Prediction Error Meta Classification in Semantic Segmentation: Detection via Aggregated Dispersion Measures of Softmax Probabilities

Matthias Rottmann*, Pascal Colling*, Thomas Paul Hack†,
Robin Chan*, Fabian Hüger‡, Peter Schlicht‡ and Hanno Gottschalk*
*University of Wuppertal, School of Mathematics and Natural Sciences
Email: {rottmann, pascal.colling, robin.chan, hanno.gottschalk}@uni-wuppertal.de
†University Leipzig, Faculty of Physics
Email: thomas-paul.hack@itp.uni-leipzig.de
‡Volkswagen Group Innovation, COI Automation, Architecture and AI-Technologies
Email: {fabian.hueger, peter.schlicht}@volkswagen.de

***Abstract*—We present a method that "meta" classifies whether segments predicted by a semantic segmentation neural network intersect with the ground truth. For this purpose, we employ measures of dispersion for predicted pixel-wise class probability distributions, like classification entropy, that yield heat maps of the input scene's size. We aggregate these dispersion measures segment-wise and derive metrics that are well correlated with the segment-wise IoU of prediction and ground truth. This procedure yields an almost plug and play post-processing tool to rate the prediction quality of semantic segmentation networks on segment level. This is especially relevant for monitoring neural networks in online applications like automated driving or medical imaging where reliability is of utmost importance. In our tests, we use publicly available state-of-the-art networks trained on the Cityscapes dataset and the BraTS2017 dataset and analyze the predictive power of different metrics as well as different sets of metrics. To this end, we compute logistic LASSO regression fits for the task of classifying $IoU = 0$ vs. $IoU > 0$ per segment and obtain AUROC values of up to 91.55%. We complement these tests with linear regression fits to predict the segment-wise IoU and obtain prediction standard deviations of down to 0.130 as well as R^2 values of up to 84.15%. We show that these results clearly outperform standard approaches.**

***Index Terms*—computer vision, convolutional neural networks, false positive detection**

I. Introduction

In recent years, deep learning has outperformed other classes of predictive models in many applications. In some of these, e.g. autonomous driving or diagnostics in medicine, the reliability of a prediction is of highest interest. In classification tasks, thresholding on the highest softmax probability or thresholding on the entropy of the classification distributions (softmax output) are commonly used approaches to detect false predictions of neural networks, see e.g. [1], [2]. Metrics like classification entropy or the highest softmax probability are usually combined with model uncertainty (Monte-Carlo (MC) dropout inference) and sometimes input uncertainty, cf. [3] and [2], respectively. These approaches have proven to be practically efficient for detecting uncertainty. Such methods have also been transferred to semantic segmentation tasks. See also [4] for further uncertainty metrics. The work presented in [5] makes use of MC dropout to model the uncertainty of segmentation networks and also shows performance improvements in terms of segmentation accuracy. This approach was applied in other works to model the uncertainty and filter out predictions with low reliability, cf. e.g. [6], [7]. In [8] this line of research was further developed to detect spacial and temporal uncertainty in the semantic segmentation of videos.

In this work we establish an approach for efficiently meta classifying whether an inferred segment (representing a predicted object) of a semantic segmentation intersects with the ground truth or not. This task was first proposed for classification problems in [1] and transferred to semantic segmentation [9], [10], however not on segment level but for estimating the quality of a segmentation for an entire image that contains only a single object of interest. Segment level quality control for brain segmentation by means of metrics computed from MC dropout inferences is introduced in [11] and another MC dropout based approach for object detection is presented in [12].

We term the task of classifying whether a predicted segment intersects with the ground truth or not as *meta classification*. This term has been used in the context of classical machine learning for learning the weights for each member of a committee of classifiers [13]. In terms of deep learning we use this term as a shorthand to distinguish between a network's own classification and the classification whether a prediction is "true" or "false". In contrast to the work cited above, we aim at judging the statistical reliability of each segment inferred by the neural network. To the best of our knowledge, this is the first work that detects false positive segments (objects) in the semantic segmentation with multiple segments per image.

For meta classification we utilize dispersion measures, like entropy, applied to the softmax probabilities (the network's output) on pixel level yielding dispersion heat maps. We aggregate these heat maps over predicted segments alongside with other quantities derived from the network's prediction like

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

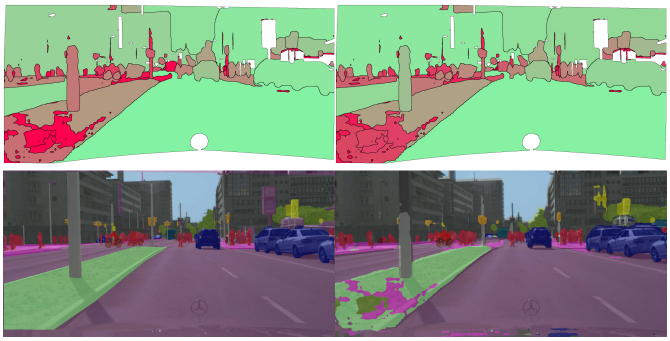


Figure 1. A demonstrations of the proposed method's performance of predicting the segment-wise IoU as a quality measure. The figure consists of ground truth (bottom left), segmentation predicted by DeepLabv3+ MobilenetV2 (bottom right), true IoU for each predicted segment (top left) and prediction of the IoU for each predicted segment obtained by our method (top right). In the top row, green color corresponds to high IoU values and red color to low ones. For the white regions there is no ground truth available. These regions are excluded from statistical evaluations.

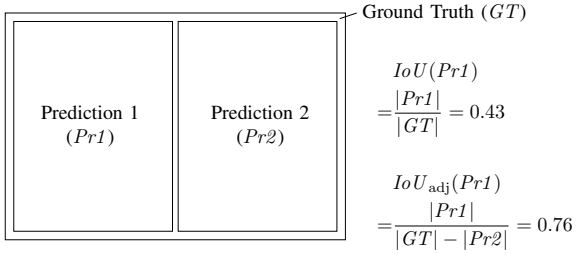


Figure 2. Illustration of different behaviors of IoU and IoU_{adj}. Two disjoint predictions (of the same size and assumed to be assigned to the same class) are enclosed by their corresponding ground truth component. Each predicted component achieves an IoU of 43%. However, this value seems rather low as the ground truth is well covered. Thus, we modify the quality measure for each prediction by excluding that part of the ground truth covered by other predicted components (of the same class), yielding an IoU_{adj} of 76%.

the segment size and predicted class. From this, we construct per-segment metrics. A commonly used performance measure for the quality of a segmentation is the intersection over union (IoU a.k.a. Jaccard index [14]) of prediction and ground truth. We use the constructed metrics as inputs to logistic regression models for *meta classifying*, whether an inferred segment's IoU vanishes or not, i.e., predicting $IoU = 0$ or $IoU > 0$. Also, we use linear regression models for predicting a segment's IoU directly, thus obtaining statements about the reliability of the network's prediction. We term this task *meta regression* and also introduce a modified version of the IoU (adjusted IoU) that is more suitable for this task. The same task is pursued in [9], [10] for images containing only a single object, instead of metrics they utilize additional CNNs. The approach presented in [11] is inherently based on MC dropout while our approach is independent of this.

Our method only uses the softmax output of a semantic segmentation network and the corresponding ground truth. It is a pure post-processing tool that is trained once and offline, there is no additional training of segmentation networks involved. The segmentation network's output is not changed, only assessed. Our approach can be equipped with any heat map obtained from pixel-wise uncertainty measures. Thus, any work on uncertainty quantification for semantic segmentation that yields new improved dispersion heat maps can be seamlessly integrated and leverages our method. Hence, we also provide a framework to evaluate the information contained in pixel-wise uncertainty measures for semantic segmentation. To the best of our knowledge, this is the first work that estimates the quality of each predicted segment in a fully segmented image. A demonstration of its performance is given in Fig. 1.

The work presented in this publication has initiated further research on resolution dependent uncertainty [15] as well as time-dynamic quality estimates [16].

In our tests we use two publicly available datasets: Cityscapes [17] for the semantic segmentation of street scenes and BraTS2017 [18], [19] for brain tumor segmentation. For each of the two datasets we employ two state-of-the-art networks. We perform tests on validation sets and demonstrate that our segment-wise metrics are well correlated with the IoU; thus they are suitable for detecting false positives on segment level. For logistic regression fits we obtain values of up to 91.55% for the area under curve corresponding to the receiver operator characteristic curve (AUROC, see [20]). Predicting the segment-wise IoU via linear regression we obtain prediction standard deviations of down to 0.130 and R^2 values of up to 84.15%.

II. False positives and segment-wise quality measures for semantic segmentation

In order to perform meta classification and regression we first define the corresponding measures that can be deduced from prediction and ground truth.

A segmentation network with a softmax output layer can be seen as a statistical model that provides for each pixel z of the image a probability distribution $f_z(y|x,w)$ on the q class labels $y \in \mathcal{C} = \{y_1, \ldots, y_q\}$, given the weights w and the data x. The predicted class in y is then given by

$$\hat{y}_z(x,w) = \arg\max_{y \in \mathcal{C}} f_z(y|x,w)\,. \tag{1}$$

For a given image x we denote by $\hat{\mathcal{K}}_x$ the set of connected components (segments) in the predicted segmentation $\hat{\mathcal{S}}_x = \{\hat{y}_z(x,w) | z \in x\}$ (omitting the dependence on the weights w). Analogously we denote by $\mathcal{K}_x$ the set of connected components in the ground truth $\mathcal{S}_x$. For each $k \in \hat{\mathcal{K}}_x$, the intersection over union IoU is defined as follows: Let $\mathcal{K}_x|_k$ be the set of all $k' \in \mathcal{K}_x$ that have non-trivial intersection with k and whose class label are equal to the predicted class of k, then

$$IoU(k) = \frac{|k \cap K'|}{|k \cup K'|}\,, \qquad K' = \bigcup_{k' \in \mathcal{K}_x|_k} k'\,. \tag{2}$$

High values of $IoU(k)$ correspond to good predictions, low values to bad predictions. The task of meta classification can

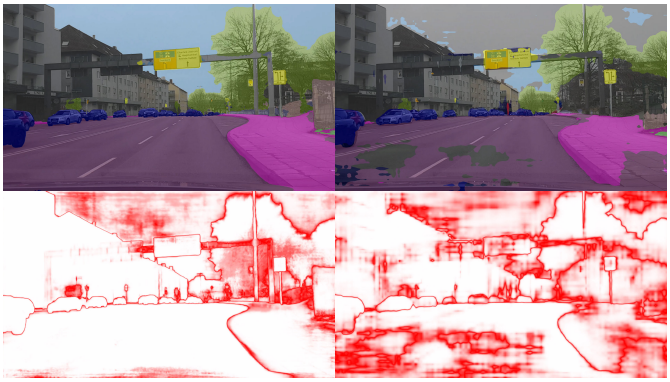


Figure 3. Segmentation example (top line) and heat map D_z (bottom line) for Xception65 (left column) and MobilenetV2 (right column). Original image is not part of the Cityscapes dataset.

now be defined as predicting for each $k \in \hat{\mathcal{K}}_x$, whether $IoU(k) = 0$ or $IoU(k) > 0$. Meta regression amounts to predicting $IoU(k)$ quantitatively. For the latter task, however, in specific scenarios the $IoU(k)$ can have low values while the prediction looks fine. This is the case, when a ground truth segment is covered by more than one predicted segment. In this case the predicted segments can have a low $IoU(k)$ although together they provide a good prediction. To this end we introduce as a segment-wise quality measure the *adjusted intersection over union* IoU_{adj}: Let $Q = \{q \in \hat{\mathcal{K}}_x \setminus \{k\} : q \cap K' \neq \emptyset\}$, then

$$IoU_{\mathrm{adj}}(k) = \frac{|k \cap K'|}{|k \cup (K' \setminus Q)|}\,. \tag{3}$$

The $IoU_{\mathrm{adj}}(k)$ does not punish different predicted segments that share a common bigger ground truth segment, for an illustration of this see Fig. 2. Clearly, we have $IoU_{\mathrm{adj}}(k) = IoU(k) = 1$ if and only if the predicted segment k and the ground truth K' match for each pixel, $IoU_{\mathrm{adj}} = IoU = |k \cap K'| = 0$ when ground truth and predicted segment do not overlap, i.e., $k \cap K' = \emptyset$, and it holds $IoU_{\mathrm{adj}} \geq IoU$. Thus, the meta classification task is invariant under interchanging IoU and IoU_{adj}. However, the meta regression task for directly predicting IoU and IoU_{adj}, respectively, is not invariant. In our experiments we found that the $IoU_{\mathrm{adj}}(k)$ is indeed more suitable for the task of meta regression which is manifested by higher performance in terms of R^2 values. For this discussion we refer to Sec. VI.

III. Pixel-wise dispersion metrics and aggregation over segments

In this section we introduce the metrics that are used as input quantities for performing meta classification and regression. They are based on dispersion measures as well as different size measures that are aggregated for each predicted segment.

Dispersion or concentration measures quantify the degree of randomness in $f_z(y|x,w)$. Here, we consider two of those measures: *entropy* E_z *(also known as Shannon information* [21]) and *difference in probability* D_z, i.e., the difference between the two largest softmax values:

$$E_z(x,w) = -\frac{1}{\log(q)} \sum_{y \in \mathcal{C}} f_z(y|x,w) \log f_z(y|x,w)\,, \tag{4}$$

$$D_z(x,w) = 1 - f_z(\hat{y}_z(x,w)|x,w) + \max_{y \in \mathcal{C} \setminus \{\hat{y}_z(x,w)\}} f_z(y|x,w)\,. \tag{5}$$

For better comparison, both quantities have been written as dispersion measures and been normalized to the interval $[0,1]$: One has $E_z = D_z = 1$ for the equiprobability distribution $f_z(y|x,w) = \frac{1}{q}$, $y \in \mathcal{C}$, and $E_z = D_z = 0$ on the deterministic probability distribution ($f_z(y|x,w) = 1$ for one class and 0 otherwise). For further discussion on dispersion measures, see [22]. The most direct method of uncertainty quantification on an image is the heat mapping of a dispersion measure as in Fig. 3. We now aggregate these measures over predicted segments. Therefore, for each $k \in \hat{\mathcal{K}}_x$, we define the following quantities:

- the interior $k_{in} \subset k$ where a pixel z is an element of k_{in} if all eight neighbouring pixels are an element of k
- the boundary $k_{bd} = k \setminus k_{in}$
- the pixel sizes $S = |k|$, $S_{in} = |k_{in}|$, $S_{bd} = |k_{bd}|$
- the mean entropies $\bar{E}$, $\bar{E}_{in}$, $\bar{E}_{bd}$ defined as

$$\bar{E}_\sharp(k) = \frac{1}{S_\sharp} \sum_{z \in k_\sharp} E_z(x)\,, \qquad \sharp \in \{\_, in, bd\}$$

- the mean distances $\bar{D}$, $\bar{D}_{in}$, $\bar{D}_{bd}$ defined in analogy to the mean entropies
- the relative sizes $\tilde{S} = S/S_{bd}$, $\tilde{S}_{in} = S_{in}/S_{bd}$
- the relative mean entropies $\tilde{E} = \bar{E}\tilde{S}$, $\tilde{E}_{in} = \bar{E}_{in}\tilde{S}_{in}$, and
- the relative mean distances $\tilde{D} = \bar{D}\tilde{S}$, $\tilde{D}_{in} = \bar{D}_{in}\tilde{S}_{in}$.

Typically, E_z and D_z are large for $z \in k_{bd}$. This motivates the separate treatment of interior and boundary measures. With the exception of IoU and IoU_{adj}, all scalar quantities defined above can be computed without the knowledge of the ground truth. Our aim is to analyze to which extent they are able to predict IoU_{adj}.

Table I
Correlation coefficients ρ of aggregated dispersion metrics with respect to IoU_{adj}. Results are computed on the Cityscapes validation set, XC: DeepLabv3+Xception65 and MN: DeepLabv3+MobilenetV2.

	XC	MN		XC	MN
$\bar{E}$	-0.70139	-0.70162	$\bar{D}$	-0.85211	-0.84858
$\bar{E}_{bd}$	-0.44065	-0.41845	$\bar{D}_{bd}$	-0.60308	-0.52163
$\bar{E}_{in}$	-0.71623	-0.69884	$\bar{D}_{in}$	-0.85458	-0.82171
$\tilde{E}$	0.31219	0.36261	$\tilde{D}$	0.22797	0.30245
$\tilde{E}_{in}$	0.39195	0.42806	$\tilde{D}_{in}$	0.29279	0.35131
S	0.30442	0.47978	$\tilde{S}$	0.50758	0.71071
S_{bd}	0.44625	0.62713	$\tilde{S}_{in}$	0.50758	0.71071
S_{in}	0.30201	0.47708			

IV. Numerical Experiments: Street Scenes

We investigate the properties of the metrics defined in the previous section for the example of a semantic segmentation of

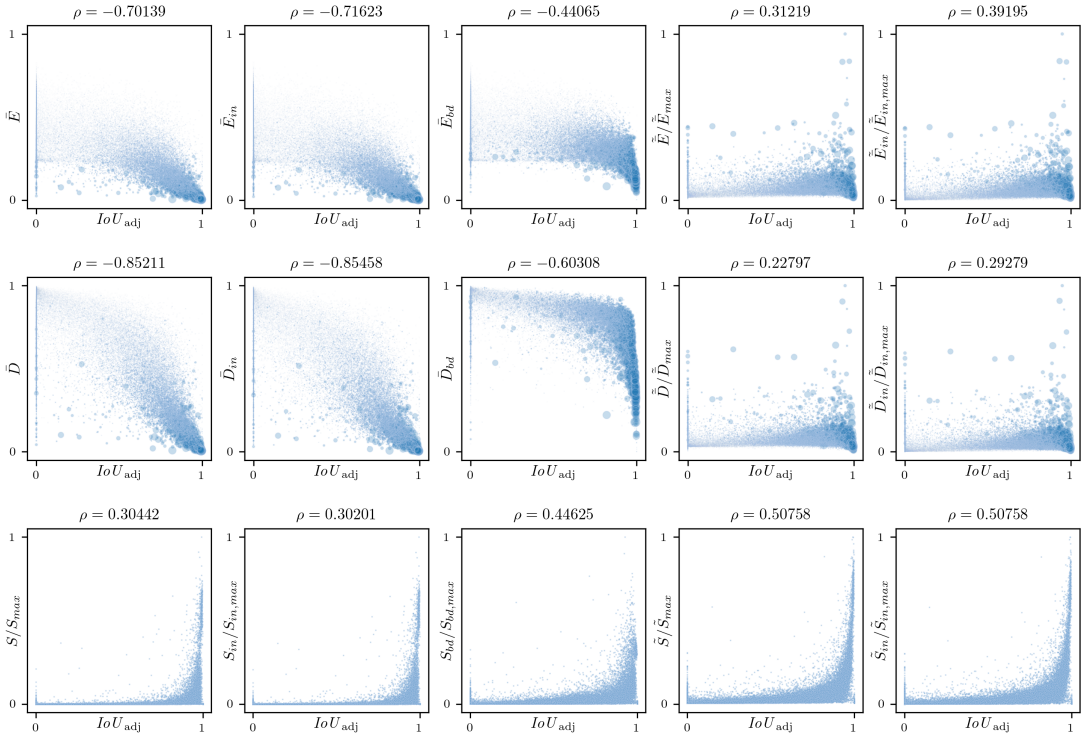


Figure 4. Correlation scatter plots of IoU_{adj} and rescaled features for the DeepLabv3+Xception65 network. Dot sizes in the first two rows are proportional to S.

street scenes. In order to investigate the predictive power of the metrics, we first compute the Pearson correlation $\rho \in [-1, 1]$ between each feature and IoU_{adj}. We report the results of this analysis in Tab. I and Fig. 4.

In our experiments we consider the DeepLabv3+ model [23] for which we use a reference implementation in Tensorflow [24] as well as weights pretrained on the Cityscapes dataset [17] that are available on GitHub. The DeepLabv3+ implementation and weights are available for two network backbones: Xception65, which is a modified version of Xception [25] and is a powerful structure intended for server-side deployment, and MobilenetV2 [26], a fast structure designed for mobile devices. Each of these implementations have parameters tuning the segmentation accuracy. We choose the following best (for Xception65) and worst (for MobilenetV2) parameters in order to perform our analysis on two very distinct networks. Note, that the parameter set for the Xception65 setting also includes the evaluation of the input on multiple scales (averaging the results) which increases the accuracy and also leverages classification uncertainty. We refer to [23] for a detailed explanation of the chosen parameters.

For both networks, we consider the output probabilities and predictions on the Cityscapes validation set, which consists of 500 street scene images at a resolution of 2048×1024. We compute the 15 constructed metrics as well as IoU_{adj} for each segment in the segmentations of the images. Note, that in all computations, we only consider connected components with non-empty interior.

- DeepLabv3+Xception65: output stride 8, decoder output stride 4, evaluation on input scales $0.75, 1.00, 1.25$ – $mIoU = 79.72\%$ on the Cityscapes validation set
- DeepLabv3+MobilenetV2: output stride 16, evaluation on input scale 1.00 – $mIoU = 61.85\%$ on the Cityscapes validation set

For both networks IoU_{adj} shows a strong correlation with the mean distances $\bar{D}$ and $\bar{D}_{in}$ as well as with the mean entropies $\bar{E}$ and $\bar{E}_{in}$. On the other hand, the relative counterparts are less correlated with IoU_{adj}. The relative segment size $\tilde{S}$ for the DeepLabv3+MobilenetV2 network shows a clear correlation whereas this is not the case for the more powerful DeepLabv3+Xception65 network.

In order to find more indicative measures, we now investigate the predictive power of the metrics when they are combined. For the Xception65 net, we obtain 45,194 segments with non-empty interior of which 11,331 have $IoU_{\text{adj}} = 0$. For the weaker MobilenetV2 this ratio is 42,261/17,671. We would first like to detect segments with $IoU_{\text{adj}} = 0$, i.e., learn the meta classification task of identifying false positive segments based on our 15 metrics and the segment-wise averaged probability distribution vectors. We term these (standardized) inputs x_k for a segment k. Further, let $y_k = ceil(IoU_{\text{adj}}) = \{0 \text{ if } IoU_{\text{adj}} = 0, 1 \text{ if } IoU_{\text{adj}} > 0\}$.

The least absolute shrinkage and selection operator (LASSO, [27]) is a popular tool for investigating the predictive power of different combinations of input variables. We compute a series of LASSO fits, i.e., ℓ_1-penalized logistic

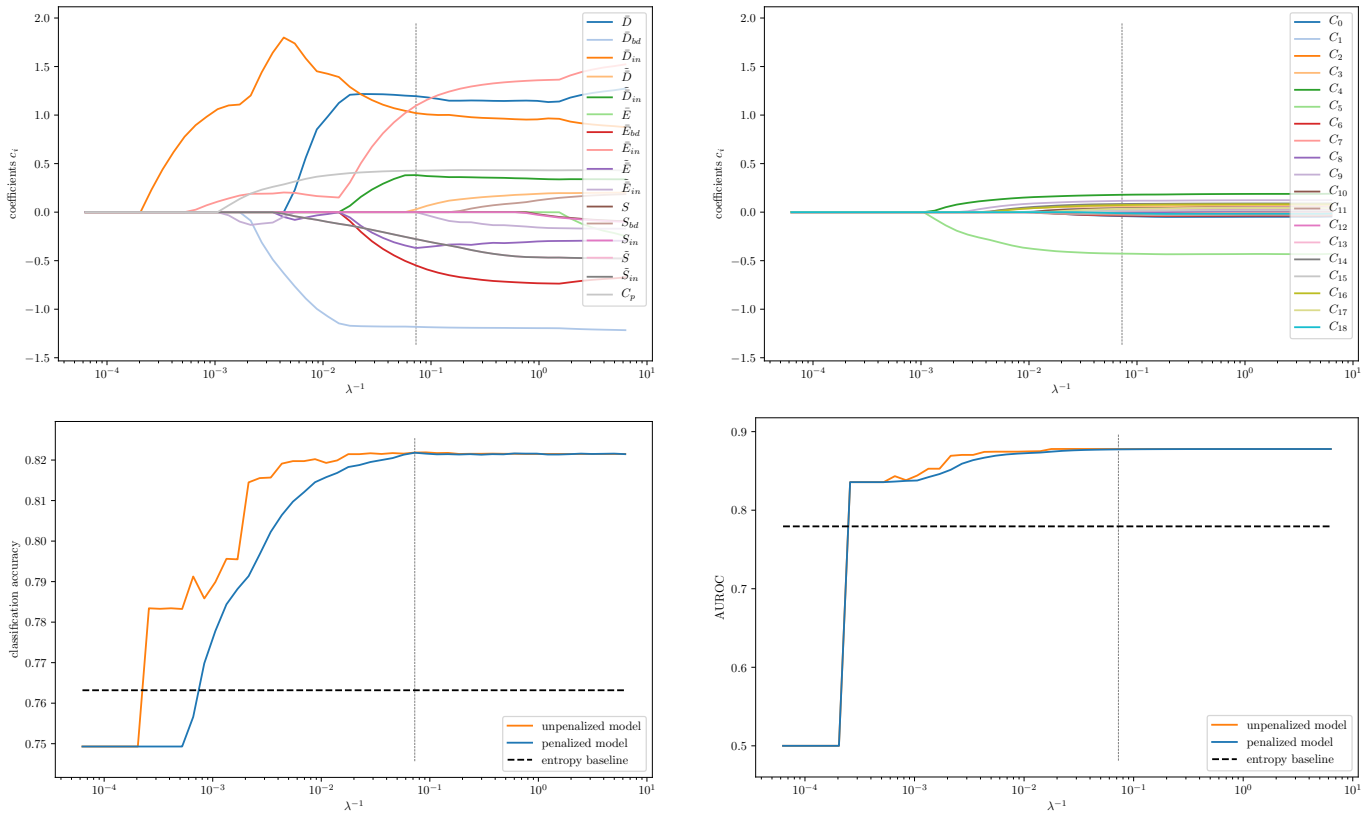


Figure 5. Results for the meta classification task $IoU_{\text{adj}} = 0, > 0$ for predictions obtained from the Xception65 net. (Top left): the weights coefficients for the 15 metrics computed with LASSO fits as function of λ^{-1}, C_p denotes the maximum of the absolute values of all weight coefficients for predicted classes. (Top right): like top left but showing coefficients for the 18 predicted classes. (Bottom left): meta classification rates for $IoU_{\text{adj}} = 0, > 0$. The blue line are the LASSO fits for different λ values, the orange line shows the performance of regular logistic regression fits ($\lambda = 0$) where the input metrics are only those that have non-zero coefficients in the LASSO fit for the current λ. (Bottom right) same as bottom left, but for AUROC. The vertical dashed lines indicate the λ value for which we obtained the best validation accuracy.

regression fits

$$\min_w \sum_i [-y_i \log(\tau(w^T x_i)) \\ - (1 - y_i)(1 - \log(\tau(w^T x_i))) + \lambda \|w\|_1] \ , \tag{6}$$

for different regularization parameters λ and standardized inputs (zero mean and unit standard deviation). Here, $\tau(\cdot)$ is the logistic function. Results for the Xception65 net are shown in Fig. 5.

The top left and top right panels show, in which order the weight coefficients w for each metric/predicted class become active. At the same time the bottom left and bottom right panels show, which weight coefficient causes which amount of increase in predictive performance in terms of meta classification rate and AUROC, respectively. The AUROC is obtained by varying the decision threshold of the logistic regression output for deciding whether $IoU = 0$ or $IoU > 0$.

The first non-zero coefficient activates the $\bar{D}_{in}$ metric, which elevates the predictive power above our reference benchmark of choice, the mean entropy per component $\bar{E}$. Another significant gain is achieved when $\bar{D}_{bd}$ and the predicted classes come into play. In the numerical experiments we randomly choose 10 50/50 training/validation data splits and average the results. Additionally, the bottom line of Fig. 5 shows that there is almost no performance loss when only incorporating some of the metrics proposed by the LASSO trajectory. For both networks the classification accuracy corresponds to a logistic regression trained with unbalanced meta classes $IoU_{\text{adj}} = 0$ and $IoU_{\text{adj}} > 0$, i.e., we did not adjust the class weights. On average (over the 10 training/validation splits) 6851 components with vanishing IoU_{adj} are detected for Xception65 while 4480 remain undetected, for MobilenetV2 this ratio is 14976/2695. These ratios can be adjusted by varying the probability thresholds for deciding between $IoU_{\text{adj}} = 0$ and $IoU_{\text{adj}} > 0$. For this reason we state results in terms of AUROC which is threshold independent.

We compare our results with two different baselines. The naive baseline is given by random guessing (randomly assigning a probability to each segment k and then thresholding on it). The best meta classification accuracy is achieved for the threshold being either 0 or 1. For $I_0 := |\{k : IoU_{\text{adj}} = 0\}|$ and $I_1 := |\{k : IoU_{\text{adj}} > 0\}|$ the naive baseline accuracy is then given by $\frac{\max(I_0, I_1)}{I_0 + I_1}$. The corresponding AUROC value is 50%. Another baseline is to equip our approach only with a single metric. For this purpose we choose the entropy as it is commonly used for uncertainty quantification.

Table II
SUMMARIZED RESULTS FOR THE META CLASSIFICATION AND REGRESSION TASK FOR CITYSCAPES. THE RESULTS ARE AVERAGED OVER 10 RUNS. THE NUMBERS IN BRACKETS DENOTE STANDARD DEVIATIONS OF THE COMPUTED MEAN VALUES.

	Xception65		MobilenetV2	
Cityscapes	training	validation	training	validation
	Classification $IoU_{\text{adj}} = 0, > 0$			
ACC, penalized	81.88%(±0.13%)	81.91%(±0.13%)	78.87%(±0.13%)	78.93%(±0.17%)
ACC, unpenalized	81.91%(±0.12%)	81.92%(±0.12%)	78.84%(±0.14%)	78.93%(±0.18%)
ACC, entropy only	76.36%(±0.17%)	76.32%(±0.17%)	68.33%(±0.27%)	68.57%(±0.25%)
ACC, naive baseline	74.93%		58.19%	
AUROC, penalized	87.71%(±0.14%)	87.71%(±0.15%)	86.74%(±0.18%)	86.77%(±0.17%)
AUROC, unpenalized	87.72%(±0.14%)	87.72%(±0.15%)	86.74%(±0.18%)	86.76%(±0.18%)
AUROC, entropy only	77.81%(±0.16%)	77.94%(±0.15%)	76.63%(±0.24%)	76.74%(±0.24%)
	Regression IoU_{adj}			
σ, all metrics	0.181(±0.001)	0.182(±0.001)	0.130(±0.001)	0.130(±0.001)
σ, entropy only	0.258(±0.001)	0.259(±0.001)	0.215(±0.001)	0.215(±0.001)
R^2, all metrics	75.06%(±0.22%)	74.97%(±0.22%)	81.50%(±0.23%)	81.48%(±0.23%)
R^2, entropy only	49.37%(±0.32%)	49.02%(±0.32%)	49.32%(±0.31%)	49.12%(±0.32%)

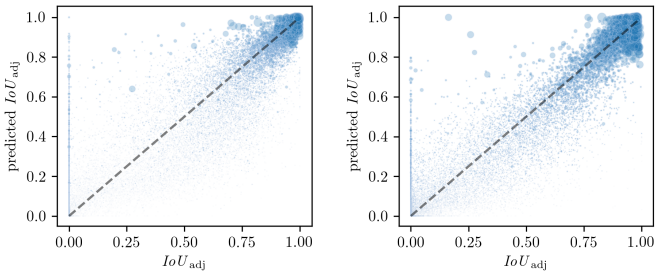


Figure 6. Relationship between IoU_{adj} and predicted IoU_{adj} for all connected components predicted by Xception65 (left) and MobilenetV2 (right). Dot sizes are proportional to connected component's size S.

The meta classification results averaged over 10 runs with different training/validation splits are reported in Tab. II. We obtain a meta classification validation accuracy of up to $81.91\%(\pm 0.13\%)$ and an AUROC of up to $87.71\%(\pm 0.15\%)$ for Xception65. And also for the weaker MobilenetV2 we obtain $78.93\%(\pm 0.17\%)$ classification accuracy and $86.77\%(\pm 0.17\%)$ AUROC. The numbers in brackets denote standard deviations of the performance scores. The classification accuracy and AUROC results are slightly biased towards the validation results as they correspond to the particular λ value that maximizes the validation accuracy. Both baselines (random guessing and entropy) are clearly outperformed and indicate that the computed set of dispersion measures contains rich information for detecting unreliably predicted segments.

Ultimately, we want to perform meta regression, i.e., predict IoU_{adj} values for all connected components and thus model a quality measure. We now resign from regularization and use a linear regression model to predict the IoU_{adj}. Fig. 6 depicts the quality of a single linear regression fit for each of the two segmentation networks.

For Xception65 we obtain an R^2 value of $74.93\%(\pm 0.22\%)$ and for MobilenetV2 $81.48\%(\pm 0.23\%)$. Averaged results over 10 runs including standard deviations σ are summarized in Tab. II. In both cases, our presented approach clearly outperforms the entropy. The linear regression models do not overfit the data and note-worthily we obtain prediction standard deviations of down to 0.130 and almost no standard deviation for the averages.

V. NUMERICAL EXPERIMENTS: BRAIN TUMOR SEGMENTATION

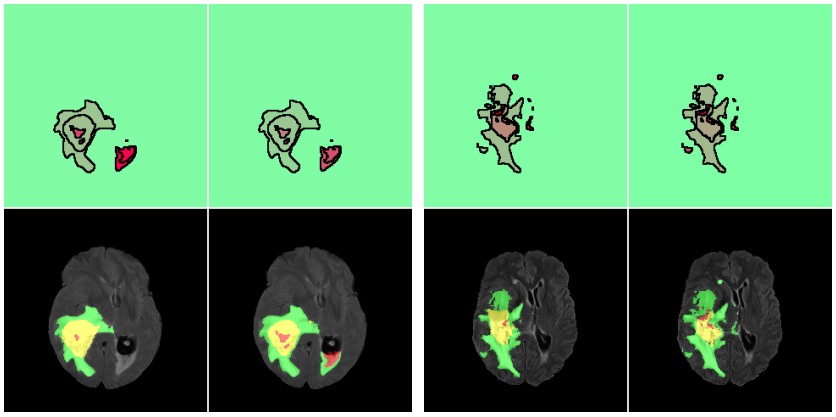
The method we propose only uses dispersion heat maps and softmax probabilities as inputs. Any additional heat map increases the performance as long as there is no overfitting. Thus, we expect our approach to generalize across different datasets even from different domains. To demonstrate this, we perform additional tests with the brain tumor segmentation dataset BraTS2017 [18], [19] and two different networks, i.e., a simple 2D network and a more complex 3D network. Compared to the segmentation of street scenes, brain tumor segmentation involves way fewer classes. The background class is usually dominant. In BraTS2017, roughly 98% of all pixels are background, the remaining classes comprise necrotic/non-enhancing tumor, peritumoral edema and enhancing tumor. For benchmarks of predictive methods, these labels are combined into three nested classes: whole tumor (WT), tumor core (TC) and enhancing tumor (ET) (see Fig. 7). The most commonly used evaluation metric is the so-called *Dice-Coefficient* [28] that is defined as

$$Dice := 2TP/(2TP + FP + FN) \tag{7}$$

where TP, FP and FN denote all true positive, false positive and false negative pixels, respectively, for a chosen class.

The BraTS data is available as magnetic resonance imaging (MRI) brain scans from three viewing angles and with four modalities of higher grade gliomas (HGG) and lower grade gliomas (LGG). For training and validation, we combine HGG and LGG images and randomly split the data 80/20. We train the networks from scratch with the different scan modalities stacked as the network's input channels. Once this is done, we perform tests analogously to the previous section. The performance of the two networks being used for our validation split are reported in Tab. III, the results for meta classification and regression are summarized in Tab. IV.

For the first test we use the network by Kermi et al. [29]. It is based on the U-Net [30] which is originally well known for



(a) 2D U-Net by Kermi et al. (b) 3D Net by Myronenko et al.

Figure 7. Two demonstrations (left and right four panels, analogously to Fig. 1) of our method's performance of predicting IoU_{adj} on BraTS2017. In the bottom row, the whole tumor (WT) includes all colored segments (union of green, yellow & red), the tumor core (TC) the yellow joined with the red colored segments and the enhancing tumor (ET) only the yellow colored segments. In the top row, green color corresponds to high IoU_{adj} values and red color to low ones. In both examples, predicted quality and true quality look very similar.

Table III

BRATS2017 PERFORMANCE SCORES ON VALIDATION DATA SPLIT FOR THE TWO NETWORKS USED IN THE NUMERICAL EXPERIMENTS. THE NESTED CLASSES WHOLE TUMOR (WT), TUMOR CORE (TC) AND ENHANCING TUMOR (ET) ARE EVALUATED WITH THE DICE COEFFICIENT. FOR COMPARISON PURPOSES, THE MEAN DICE SCORE IS REPORTED AS WELL AS MEAN INTERSECTION OVER UNION FOR NESTED CLASSES AND SINGLE CLASSES (BACKGROUND, NON-ENHANCING TUMOR, PERITUMORAL EDEMA AND ENHANCING TUMOR).

Metric	Dice Coefficient				Intersection over Union	
	WT	TC	ET	$mDice$	$mIoU$ nested	$mIoU$ single
2D U-Net by Kermi et al.	88.09%	77.38%	78.89%	81.45%	68.99%	67.14%
3D Net by Myronenko et al.	88.83%	81.07%	79.63%	83.18%	71.40%	69.64%

Table IV

SUMMARIZED RESULTS FOR THE META CLASSIFICATION AND REGRESSION TASK FOR BRATS2017. THE RESULTS ARE AVERAGED OVER 10 RUNS. THE NUMBERS IN BRACKETS DENOTE STANDARD DEVIATIONS OF THE COMPUTED MEAN VALUES.

	2D U-Net by Kermi et al.		3D Net by Myronenko et al.	
BraTS2017	training	validation	training	validation
	Classification $IoU_{\text{adj}} = 0, > 0$			
ACC, penalized	89.30%(±0.18%)	89.39%(±0.17%)	88.40%(±0.27%)	88.42%(±0.27%)
ACC, unpenalized	89.29%(±0.19%)	89.40%(±0.18%)	88.38%(±0.27%)	88.40%(±0.28%)
ACC, entropy only	87.96%(±0.12%)	87.96%(±0.12%)	86.69%(±0.20%)	86.69%(±0.20%)
ACC, naive baseline	88.30%		86.35%	
AUROC, penalized	91.84%(±0.25%)	91.93%(±0.24%)	91.51%(±0.22%)	91.55%(±0.22%)
AUROC, unpenalized	91.83%(±0.25%)	91.93%(±0.24%)	91.49%(±0.22%)	91.53%(±0.22%)
AUROC, entropy only	86.68%(±0.25%)	86.73%(±0.25%)	86.59%(±0.28%)	86.74%(±0.28%)
	Regression IoU_{adj}			
σ, all metrics	0.148(±0.001)	0.149(±0.001)	0.171(±0.001)	0.171(±0.001)
σ, entropy only	0.178(±0.001)	0.178(±0.001)	0.198(±0.001)	0.197(±0.001)
R^2, all metrics	84.22%(±0.21%)	84.15%(±0.21%)	79.53%(±0.28%)	79.64%(±0.28%)
R^2, entropy only	77.18%(±0.18%)	77.30%(±0.17%)	72.63%(±0.27%)	72.91%(±0.27%)

its performance on biomedical image segmentation. We train the network on randomly sampled 2D patches from axial (top view) slices of the brain scans. The results of our prediction rating methods are computed for 22,242 non-empty segments of which 2,603 have $IoU_{\text{adj}} = 0$. Indeed, we obtain higher accuracy values compared to Cityscapes, however the gain over the single metric baseline is not as big. This is primarily due to a strong correlation between E and IoU_{adj} (-0.87794). In this case, the gain over the naive baseline is marginal. This may be misleading to the disadvantage of our method as the high naive accuracy is caused by the strong sample imbalance of the meta classes. The corresponding AUROC value of 91.93% shows that our method meta classifies with significantly higher confidence when incorporating all metrics. Regarding the R^2 value of our regression model for predicting IoU_{adj}, we observe an increase from $77.30\%(\pm 0.17\%)$ to $84.15\%(\pm 0.21\%)$ when incorporating all metrics instead of only the entropy.

Next, we compare the U-Net's performance to the state-of-the-art network by Myronenko et al. [31]. One main difference is that the latter network considers the MRI scans' 3D contextual information by processing multiple contiguous

Table V
COMPARISON OF REGRESSION RESULTS FOR SEGMENT-WISE FITTING IoU_{adj} AND IoU, AVERAGED OVER 10 RUNS. THE NUMBERS IN BRACKETS DENOTE STANDARD DEVIATIONS OF THE COMPUTED MEAN VALUES.

	Xception65		MobilenetV2	
	training	validation	training	validation
	Regression IoU_{adj}			
σ, all metrics	$0.181(\pm 0.001)$	$0.182(\pm 0.001)$	$0.130(\pm 0.001)$	$0.130(\pm 0.001)$
σ, entropy only	$0.258(\pm 0.001)$	$0.259(\pm 0.001)$	$0.215(\pm 0.001)$	$0.215(\pm 0.001)$
R^2, all metrics	$75.06\%(\pm 0.22\%)$	$74.97\%(\pm 0.22\%)$	$81.50\%(\pm 0.23\%)$	$81.48\%(\pm 0.23\%)$
R^2, entropy only	$49.37\%(\pm 0.32\%)$	$49.02\%(\pm 0.32\%)$	$49.32\%(\pm 0.31\%)$	$49.12\%(\pm 0.32\%)$
	Regression IoU			
σ, all metrics	$0.192(\pm 0.001)$	$0.192(\pm 0.001)$	$0.135(\pm 0.001)$	$0.135(\pm 0.001)$
σ, entropy only	$0.267(\pm 0.001)$	$0.268(\pm 0.001)$	$0.217(\pm 0.001)$	$0.217(\pm 0.001)$
R^2, all metrics	$72.90\%(\pm 0.21\%)$	$72.77\%(\pm 0.21\%)$	$79.63\%(\pm 0.27\%)$	$79.58\%(\pm 0.27\%)$
R^2, entropy only	$47.43\%(\pm 0.28\%)$	$47.07\%(\pm 0.28\%)$	$47.73\%(\pm 0.37\%)$	$47.50\%(\pm 0.38\%)$

(a) (b) (c)

Figure 8. Illustration of the different behaviors of IoU and IoU_{adj}. We have (a): IoU per predicted segment, (b) left: ground truth, right: detail views for the crucial area of the predicted segmentation (top) and the corresponding ground truth (bottom) and (c): IoU_{adj} per segment. Green stands for high IoU and IoU_{adj} values, red for low ones, respectively. The top right panel in (b) shows that the prediction for the class 'nature' is decoupled into two components by the traffic light's prediction. The IoU rates this small part on the left from the traffic light very badly even though the prediction is absolutely fine. The adjusted IoU_{adj} circumvents this type of problems.

2D slices at once, i.e., we train the network on randomly sampled 3D patches. As a consequence, the model is more complex and the number of trainable parameters is noticeably increased (10.1M vs. 17.3M). We perform the evaluation in the same 2D slice-wise manner as for the U-Net. The results are now computed for 24,397 non-empty segments of which 3331 have $IoU_{\text{adj}} = 0$. Again, we observe a strong correlation between E and IoU_{adj} of -0.84294 which results in a nearly identical gain in terms of percent points over the single metric baseline as for the U-Net. Also with respect to the R^2 value of our regression model, the gain is again around 7 percent points, whereas the absolute value with $79.64\%(\pm 0.28\%)$ for all metrics is not as high as for the U-Net.

VI. COMPARISON OF IoU AND IoU_{adj}

Recall from Sec. II, the $IoU_{\text{adj}}(k)$ does not punish differently predicted segments that share a common bigger ground truth segment, whereas the standard IoU measure does. As the meta regression task is not invariant under interchanging IoU and IoU_{adj}, we compare performance differences when using either of these measures. Carrying out the regression tests from Sec. IV for the IoU_{adj} with the IoU as well, we observe that the regression fit for the IoU_{adj} achieves R^2 values that are roughly 2% higher than those for the IoU, cf. Tab. V. Usually, for performance measures in semantic segmentation, the IoU is computed for a chosen class over the whole image. This means that each pixel of the union of prediction and ground truth is only counted once in the denominator of the image-wise IoU. On the other hand, a ground truth pixel may contribute to segment-wise IoUs of several segments, a practical example is given in Fig. 8. In this sense, in the context of semantic segmentation, the adjusted IoU_{adj} is in spirit closer to the regular image-wise IoU.

VII. CONCLUSION AND OUTLOOK

We have shown statistically that per-segment metrics derived from entropy, probability difference, segment size and the predicted class clearly contain information about the reliability of the segments and constructed an approach for detecting unreliable segments in the network's prediction. In our tests with publicly available networks and datasets, Cityscapes and BraTS2017, the computed logistic LASSO fits for meta classification task $IoU_{\text{adj}} = 0$ vs. $IoU_{\text{adj}} > 0$ achieve AUROC values of up to 91.55%. When predicting the IoU_{adj} with a linear regression fit we obtain a prediction standard deviation of down to 0.130, as well as R^2 values of up to 84.15%. These results could be further improved when incorporating model uncertainty in heat map generation. We believe that using MC dropout will further improve these results, just like the development of ever more accurate networks. We plan to use our method for detecting labeling errors, for label acquisition in active learning and we plan to investigate further metrics that may leverage detection accuracy. Apart from that, detection mechanisms built on the

softmax input and even earlier layers could be thought of. Furthermore, when reducing the number of false negatives for a chosen class by adjusting softmax thresholds, our method can be used to keep the production of new false positives under control. The source code of our method is publicly available at https://github.com/mrottmann/MetaSeg.

Acknowledgment

This work is in part funded by Volkswagen Group Innovation.

References

[1] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *CoRR*, vol. abs/1610.02136, 2016. [Online]. Available: http://arxiv.org/abs/1610.02136

[2] S. Liang, Y. Li, and R. Srikant, "Principled detection of out-of-distribution examples in neural networks," *CoRR*, vol. abs/1706.02690, 2017. [Online]. Available: http://arxiv.org/abs/1706.02690

[3] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, pp. 1050–1059. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045390.3045502

[4] P. Oberdiek, M. Rottmann, and H. Gottschalk, "Classification uncertainty of deep neural networks based on gradient information," in *Artificial Neural networks and Pattern Recognition (ANNPR)*, 2018.

[5] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *CoRR*, vol. abs/1511.02680, 2015. [Online]. Available: http://arxiv.org/abs/1511.02680

[6] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 680–688, 2016.

[7] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps," *CoRR*, vol. abs/1807.10584, 2018. [Online]. Available: http://arxiv.org/abs/1807.10584

[8] P.-Y. Huang, W.-T. Hsu, C.-Y. Chiu, T.-F. Wu, and M. Sun, "Efficient uncertainty estimation for semantic segmentation in videos," in *European Conference on Computer Vision (ECCV)*, 2018.

[9] T. DeVries and G. W. Taylor, "Leveraging uncertainty estimates for predicting segmentation quality," *CoRR*, vol. abs/1807.00502, 2018. [Online]. Available: http://arxiv.org/abs/1807.00502

[10] C. Huang, Q. Wu, and F. Meng, "Qualitynet: Segmentation quality evaluation with deep convolutional networks," in *2016 Visual Communications and Image Processing (VCIP)*, Nov 2016, pp. 1–4.

[11] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger, "Inherent brain segmentation quality control from fully convnet monte carlo sampling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 664–672.

[12] O. Ozdemir, B. Woodward, and A. A. Berlin, "Propagating uncertainty in multi-stage bayesian convolutional neural networks with application to pulmonary nodule detection," *CoRR*, vol. abs/1712.00497, 2017. [Online]. Available: http://arxiv.org/abs/1712.00497

[13] W. Lin and A. Hauptmann, "Meta-classification: Combining multimodal classifiers," *Mining Multimedia and Complex Data. PAKDD 2002. Lecture Notes in Computer Science*, vol. 2797, 2003.

[14] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, Feb. 1912. [Online]. Available: http://www.jstor.org/stable/2427226?seq=3

[15] M. Rottmann and M. Schubert, "Uncertainty measures and prediction quality rating for the semantic segmentation of nested multi resolution street scene images," *CoRR*, vol. abs/1904.04516, 2019. [Online]. Available: http://arxiv.org/abs/1904.04516

[16] K. Maag, M. Rottmann, and H. Gottschalk, "Time-dynamic estimates of the reliability of deep semantic segmentation networks," *CoRR*, vol. abs/1911.05075, 2019. [Online]. Available: http://arxiv.org/abs/1911.05075

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, Sep 2017, data Descriptor. [Online]. Available: https://doi.org/10.1038/sdata.2017.117

[19] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct 2015.

[20] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, 2006, pp. 233–240. [Online]. Available: http://doi.acm.org/10.1145/1143844.1143874

[21] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948. [Online]. Available: http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

[22] F. Cowell, *Measuring Inequality*, 3rd ed. Oxford University Press, 2011.

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *CoRR*, vol. abs/1802.02611, 2018.

[24] M. Abadi, A. Agarwal, P. Barham *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[25] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, 2017.

[26] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," *CoRR*, vol. abs/1801.04381, 2018.

[27] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, pp. 267–288, 1996. [Online]. Available: https://www.bibsonomy.org/bibtex/290e648276aa6cd3c601e7c0a54366233/dieudonnew

[28] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. C. Tempany, M. R. Kaus, S. J. Haker, W. M. r. Wells, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index," *Academic radiology*, vol. 11, no. 2, pp. 178–189, Feb 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/14974593

[29] A. Kermi, I. Mahmoudi, and M. T. Khadir, "Deep convolutional neural networks using u-net for automatic brain tumor segmentation in multimodal mri volumes," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 37–48.

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

[31] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham: Springer International Publishing, 2019, pp. 311–320.