaging* **37**(5), 1127–1139 (2018).
17. S. E. Dreyfus, "Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure," *J. Guidance Control Dyn.* **13**(5), 926–928 (1990).
18. S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge (2004).
19. F. Dubost et al., "Gp-Unet: lesion detection from weak labels with a 3D regression network," in *Int. Conf. Med. Image Comput. and Comput. -Assisted Intervention*, Springer, pp. 214–221 (2017).

20. R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 580–587 (2014).
21. M. Oquab et al., "Is object localization for free?-Weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 685–694 (2015).
22. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2014).
23. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105 (2012).
24. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: visualising image classification models and saliency maps," arXiv:1312.6034 (2013).
25. A. Bergamo et al., "Self-taught object localization with deep networks," in *2016 IEEE Winter Conf. Appl. Comput. Vision (WACV)*, IEEE, pp. 1–9 (2016).
26. V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.* **54**, 280–296 (2019).
27. K. Kamnitsas et al., "Semi-supervised learning via compact latent space clustering," in *Proc. 35th Int. Conf. Mach. Learn.*, PMLR, pp. 2459–2468 (2018).
28. Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imaging* **20**(1), 45–57 (2001).
29. G. Papandreou et al., "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 1742–1750 (2015).
30. M. Rajchl et al., "DeepCut: object segmentation from bounding box annotations using convolutional neural networks," *IEEE Trans. Med. Imaging* **36**(2), 674–683 (2016).
31. W.-C. Hung et al., "Adversarial learning for semi-supervised semantic segmentation," arXiv:1802.07934 (2018).
32. I. Goodfellow et al., "Generative adversarial nets," in *Adv. Neural Inf. Process. Syst.*, pp. 2672–2680 (2014).
33. S. Hong, H. Noh, and B. Han, "Decoupled deep neural network for semi-supervised semantic segmentation," in *Adv. Neural Inf. Process. Syst.*, pp. 1495–1503 (2015).
34. S. Hong et al., "Learning transferrable knowledge for semantic segmentation with deep convolutional neural network," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3204–3212 (2016).
35. T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *Proc. tenth ACM SIGKDD Int. Conf. Knowl. Discovery and Data Min.*, ACM, pp. 109–117 (2004).
36. M. P. Shah, S. Merchant, and S. P. Awate, "MS-Net: mixed-supervision fully-convolutional networks for full-resolution segmentation," *Lect. Notes Comput. Sci.* **11073**, 379–387 (2018).
37. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, PMLR, pp. 448–456 (2015).
38. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
39. P. Mlynarski et al., "3D convolutional neural networks for tumor segmentation using long-range 2D context," *Computerized Med. Imaging and Graphics* **73**, 60–72 (2019).
40. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive Model.* **5**(3), 1 (1988).
41. J. Bergstra et al., "Theano: a CPU and GPU math compiler in Python," in *Proc. 9th Python Sci. Conf.*, pp. 1–7 (2010).
42. M. Abadi et al., "Tensorflow: large-scale machine learning on heterogeneous distributed systems," arXiv:1603.04467 (2016).

Pawel Mlynarski is a PhD student at Inria Sophia Antipolis, France. His research work addresses problems of automatic analysis of medical images in oncology. His thesis focuses on segmentation of brain tumors and organs at risk in the context of radiotherapy planning. He obtained his master's degree of applied mathematics at the École Normale Supérieure (Cachan, France).

Hervé Delingette is a research director at Inria, director of an academy of excellence at the Université Côte d'Azur, a member of the board of the MICCAI Society, and he holds a chair at the new AI institute 3IA Côte d'Azur. He received his engineering degree and PhD from the École Centrale Paris. His research focuses on various aspects of artificial intelligence in medical image analysis, computational physiology, and surgery simulation.

Antonio Criminisi recently left Microsoft Research Cambridge and is now a science manager at Amazon in Cambridge, UK. His interests are in the fields of machine learning, medical image analysis, and computer vision. Antonio received his PhD from the University of Oxford in the year 2000. Antonio has written hundreds of scientific papers and books. He has won multiple prizes and awards, including the prestigious David Marr Best Paper Prize at ICCV 2015 in Chile.

Nicholas Ayache is research director at Inria, head of the Epione research team dedicated to E-patients for E-medicine, and scientific director of the new AI institute 3IA Côte d'Azur. He is a member of the French Academy of Sciences and Academy of Surgery. His current research focuses on AI methods to improve diagnosis, prognosis and therapy from medical images and clinical, biological, behavioral, and environmental data available on the patient.