Detection of False Positive and False Negative Samples in Semantic Segmentation

Matthias Rottmann
University of Wuppertal & ICMD
rottmann@math.uni-wuppertal.de

Kira Maag
University of Wuppertal & ICMD
maag@math.uni-wuppertal.de

Robin Chan
University of Wuppertal & ICMD
chan@math.uni-wuppertal.de

Fabian Hüger
Volkswagen Group Innovation
fabian.hueger@volkswagen.de

Peter Schlicht
Volkswagen Group Innovation
peter.schlicht@volkswagen.de

Hanno Gottschalk
University of Wuppertal & ICMD
hanno.gottschalk@uni-wuppertal.de

Abstract—In recent years, deep learning methods have outperformed other methods in image recognition. This has fostered imagination of potential application of deep learning technology including safety relevant applications like the interpretation of medical images or autonomous driving. The passage from assistance of a human decision maker to ever more automated systems however increases the need to properly handle the failure modes of deep learning modules. In this contribution, we review a set of techniques for the self-monitoring of machine-learning algorithms based on uncertainty quantification. In particular, we apply this to the task of semantic segmentation, where the machine learning algorithm decomposes an image according to semantic categories. We discuss false positive and false negative error modes at instance-level and review techniques for the detection of such errors that have been recently proposed by the authors. We also give an outlook on future research directions.

Index Terms—deep learning, semantic segmentation, false positive and false negative detection

I. INTRODUCTION

The stunning success of deep learning technology, convolutional neural networks (CNN) in particular [1]–[3], has led to a rush towards technology development for new applications that ten years ago would have been considered unrealistic. In particular, fully automated driving systems are intensively developed in the automotive industry including also new competitors [4], [5]. While the industry strives to advance such systems from driving assistance for a human driver (level 1 and 2 of automated driving) to higher levels where the human as the ultimate redundancy for the technology can be temporally (level 3 and 4) or entirely replaced (level 5), the question of how to design automated driving systems based on deep learning technology still poses a number of unresolved questions, in particular with respect to reliability and safety [5], [6]. A similar set of problems exists when AI-driven systems assist the interpretation of medical images [7], although there is no intention to fully automate this process.

In the following, we focus on the semantic interpretation of street scenes based on camera data which is an important prerequisite for any automated driving strategy. For the sake of

concreteness, we focus on semantic segmentation in contrast to object detection [8]. In semantic segmentation, an image is decomposed into a number of masks, each of which unifies the pixels that adhere to a specific category in a predefined semantic space [9]. Despite there also exist instance segmentation networks [10], we here consider each connected component of a mask as one instance. Based on such instances, the following failure modes have to be taken into account:

- False Positive (FP): An instance of a given category that is present in the predicted mask has zero intersection with the same category in the ground truth mask.
- False Negative (FN): An instance of a given category that is present in the ground truth mask is completely overlooked, i.e., has zero intersection with the same category in the predicted mask.
- Out of Distribution (OOD): An object that is outside the semantic space on which the perception algorithm has been trained nevertheless occurs in the input data and therefore is misclassified [11], [12].
- Adversarial Attack (AA): The perception module is intentionally forced to commit an FP or FN error by manipulation of the input of the sensor [3], [13].

In the following, we focus on the first two 'FP' and 'FN' failure modes. In particular, we discuss methods for self-monitoring of segmentation networks. Improved reliability due to redundancies in the architecture of autonomous cars is not considered here.

While the detection of false positives in semantic segmentation is mostly considered on a pixel level and is measured with global indices like the global accuracy over frames or the averaged intersection over union (IoU) on class mask level [14], [15], here we pass on to connected components in the predicted masks of segmentation networks, which is often more relevant in practice. Meta classification then is the machine learning task to infer from the aggregated uncertainty metrics whether the predicted segment has intersection with the ground truth, or is a false positive in the sense given above. While this results in a $0 - 1$ decision, the IoU score on *a single connected component* gives a gradual quality measure.

Meta regression then is the task to predict this score from the same uncertainty metrics in the absence of ground truth. In this article we give an overview over recent progress in meta classification [16]–[18] for the semantic segmentation of street scenes [19].

We also deal with class imbalance as one of the reasons for false negative predictions for which the corresponding ground truth is underrepresented in the training data. In semantic segmentation this is often unavoidable, as e.g. pedestrians are underrepresented in terms of their pixel count even on images with several individuals. Here we propose methods to correct the bias from the maximum *a posteriori* (or Bayes) decision principle that is mostly applied in machine learning. As alternatives we propose a decision principle – the maximum likelihood (ML) decision rule [20] – that looks out for the best fit of the data to a given semantic class. We review the false negative detection [21] using the ML decision rule for semantic segmentation and also discuss cost based decision rules in general along with the problems of setting the cost structure up [22].

The paper is organised as follows: In section II we discuss the detection of false positive instances by a meta classification procedure that involves uncertainty heatmaps aggregated over predicted segments. The following section III extends this procedure to video stream data using time series of uncertainty metrics for meta-classification. In section IV we discuss the reduction of false negatives especially for rare classes of high importance. Here we use cost based decision rules and discuss some of the ethical issues connected with setting up the cost structure. Finally, section V gives a summary and outlook to future research.

II. FALSE POSITIVE DETECTION VIA META CLASSIFICATION

In semantic segmentation, a network with a softmax output layer 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 input image $x \in \mathcal{X}$. Using the maximum *a posteriory* probability (MAP) principle, also called Bayes decision rule, predicted class in z is then given by

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

We denote by $\hat{\mathcal{K}}_x$ the set of connected components (segments) in the predicted segmentation $\hat{\mathcal{S}}_x = \{\hat{y}_z(x,w) | z \in x\}$. Analogously we denote by $\mathcal{K}_x$ the set of connected components in the ground truth $\mathcal{S}_x$.

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 equals the predicted class for k, then the intersection over union (IoU) is defined as

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

In the given context, false positive detection (cf. [23] for classification tasks with neural networks) corresponds to the binary classification task $IoU(k) = 0$ or $IoU(k) > 0$ for a given segment $k \in \hat{\mathcal{K}}_x$, i.e., k intersect with ground truth or not. We term this task *meta classification*, see [16]. In analogy we term the regression task of estimating $IoU(k)$ directly as *meta regression*, this can also be viewed as a quality measure. Quality estimates for neural networks were first proposed for one object per image in [24], [25]. While these works rely on neural networks as post processors, we introduced a light-weight and transparent approach in [16] that deals with multiple segments per image for both meta tasks.

In our approach presented in [16] we proceed as follows: for each $k \in \hat{\mathcal{K}}_x$ we construct metrics based on dispersion measures of $f_z(y|x,w)$ (entropy, probability margin) as well as fractality measures of k. The dispersion measures are aggregated over the predicted segment by computing their averages, fractality is measured by the quotient of volume and boundary length of k. We observe that these metrics are strongly correlated with the $IoU(k)$, yielding Pearson correlation coefficients R of up to 0.85 (in absolute values) for two different state-of-the-art DeepLabv3+ [26] networks (Xception65 [14] and MobilenetV2 [15]). Hence, the constructed metrics are suitable for both meta tasks. The construction of metrics can be seen as a map $\mu : \hat{\mathcal{K}}_x \to \mathbb{R}^m$ that maps k to a vector of metrics. Thus,

$$M = \{\mu(k) : x \in \mathcal{X}, \, k \in \hat{\mathcal{K}}_x\}, \, M_i = \{\mu_i(k)\} \tag{3}$$

is a structured dataset, for further details on the construction of M we refer to [16]. We perform meta tasks by training linear models, i.e., a linear regression model for meta regression and a logistic one for meta classification, both of them based on M. We split the set of all predicted segments and their corresponding metrics obtained from the Cityscapes [27] validation set into meta training and meta test sets (80%/20%) and compare our approach with the following baselines: for the entropy baseline we employ for both meta tasks a single metric, i.e., the mean entropy over a predicted segment k as the entropy is a commonly used uncertainty measure. Furthermore, for the classification task a naive random guessing baseline can be formulated by randomly assigning a probability to each segment k and then thresholding on it. A comparison of our meta classification approach is given in table I. Noteworthily, we obtain AUROC values of up to 87.72% (roughly 10 percent points (pp.) above the entropy baseline) for meta classification and R^2 values of up 81.48% (more than 30 pp. above the entropy baseline) for meta regression. A visualization demonstrating the performance of our approach is given in fig. 1. In [16], we also present results for the BraTS2017 brain tumor segmentation dataset [28].

From now on, we assign the term *MetaSeg* to the introduced method. In [18] we extended this approach by taking resolution dependent uncertainty into account. As neural networks with their fixed filter sizes are not scale invariant, it makes a difference whether we infer the original input image or a resized one with the same network. Consequently, we introduced a pyramid-type of approach where a sequence of nested image crops with common center point are resized to
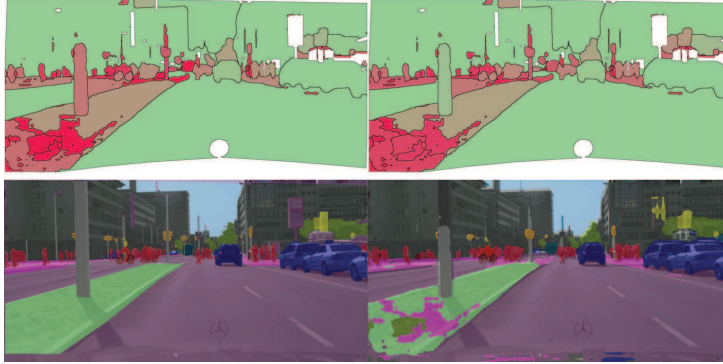
Fig. 1. Prediction of the IoU with linear regression. The figure consists of ground truth (bottom left), predicted segments (bottom right), true IoU for the predicted segments (top left) and predicted IoU for the predicted segments (top right). In the top row, green color corresponds to high IoU values and red color to low ones, for the white regions there is no ground truth available. These regions are excluded from the statistical evaluation.

TABLE I
SUMMARIZED RESULTS FOR CLASSIFICATION AND REGRESSION FOR
CITYSCAPES, AVERAGED OVER 10 RUNS. THE NUMBERS IN BRACKETS
DENOTE STANDARD DEVIATIONS OF THE COMPUTED MEAN VALUES.

Cityscapes	Xception65		MobilenetV2	
	training	validation	training	validation
Meta Classification $IoU = 0, > 0$				
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%)
Meta Regression IoU				
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%)

a common size, than as a whole batch of input data inferred by the neural network, resized to their original size and than treated as an ensemble of predictions. Of this ensemble we can investigate mean and variance of dispersion measures and introduce further metrics, see [18]. Due to this modification we gain roughly 3 pp. for both meta tasks. A part of the effect is accounted to the introduction of resolution dependent uncertainty measures, while roughly an equal share stems from the deployment of neural networks for meta classification and regression.

III. TIME-DYNAMIC META CLASSIFICATION

In online applications like automated driving, video streams of images are usually available. When inferring videos with single frame based convolutional neural networks, time dynamic uncertainties such as flickering segments can be observed. Therefore, as an extension of the previously introduced MetaSeg method, we present a time-dynamic approach for investigating uncertainties and assessing the prediction quality of neural networks over series of frames (meta regression) as well as for performing false positive detection (meta classification). In order to extend the single frame metrics to time

series of metrics, we develop a light-weight tracking algorithm based on semantic segmentation, since by assumption the latter is already available. Segments in consecutive frames are matched according to their overlap in multiple frames. These measures are improved by shifting segments according to their expected location in the subsequent frame. By means of the identification of segments over time, we can extend each metric M_i (defined as a scalar quantity in the previous section) to a time series. These time series are then presented to meta classifiers and regressors to perform both meta tasks. The set of metrics used in [16] is extended in [18], this extension is deployed in the time-dynamic MetaSeg. A precise description of these metrics and of our tracking algorithm can be found in [17]. All numerical tests in this section are performed using the updated set of metrics.

Let $\{x_1, \ldots, x_T\}$ denote an image sequence with a length of T and x_t corresponds to the tth image. In what follows, we analyze the influence of the time series length on the models that perform meta classification and regression. In case when only using single frames (this corresponds to plain MetaSeg introduced in the previous section), we only present the segment-wise metrics M^t, where M^t denotes the metrics of a single frame t, to the meta classifier/regressor. For the time dynamic approach, we extend the metrics to time series considering – frame by frame – up to 10 previous frames and their metrics M^j, $j = t - 10, \ldots, t - 1$. In total, we obtain 11 different sets of metrics that are inputs for the meta classification and regression models. The presented results are averaged over 10 runs obtained by random sampling of the train/validation/test splitting. In fig. 2 and table II, the corresponding standard deviations are given by shades and in brackets, respectively. In addition to linear models used in the previous section, we also perform tests with gradient boosting and shallow neural networks with ℓ_2-penalization for both meta tasks.
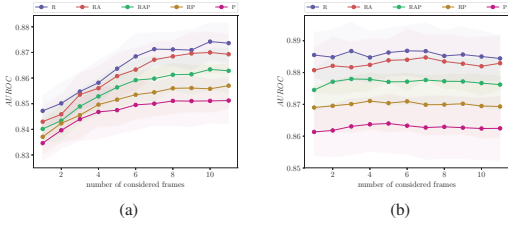
We perform tests with the KITTI dataset [19] containing

Fig. 2. A selection of results for meta classification AUROC as functions of the number of frames and for different compositions of training data. (a): meta classification via a neural network with ℓ_2-penalization, (b): meta classification via gradient boosting.

TABLE II
RESULTS FOR META CLASSIFICATION AND REGRESSION FOR DIFFERENT COMPOSITIONS OF TRAINING DATA AND METHODS. THE SUPER SCRIPT DENOTES THE NUMBER OF FRAMES WHERE THE BEST PERFORMANCE AND THUS THE GIVEN VALUE IS REACHED. THE BEST RESULTS FOR EACH DATA COMPOSITION ARE HIGHLIGHTED.

| | Meta Classification $IoU = 0, > 0$ | | | |
| | Gradient Boosting | | Neural Network with ℓ_2-penalization | |
	ACC	AUROC	ACC	AUROC
R	**81.20%**($\pm$1.02%)^4	**88.68%**($\pm$0.80%)^6	**79.67%**($\pm$0.93%)^{10}	**87.42%**($\pm$0.75%)^{10}
RA	80.73%($\pm$1.03%)^9	88.47%($\pm$0.73%)^7	78.62%($\pm$0.61%)^{11}	87.00%($\pm$0.81%)^{10}
RAP	79.64%($\pm$1.03%)^7	87.80%($\pm$0.82%)^3	77.08%($\pm$1.05%)^9	86.34%($\pm$0.84%)^{10}
RP	78.45%($\pm$0.88%)^8	87.11%($\pm$0.90%)^4	76.35%($\pm$0.67%)^9	85.70%($\pm$0.88%)^{11}
P	77.56%($\pm$0.95%)^5	86.40%($\pm$0.93%)^5	75.68%($\pm$0.67%)^{11}	85.12%($\pm$0.92%)^{11}
	Meta Regression IoU			
	Gradient Boosting		Neural Network with ℓ_2-penalization	
	σ	R^2	σ	R^2
R	0.114($\pm$0.004)^5	87.02%($\pm$1.00%)^5	**0.113**($\pm$0.005)^1	**87.16%**($\pm$1.25%)^1
RA	0.116($\pm$0.004)^3	86.39%($\pm$1.11%)^3	0.116($\pm$0.005)^1	86.46%($\pm$1.32%)^1
RAP	**0.112**($\pm$0.003)^7	**87.51%**($\pm$0.61%)^7	0.114($\pm$0.005)^1	86.97%($\pm$1.10%)^1
RP	**0.112**($\pm$0.002)^9	87.45%($\pm$0.72%)^9	0.115($\pm$0.003)^2	86.69%($\pm$0.85%)^2
P	0.114($\pm$0.002)^{11}	86.88%($\pm$0.67%)^{11}	0.117($\pm$0.004)^3	86.24%($\pm$0.99%)^3

street scene images from Karlsruhe, Germany. This dataset contains plenty of video sequences of which 29 contain ground truth. The tests we perform are based on these 29 sequences (yielding ~12K images) containing 142 labeled (semanticly segmented) single frames in total. We use the same DeepLabv3+ networks like in the previous section (pre-trained on the Cityscapes dataset) to generate the output probabilities on the KITTI dataset. In our tests we mainly use the MobilenetV2 while the stronger Xception65 network serves as a reference network as to be explained subsequently. Since an evaluation of meta regression and classification requires a train/validation/test splitting, the small amount of 142 labeled images seems almost insufficient. Hence, we acquire alternative sources of useful information besides the (real) ground truth. First, we apply a variant of SMOTE for continuous target variables for *data augmentation* (see [29], [30]) to augment the structured dataset of metrics. In addition, we utilize the Xception65 net with high predictive performance, its predicted segmentations we term *pseudo ground truth*. We generate pseudo ground truth for all images where no ground truth is available. The train/val/test splitting of the data with ground truth available is 70%/10%/20%. We use the alternative sources of information to create different compositions of training data, i.e., R (real), RA (real and augmented), RAP (real, augmented and pseudo), RP (real and pseudo) and P (pseudo). The shorthand "real" refers to ground truth obtained from a human annotator, "augmented" refers to data obtained from SMOTE and "pseudo" refers to pseudo ground truth obtained from the Xception65 net. These additions are only used during training. We utilize the Xception65 network exclusively for the generation of pseudo ground truth, all tests are performed using the MobilenetV2.

A selection of results for meta classification AUROC as functions of the number of frames, i.e., the maximum time series length, is given in fig. 2. The meta classification results for neural networks presented in subfigure (a) indeed show, that an increasing length of time series has a positive effect on meta classification. On the other hand, the results in subfigure (b) show that gradient boosting does not benefit as much from time series. In part this can be accounted to overfitting which we observe in our tests when using gradient boosting. Results

for meta regression and meta classification are summarized in table II. For gradient boosting as regression method we observe that the incorporation of pseudo ground truth slightly increases the performance. Noteworthily, we achieve almost the same performance when training gradient boosting either with pseudo ground truth exclusively or with real ground truth exclusively. This shows that meta regression can also be learned when there is no ground truth but a strong reference model available. We provide video sequences that visualize the IoU prediction and the segment tracking[1]. For further results, especially those of the linear models, we refer to [17]. The results of the linear models are below those of gradient boosting in both meta tasks and are therefore not discussed in detail, here. In contrast to the single frame approach using only linear models, we increase the AUROC by 5.04 pp. for meta classification and the R^2 by 5.63 pp. for meta regression.

IV. FALSE NEGATIVE DETECTION BY DECISION RULES

In this section, we draw attention to false-negative detection and the issue connected to the probabilistic output of segmentation networks when trained on unbalanced data, i.e., a dominant portion of pixels is assigned to only a few classes. As the softmax output of a segmentation network gives a pixel-wise class distribution over all q predefined classes, the most commonly used decision rule, also known as maximum a-posteriori probability (MAP) principle, selects the class of highest probability. This is however merely one example of a cost-based decision rule and it is by far not the only possible selection principle. One could also penalize each confusion event by a specific quantity

$$c_z(\hat{y}, y) := \begin{cases} 0 & , \text{ if } \hat{y} = y \\ \psi_z(\hat{y}, y) & , \text{ if } \hat{y} \neq y \end{cases}, \ \psi_z(\hat{y}, y) \in \mathbb{R}_{\geq 0} \quad (4)$$

that valuates the aversion of a decision maker towards the confusion of the predicted class $\hat{y}$ with the actual class y. The

---

[1]See https://youtu.be/YcQ-i9cHjLk

decision on the predicted class for pixel z given image x now minimizes the expected cost:

$$\hat{y}_z(x) = \operatorname*{arg\,min}_{y' \in \mathcal{C}} \mathbb{E}_z[\, c_z(y', Y) \mid X = x\,] \qquad (5)$$

$$= \operatorname*{arg\,min}_{y' \in \mathcal{C}} \sum_{y \in \mathcal{C} \setminus \{y'\}} \psi_z(y', y)\, f_z(y|x)\,. \qquad (6)$$

Seen from this angle, the standard MAP principle corresponds to cost functions that attribute equal cost to any confusion event, cf. (1). Although it seems reasonable, according to common human sense, to assume that $\psi_z(y', y)$ should be different depending on the type of confusion, another decision policy may reveal ethical problems when it comes down to providing explicit numbers [22]. Therefore, the choice of cost functions to increase the sensitivity towards rare objects is subjected to constraints. A way out is offered by the the mathematically appealing "natural" Maximum Likelihood (ML) decision rule which is known for its strength in finding instances of under-represented classes in unbalanced datasets [20]. The latter rule assigns costs inverse proportional to the class frequencies, i.e.,

$$\psi_z(y', y) = \frac{1}{\hat{p}_z(y)} = |\mathcal{X}| \left(\sum_{x \in \mathcal{X}} 1_{\{y_z(x)=y\}} \right)^{-1} \forall\, y' \neq y \quad (7)$$

with $\hat{p}_z(y)$ being the estimated a-priori probability (prior) from data $\mathcal{X}$ for class $y \in \mathcal{C}$ at location z. Considering a segmentation network as statistical model, the softmax output $f_z(y|x)$ can then be interpreted as a-posteriori probability of pixel z in x belonging to class y. Via the softmax adjustment with the priors

$$\hat{y}_z(x) = \operatorname*{arg\,min}_{y' \in \mathcal{C}} \sum_{y \in \mathcal{C} \setminus \{y'\}} \frac{1}{\hat{p}_z(y)}\, f_z(y|x) \qquad (8)$$

$$= \operatorname*{arg\,max}_{y \in \mathcal{C}} \frac{f_z(y|x)}{\hat{p}_z(y)} \overset{\text{(Bayes' Th.)}}{=} \operatorname*{arg\,max}_{y \in \mathcal{C}} f_z(x|y) \qquad (9)$$

the class affiliation y becomes an unknown parameter that needs to be estimated using the principle of maximum likelihood. The ML rule aims at finding the class y for which the features x are most typical, independent of any prior belief about the semantic classes such as the class frequency. We apply the ML rule in a position-specific manner in order to handle pixel-wise class imbalance, see fig. 3 and fig. 4. Results for DeepLabv3+ (Xception65 and MobilenetV2) models on Cityscapes data are reported as empirical cumulative distribution functions (CDFs) of the category *human* for segment-wise precision (F^p) and recall (F^r) in fig. 5.

We observe an advantage of Bayes in terms of precision since $F^p_{ML} \prec F^p_B$ for both models, where $\prec$ stands for 1st order stochastic dominance [31] saying that typical precision values for Bayes are right shifted compared with ML. For any precision value v, in particular for low precision values, the frequency with which an instance's precision is below v is significantly less with Bayes than with ML. In terms of recall, we observe the opposite behavior, i.e., ML is superior over Bayes in this metric. The steep ascent of the ML curves additionally indicates that most ground truth
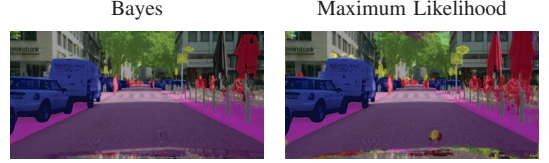
Bayes   Maximum Likelihood

Fig. 3. Illustration of two segmentation masks obtained with the Bayes decision rule (left) and the Maximum Likelihood decision rule (right).
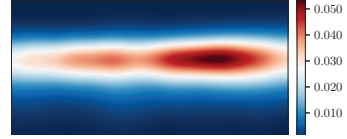

Fig. 4. Estimated pixel-wise prior probabilities of class *human* in Cityscapes. For every other category, there is another heatmap with the property that the values at each pixel position over all heatmaps sum up to 1.

segments are predicted with high recall. More relevantly, ML significantly reduces the number of non-detected segments, i.e., $F^r_B(0) > F^r_{ML}(0)$. Hence, the ML prediction can serve as uncertainty mask revealing image regions where an rare class object might be overlooked. For further reading, we refer to [21].

V. Outlook

The presented methods have clearly demonstrated their performance for false positive and false negative detection in semantic segmentation. Within this line of research, we plan to continue our work on constructing further time-dynamical metrics that quantify temporal uncertainty, as well as methods for false negative detection that do not overproduce false positives. Furthermore, transferring meta classification and regression to the task of object detection is a logical next step for future research.

Besides the mentioned algorithmic activities, several topics that can be considered as applications of false positive detection / false negative detection or segmentation quality assessment should be developed in the future. One application of meta regression is active learning for semantic segmentation. Here MetaSeg can be used as part of the query strategy.

As a future direction of reseach, work on out-of-distribution detection is a necessary next step to address the OOD failure mode. We believe that the concept of meta classification and our method MetaSeg will play a significant role. Furthermore, we expect that the development of new uncertainty measures will play a crucial role [32], [33]. It is also an interesting question, in as much approaches of uncertainty quantification that guarantee OOD - detection can be transferred to machine learning with high dimensional input data [12], [34].

On the other hand, synthetic data should be integrated in this line of method development. We expect that work on domain adaptation, active transfer learning methods as well as
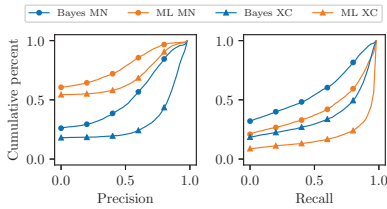
Fig. 5. Empirical cumulative distribution functions in Cityscapes for segment-wise precision and recall of class *human*.

the generation of synthetic corner cases will play a vital role in the future. For all these applications, we believe that expressive uncertainty quantification and well-performing meta classification frameworks are key-components to establish new approaches or leverage existing ones.

Source codes for our frameworks are available on GitHub, see https://github.com/mrottmann/MetaSeg.

REFERENCES

[1] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger *et al.*, "Comparison of learning algorithms for handwritten digit recognition," in *International conference on artificial neural networks*, vol. 60. Perth, Australia, 1995, pp. 53–60. 1

[2] Y. LeCun *et al.*, "Lenet-5, convolutional neural networks," *URL: http://yann. lecun. com/exdb/lenet*, vol. 20, p. 5, 2015. 1

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016. 1

[4] R. Berger, "Autonomous driving," *Think Act*, 2014. 1

[5] B. W. Smith and J. Svensson, "Automated and autonomous driving: regulation under uncertainty," 2015. 1

[6] M. Wood, P. Robbel, M. Maass, R. D. Tebbens, M. Meijs, M. Harb *et al.*, "Safety first for automated driving." [Online]. Available: https://www.daimler.com/innovation/case/autonomous/safety-first-for-automated-driving.html 1

[7] 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. 1

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788. 1

[9] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *International Journal of Multimedia Information Retrieval*, vol. 7, no. 2, pp. 87–93, Jun 2018. [Online]. Available: https://doi.org/10.1007/s13735-017-0141-z 1

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587. 1

[11] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017. 1

[12] A. Meinke and M. Hein, "Towards neural networks that provably know when they don't know," *arXiv preprint arXiv:1909.12180*, 2019. 1, 5

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *CoRR*, vol. abs/1312.6199, 2013. 1

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, 2017. 1, 2

[15] 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. 1, 2

[16] M. Rottmann, P. Colling, T. Hack, F. Hüger, P. Schlicht, and H. Gottschalk, "Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities," *CoRR*, vol. abs/1811.00648, 2018. [Online]. Available: http://arxiv.org/abs/1811.00648 2, 3

[17] 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 2, 3, 4

[18] 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 2, 3

[19] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. [Online]. Available: https://doi.org/10.1177/0278364913491297 2, 3

[20] L. Fahrmeir, A. Hamerle, and W. Häussler, *Multivariate statistische Verfahren (in German)*, 2nd ed. Walter De Gruyter, 1996. 2, 5

[21] R. Chan, M. Rottmann, F. Hüger, P. Schlicht, and H. Gottschalk, "Application of decision rules for handling class imbalance in semantic segmentation," *CoRR*, vol. abs/1901.08394, 2019. [Online]. Available: http://arxiv.org/abs/1901.08394 2, 5

[22] R. Chan, M. Rottmann, R. Dardashti, F. Huger, P. Schlicht, and H. Gottschalk, "The ethical dilemma when (not) setting up cost-based decision rules in semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 6 2019. 2, 5

[23] 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

[24] 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 2

[25] 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. 2

[26] 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. 2

[27] 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. 2

[28] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct 2015. 2

[29] N. Chawla, K. Bowyer, L. O. Hall, and W. Philip Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Intell. Res. (JAIR)*, vol. 16, pp. 321–357, 01 2002. 4

[30] L. Torgo, R. P. Ribeiro, B. Pfahringer, e. L. Branco, Paula", L. P. Reis, and J. Cascalho, "Smote for regression," in *Progress in Artificial Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 378–389. 4

[31] C. G. Pflug and R. Werner, *Modeling, measuring and managing risk*. World Scientific, 2007. 5

[32] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059. 5

[33] P. Oberdiek, M. Rottmann, and H. Gottschalk, "Classification uncertainty of deep neural networks based on gradient information," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer, 2018, pp. 113–125. 5

[34] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction," *arXiv preprint arXiv:1910.09457*, 2019. 5