Context-sensitive Classification Forests for Segmentation of Brain Tumor Tissues

D. Zikic[1], B. Glocker[1], E. Konukoglu[1], J. Shotton[1], A. Criminisi[1],
D. H. Ye[2], C. Demiralp[3],
O. M. Thomas[4,5], T. Das[4], R. Jena[4], S. J. Price[4,6]

[1]Microsoft Research Cambridge, UK
[2]Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA
[3]Brown University, Providence, RI, USA
[4]Cambridge University Hospitals, Cambridge, UK
[5]Department of Radiology, Cambridge University, UK
[6]Department of Clinical Neurosciences, Cambridge University, UK

Abstract. We describe our submission to the Brain Tumor Segmentation Challenge (BraTS) at MICCAI 2012, which is based on our method for tissue-specific segmentation of high-grade brain tumors [3].
The main idea is to cast the segmentation as a classification task, and use the discriminative power of context information. We realize this idea by equipping a classification forest (CF) with spatially non-local features to represent the data, and by providing the CF with initial probability estimates for the single tissue classes as additional input (along-side the MRI channels). The initial probabilities are patient-specific, and computed at test time based on a learned model of intensity. Through the combination of the initial probabilities and the non-local features, our approach is able to capture the context information for each data point. Our method is fully automatic, with segmentation run times in the range of 1-2 minutes per patient. We evaluate the submission by cross-validation on the real and synthetic, high- and low-grade tumor BraTS data sets.

1 Introduction

This BraTS submission is based on our work presented in [3]. We approach the segmentation of the tumor tissues as a classification problem, where each point in the brain is assigned a certain tissue class. The basic building block of our approach is a standard classification forest (CF), which is a discriminative multi-class classification method. Classification forests allow us to describe brain points to be classified by very high-dimensional features, which are able to capture information about the spatial context. These features are based on the multi-channel intensities and are spatially non-local. Furthermore, we augment the input data to the classification forest with initial tissue probabilities, which are estimated as posterior probabilities resulting from a generative intensity-based model, parametrized by Guassian Mixture models (GMM). Together with the

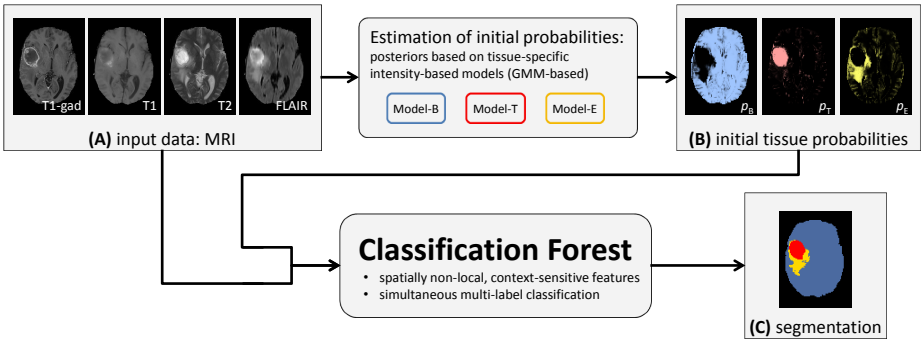


Fig. 1: Schematic Method Overview: Based on the input data **(A)**, we first roughly estimate the initial probabilities for the single tissues **(B)**, based on the local intensity information alone. In a second step, we combine the initial probabilities **(B)** with the input data from **(A)**, resulting in a higher-dimensional multi-channel input for the classification forest. The forest computes the segmentation **(C)** by a simultaneous multi-label classification, based on non-local and context-sensitive features.

context-sensitive features, the initial probabilities as additional input increase the amount of context information and thus improve the classification results.

In this paper, we focus on describing our BraTS submission. For more details on motivation for our approach and relation to previous work, please see [3].

2 Method: Context-sensitive Classification Forests

An overview of our approach is given in Figure 1. We use a standard classification forest [1], based on spatially non-local features, and combine it with initial probability estimates for the individual tissue classes. The initial tissue probabilities are based on local intensity information alone. They are estimated with a parametric GMM-based model, as described in Section 2.1. The initial probabilities are then used as additional input channels for the forest, together with the MR image data I.

In Section 2.2 we give a brief description of classification forests. The types of the context-sensitive features are described in Section 2.3.

We classify three classes $\mathcal{C} = \{\mathrm{B}, \mathrm{T}, \mathrm{E}\}$ for background (B), tumor (T), and edema (E). The MR input data is denoted by $I = (I_{\mathrm{T1C}}, I_{\mathrm{T1}}, I_{\mathrm{T2}}, I_{\mathrm{FLAIR}})$.

2.1 Estimating Initial Tissue Probabilities

As the first step of our approach, we estimate the initial class probabilities for a given patient based on the intensity representation in the MRI input data.

The initial probabilities are computed as posterior probabilities based on the likelihoods obtained by training a set of GMMs on the training data. For each

class $c \in \mathcal{C}$, we train a single GMM, which captures the likelihood $p_{\text{lik}}(\mathbf{i}|c)$ of the multi-dimensional intensity $\mathbf{i} \in \mathbb{R}^4$ for the class c. With the trained likelihood p_{lik}, for a given test patient data set I, the GMM-based posterior probability $p^{\text{GMM}}(c|\mathbf{p})$ for the class c is estimated for each point $\mathbf{p} \in \mathbb{R}^3$ by

$$p^{\text{GMM}}(c|\mathbf{p}) = \frac{p_{\text{lik}}(I(\mathbf{p})|c)\, p(c)}{\sum_{c_j} p_{\text{lik}}(I(\mathbf{p})|c_j)\, p(c_j)} \ , \tag{1}$$

with $p(c)$ denoting the prior probability for the class c, computed as a normalized empirical histogram. We can now use the posterior probabilities directly as input for the classification forests, in addition to the multi-channel MR data I. So now, with $p_c^{\text{GMM}}(\mathbf{p}) := p^{\text{GMM}}(c|\mathbf{p})$, our data for one patient consists of the following channels

$$C = (I_{\text{T1-gad}}, I_{\text{T1}}, I_{\text{T2}}, I_{\text{FLAIR}}, p_{\text{AC}}^{\text{GMM}}, p_{\text{NC}}^{\text{GMM}}, p_{\text{E}}^{\text{GMM}}, p_{\text{B}}^{\text{GMM}}) \ . \tag{2}$$

For simplicity, we will denote single channels by C_j.

2.2 Classification Forests

We employ a classification forest (CF) to determine a class $c \in \mathcal{C}$ for a given spatial input point $\mathbf{p} \in \Omega$ from a spatial domain Ω of the patient. Our classification forest operates on the representation of a spatial point $\mathbf{p}$ by a corresponding feature vector $x(\mathbf{p}, C)$, which is based on spatially non-local information from the channels C. CFs are ensembles of (binary) classification trees, indexed and referred to by $t \in [1, T]$. As a supervised method, CFs operate in two stages: training and testing.

During training, each tree t learns a weak class predictor $p_t(c|x(\mathbf{p}, C))$. The input training data set is $\{(x(\mathbf{p}, C^{(k)}), c^{(k)}(\mathbf{p})) : \mathbf{p} \in \Omega^{(k)}\}$, that is, the feature representations of all spatial points $\mathbf{p} \in \Omega^{(k)}$, in all training patient data sets k, and the corresponding manual labels $c^{(k)}(\mathbf{p})$.

To simplify notation, we will refer to a data point at $\mathbf{p}$ by its feature representation x. The set of all data points shall be X.

In a classification tree, each node i contains a set of training examples X_i, and a class predictor $p_t^i(c|x)$, which is the probability corresponding to the fraction of points with class c in X_i (normalized empirical histogram). Starting with the complete training data set X at the root, the training is performed by successively splitting the training examples at every node based on their feature representation, and assigning the partitions X_{L} and X_{R} to the left and right child node. At each node, a number of splits along randomly chosen dimensions of the feature space is considered, and the one maximizing the Information Gain is applied (i.e., an axis-aligned hyperplane is used in the split function). Tree growing is stopped at a certain tree depth D.

At testing, a data point x to be classified is pushed through each tree t, by applying the learned split functions. Upon arriving at a leaf node l, the leaf probability is used as the tree probability, i.e. $p_t(c|x) = p_t^l(c|x)$. The overall probability

is computed as the average of tree probabilities, i.e. $p(c|x) = \frac{1}{T}\sum_{t=1}^{T} p_t(c|x)$. The actual class estimate $\hat{c}$ is chosen as the class with the highest probability, i.e. $\hat{c} = \arg\max_c p(c|x)$.

For more details on classification forests, see for example [1].

2.3 Context-sensitive Feature Types

We employ three features types, which are intensity-based and parametrized. Features of these types describe a point to be labeled based on its non-local neighborhood, such that they are context-sensitive. The first two of these feature types are quite generic, while the third one is designed with the intuition of detecting structure changes. We denote the parametrized feature types by $x^{\text{type}}_{\text{params}}$. Each combination of type and parameter settings generates one dimension in the feature space, that is $x_i = x^{\text{type}_i}_{\text{params}_i}$. Theoretically, the number of possible combinations of type and parameter settings is infinite, and even with exhaustive discrete sampling it remains substantial. In practice, a certain predefined number d' of combinations of feature types and parameter settings is randomly drawn for training. In our experiments, we use $d' = 2000$.

We use the following notation: Again, $\mathbf{p}$ is a spatial point, to be assigned a class, and C_j is an input channel. $R^s_j(\mathbf{p})$ denotes an $\mathbf{p}$-centered and axis aligned 3D cuboid region in C_j with edge lengths $\mathbf{l} = (l_x, l_y, l_z)$, and $\mathbf{u} \in \mathbb{R}^3$ is an offset vector.

- **Feature Type 1:** measures the intensity difference between $\mathbf{p}$ in a channel C_{j_1} and an offset point $\mathbf{p}+\mathbf{u}$ in a channel C_{j_2}

$$x^{\text{t1}}_{j_1,j_2,\mathbf{u}}(\mathbf{p}, C) = C_{j_1}(\mathbf{p}) - C_{j_2}(\mathbf{p}+\mathbf{u}) \ . \tag{3}$$

- **Feature Type 2:** measures the difference between intensity means of a cuboid around $\mathbf{p}$ in C_{j_1}, and around an offset point $\mathbf{p}+\mathbf{u}$ in C_{j_2}

$$x^{\text{t2}}_{j_1,j_2,\mathbf{l}_1,\mathbf{l}_2,\mathbf{u}}(\mathbf{p}, C) = \mu(R^{\mathbf{l}_1}_{j_1}(\mathbf{p})) - \mu(R^{\mathbf{l}_2}_{j_2}(\mathbf{p}+\mathbf{u})) \ . \tag{4}$$

- **Feature Type 3:** captures the intensity range along a 3D line between $\mathbf{p}$ and $\mathbf{p}+\mathbf{u}$ in one channel. This type is designed with the intuition that structure changes can yield a large intensity change, e.g. NC being dark and AC bright in T1-gad.

$$x^{\text{t3}}_{j,\mathbf{u}}(\mathbf{p}, C) = \max_{\lambda}(C_j(\mathbf{p}+\lambda\mathbf{u})) - \min_{\lambda}(C_j(\mathbf{p}+\lambda\mathbf{u})) \quad \text{with} \ \ \lambda \in [0,1] \ . \tag{5}$$

In the experiments, the types and parameters are drawn uniformly. The offsets u_i originate from the range $[0, 20]$mm, and the cuboid lengths l_i from $[0, 40]$mm.

Dice score	High-grade (real)		Low-grade (real)		High-grade (synth)		Low-grade (synth)	
	Edema	Tumor	Edema	Tumor	Edema	Tumor	Edema	Tumor
mean	0.70	0.71	0.44	0.62	0.65	0.90	0.55	0.71
std. dev.	0.09	0.24	0.18	0.27	0.27	0.05	0.23	0.20
median	0.70	0.78	0.44	0.74	0.76	0.92	0.65	0.78

Table 1: Evaluation summary. The Dice scores are computed by the online evaluation tool provided by the organizers of the BraTS challenge.

3 Evaluation

We evaluate our approach on the real and synthetic data from the BraTS challenge. Both real and synthetic examples contain separate high-grade (HG) and low-grade (LG) data sets. This results in 4 data sets (Real-HG, Real-LG, Synth-HG, Synth-LG). For each of these data sets, we perform the evaluation independently, i.e., we use only the data from one data set for the training and the testing for this data set.

In terms of sizes, Real-HG contains 20 patients, Synth-LG has 10 patients, and the two synthetic data sets contain 25 patients each. For the real data sets, we test our approach on each patient by leave-one-out cross-validation, meaning that for each patient, the training is performed on all other images from the data set, excluding the tested image itself. For the synthetic images, we perform a leave-5-out cross-validation.

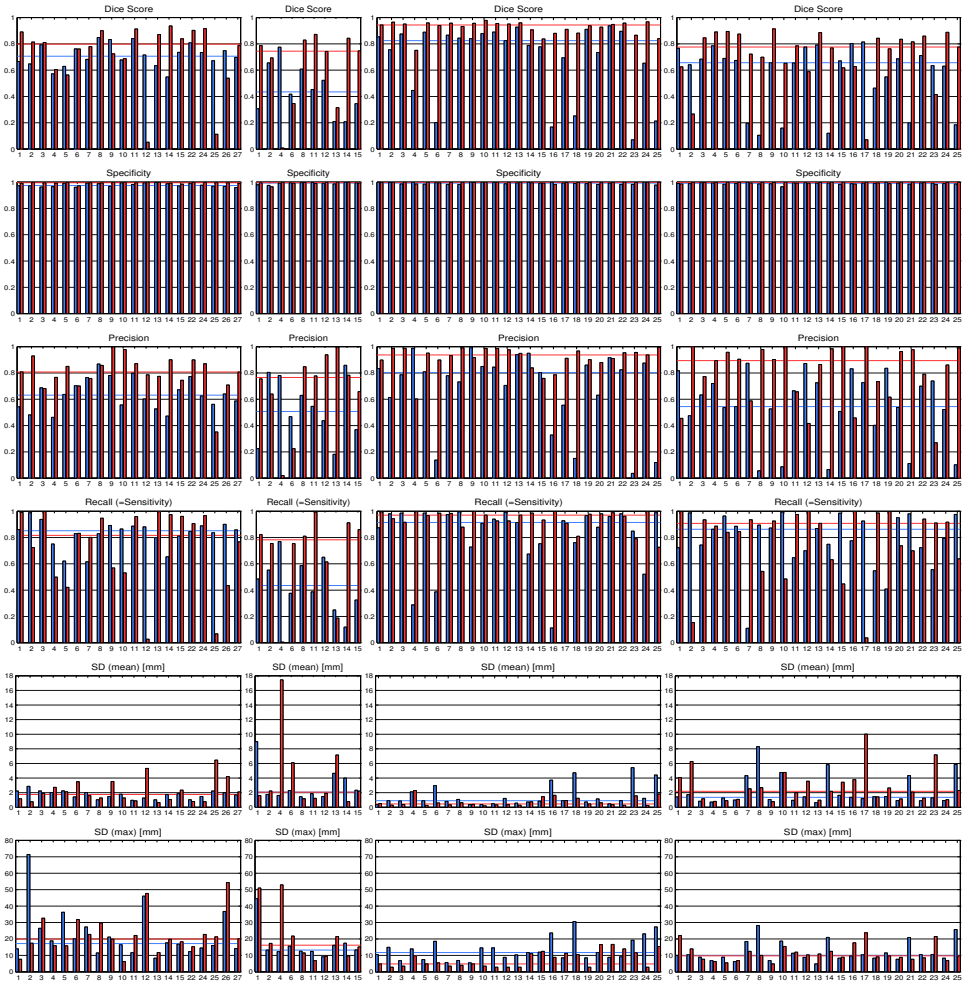
Pre-processing. We apply bias-field normalization by the ITK N3 implementation from [2]. Then, we align the mean intensities of the images within each channel by a global multiplicative factor. For speed reasons, we run the evaluation on a down-sampled version of the input images, with isotropic spatial resolution of 2mm. The computed segmentations are up-sampled back to 1mm for the evaluation.

Settings. In all tests, we employ forests with $T = 40$ trees of depth $D = 20$.

Runtime. Our segmentation method is fully automatic, with segmentation run times in the range of 1-2 minutes per patient. The training of one tree takes approximately 20 minutes on a single desktop PC.

Results. We evaluated our segmentations by the BraTS online evaluation tool, and we summarize the results for the Dice score in Table 1.

Overall, the results indicate a higher segmentation quality for the high-grade tumors than for the low-grade cases, and a better performance on the synthetic data than the real data set.



(a) Real-HG (b)Real-LG (c) Synth-HG (d) Synth-LG

Fig. 2: Per patient evaluation for the four BraTS data sets (Real-HG, Real-LG, Synth-HG, Synth-LG). We show the results for edema (blue) and tumor tissue (red) per patient, and indicate the respective median results with the horizontal lines. We report the following measures: Dice, Specificity, Precision, Recall(=Sensitivity), Mean Surface Distance (SD), and Maximal SD.

Further Evaluation. Furthermore, we reproduce most of the BraTS measures (except Kappa) by our own evaluation in Figure 2. It can be seen in Figure 2, that the *Specificity* is not a very discriminative measure in this application. Therefore, we rather evaluate *Precision*, which is similar in nature, but does

not take the background class into account (TN), and is thus more sensitive to errors.

In order to obtain a better understanding of the data and the performance of our method we perform three further measurements.

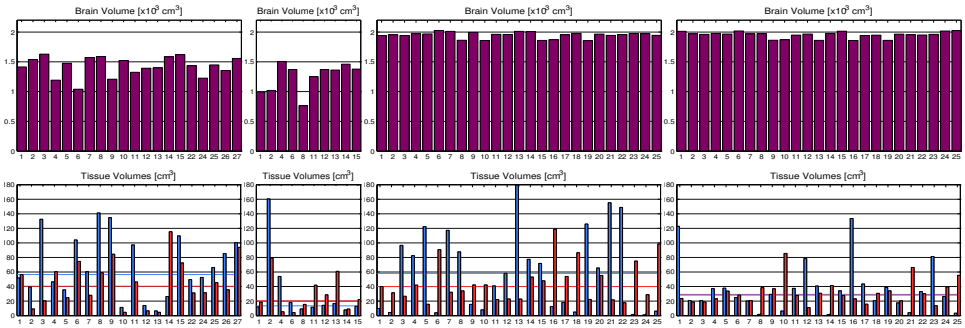
1. In Figure 3, we measure the volumes of the brain, and the edema and tumor tissues for the individual patients. This is done in order to be able to evaluate how target volumes influence the segmentation quality.
2. In Figure 4, we report the results for the basic types of classification outcomes, i.e. true positives (TP), false positives (FP), and false negatives (FN). It is interesting to note the correlation of the TP values with the tissue volumes (cf. Fig. 3). Also, it seems that for edema, the error of our method consists of more FP estimates (wrongly labeled as edema) than FN estimates (wrongly not labeled as edema), i.e. it performs an over-segmentation.
3. In Figure 5, we report additional measures, which might have an application-specific relevance. We compute the overall *Error*, i.e. the volume of all misclassified points FN + FP, and the corresponding relative version, which relates the error to the target volume T, i.e. (FN + FP)/T. Also, we compute the absolute and the relative *Volume Error* $|\mathrm{T} - (\mathrm{TP} + \mathrm{FP})|$, and $|\mathrm{T} - (\mathrm{TP} + \mathrm{FP})|/\mathrm{T}$, which indicate the potential performance for volumetric measurements. The volume error is less sensitive than the error measure, since it does not require an overlap of segmentations but only that the estimated volume is correct (volume error can be expressed as $|\mathrm{FN} - \mathrm{FP}|$).

Acknowledgments

S. J. Price is funded by a Clinician Scientist Award from the National Institute for Health Research (NIHR). O. M. Thomas is a Clinical Lecturer supported by the NIHR Cambridge Biomedical Research Centre.

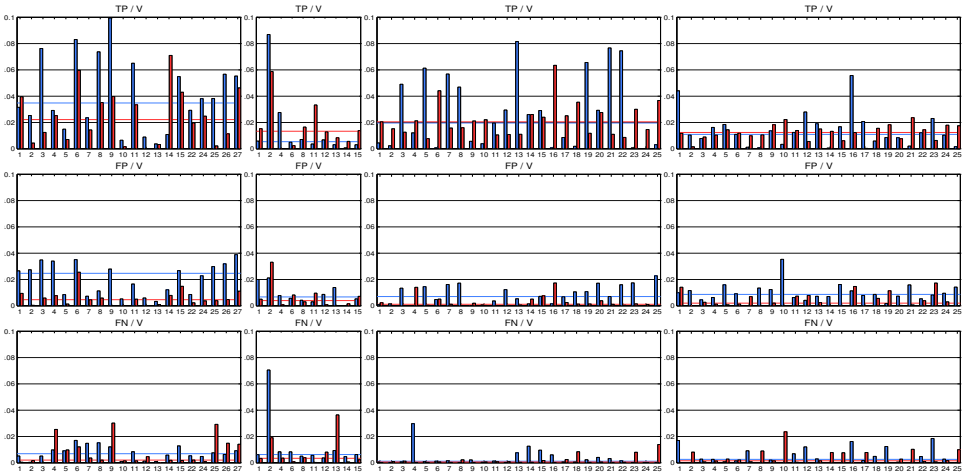
References

1. A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2-3), 2012.
2. N. Tustison and J. Gee. N4ITK: Nick's N3 ITK implementation for MRI bias field correction. *The Insight Journal*, 2010.
3. D. Zikic, B. Glocker, E. Konukoglu, A. Criminisi, C. Demiralp, J. Shotton, O. M. Thomas, T. Das, R. Jena, and Price S. J. Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel mr. In *Proc. Medical Image Computing and Computer Assisted Intervention*, 2012.



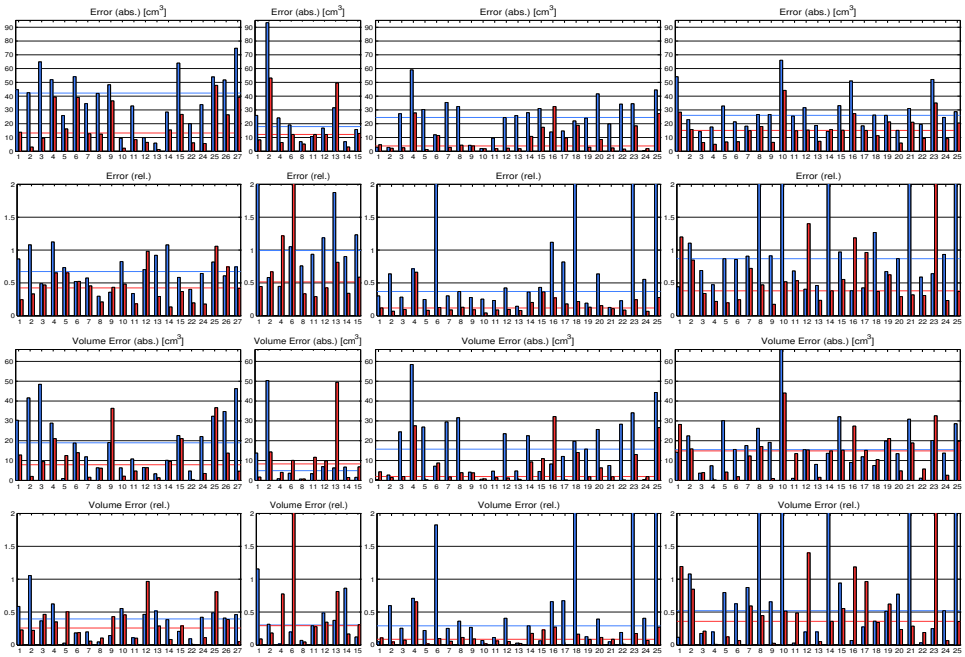
(e) Real-HG (f) Real-LG (g) Synth-HG (h) Synth-LG

Fig. 3: Volume statistics of the BraTS data sets. We compute the brain volumes (top row), and the volumes of the edema (blue) and tumor (red) tissues per patient.



(a) Real-HG (b) Real-LG (c) Synth-HG (d) Synth-LG

Fig. 4: We report the values of true positives (TP), false positives (FP), and false negatives (FN), for edema (blue), and tumor (red) tissues. To make the values comparable, we report them as percentage of the patient brain volume (V). Again, horizontal lines represent median values. It is interesting to note the correlation of the TP values with the tissue volumes (cf. Fig. 3). Also, it seems that for edema, the error of our method consists of more FP estimates (wrongly labeled as edema) than FN estimates (wrongly not labeled as edema), i.e. it performs an over-segmentation.



(a) Real-HG (b)Real-LG (c) Synth-HG (d) Synth-LG

Fig. 5: We further evaluate additional measures which might have application-specific relevance. Again, we have blue=edema, red=tumor, and horizontal line=median. In the two top rows, we compute the *Error*, i.e. the volume of all misclassified points FN + FP, and the relative version, which relates the error to the target volume T, i.e. (FN + FP)/T. In the bottom two rows, we compute the absolute and the relative *Volume Error* |T − (TP + FP)|, and |T − (TP + FP)|/T.