Uncertainty estimates for semantic segmentation: providing enhanced reliability for automated motor claims handling

Jan Küchler[1] · **Daniel Kröll**[1] · **Sebastian Schoenen**[1] · **Andreas Witte**[1]

Abstract
Deep neural network models for image segmentation can be a powerful tool for the automation of motor claims handling processes in the insurance industry. A crucial aspect is the reliability of the model outputs when facing adverse conditions, such as low quality photos taken by claimants to document damages. We explore the use of a meta-classification model to empirically assess the precision of segments predicted by a model trained for the semantic segmentation of car body parts. Different sets of features correlated with the quality of a segment are compared, and an AUROC score of 0.915 is achieved for distinguishing between high- and low-quality segments. By removing low-quality segments, the average *mIoU* of the segmentation output is improved by 16 percentage points and the number of wrongly predicted segments is reduced by 77%.

1 Introduction

In the rapidly evolving world of automotive insurance, technological advancements are reshaping the landscape. Efficient and accurate claims handling remain key success factors for the insurance industry. At the heart of this process is damage assessment, traditionally reliant on manual methods. This procedure often required experts to either make on-site visits to inspect damaged cars or, increasingly common today, review photographs provided by claimants. While this approach is thorough, it is also time-consuming and vulnerable to human biases and errors.

Daniel Kröll, Sebastian Schoenen and Andreas Witte contributed equally to this work.

✉ Jan Küchler
j.kuechler@controlexpert.com

Daniel Kröll
d.kroell@controlexpert.com

Sebastian Schoenen
s.schoenen@controlexpert.com

Andreas Witte
a.witte@controlexpert.com

1   ControlExpert GmbH, Marie-Curie-Straße 3, 40764 Langenfeld, Germany

The advent of new computer vision techniques, particularly semantic segmentation [1], opens up possibilities to automate and streamline the damage assessment process. By segmenting images into categorized car parts and damages, this holds the potential to identify, classify and localize car damages. Embracing these techniques could empower the insurance industry to cut operational costs, expedite claim processing, and crucially, boost accuracy.

However, any technology-driven solution requires rigorous validation of its reliability. While deep neural networks (DNNs) have demonstrated exceptional performance in semantic segmentation tasks [2, 3], the variability in images of damaged cars—influenced by factors like lighting conditions, vehicle models, capture angles, and other variables—can introduce uncertainties. Addressing this challenge is of paramount importance.

Figure 1 shows an example for an image with some of the aforementioned issues, together with the semantic segmentation mask of car body parts. Among other mistakes, a small area at the rim of the rear left wheel is identified as an air intake, likely due to dirt obscuring the usual features expected for a wheel.

To ensure a reliable and trustworthy damage assessment leveraging these technologies, the incorporation of uncertainty estimates into semantic segmentation is indispensable. By doing so, the industry can not only revolutionize the dam-

🙋 Springer

Fig. 1 Photograph of a car (left), taken to highlight issues that can negatively affect the performance of DNN segmentation models: reflections, dirt and bad exposure. Result of our semantic segmentation model (right), trained to segment car body parts. Predicted segments are shown as colored overlays. A few mistakes in the prediction are highlighted by red boxes: a reflection on the door is segmented as a molding part, and a part of the rear left rim is identified as an air intake

age assessment but also make it transparent, consistent, and trustworthy, truly elevating the standards of automotive insurance claims handling.

Various approaches have been proposed to provide a measure of uncertainty in the model results for semantic segmentation. Modern architectures steadily improve the robustness of segmentation models, but they do not improve in terms of uncertainty estimation and calibration [4]. While the output scores of a DNN are correlated with the accuracy of the result, models are often overconfident and output high probabilities even for wrong results [5–7]. In general, uncertainty quantification for deep learning is a widely studied topic [8], with techniques comprising primarily Bayesian approaches and ensemble methods, but also empirical methods to estimate uncertainties. Monte Carlo dropout is used in a Bayesian framework to estimate model uncertainties [9, 10], and can be combined with test-time image augmentation to also encompass data uncertainties [11]. A technique called 'Bayes by Backprop' is an alternative approach principled in the minimization of the variational free energy and used to quantify the uncertainty in the learned weights [12]. Ensemble methods assess the uncertainty by comparing the results of multiple, slightly different models trained for the same task [13], and have been found to give a well calibrated result probability [14]. Using distillation techniques, even single models can be trained to predict the pixel-wise uncertainty in a segmentation result [15, 16], thus reducing the computational demands at inference time.

In this work, we explore the use of a meta-classification [17] model to empirically estimate the uncertainty of individual segments [18]. Although this approach is not based on a theoretical foundation, it has the advantage of neither requiring modifications to the segmentation model, nor to its training, and has a relatively low computational overhead during inference.

As detailed in the following, uncertainty measures are first defined pixel by pixel, based on the softmax probability output of the segmentation network together with the loss gradient of the last convolutional layer. They are aggregated over predicted segments, and used, together with the predicted class of a segment and its size, to build a classification model that distinguishes between well and wrongly predicted segments. The score of this classifier is used as a measure of the uncertainty in the prediction. A low uncertainty result can be automatically processed with high confidence, while a high uncertainty score can indicate the need of human oversight. In special cases, the uncertainty score can be used to improve the segmentation mask. By removing segments with a high uncertainty from the segmentation mask, the precision of the segmentation output can be improved for the cost of reducing the recall. Figure 2 shows a schematic diagram of the method.

2 Pixel- and segment-wise uncertainty measures

The output of a semantic segmentation network with a final softmax layer are the pixel-wise probabilities p_i^k for every semantic class $k = 1, \ldots, N$, with the index i running over all pixel coordinates. The predicted class for every pixel is the one with the highest probability, $\hat{c}_i = \arg\max_k p_i^k$.

The probability of the predicted class for a pixel, $\hat{p}_i = \max_k p_i^k$ quantifies the confidence in the result [19], thus $1 - \hat{p}_i$ is used as one measure of the pixel-wise uncertainty.

Following[1] [18], two further quantities are defined, measuring the dispersion of the pixel-wise probabilities:

---

[1] Rottmann et al. make the source code of their method available at https://github.com/mrottmann/MetaSeg.
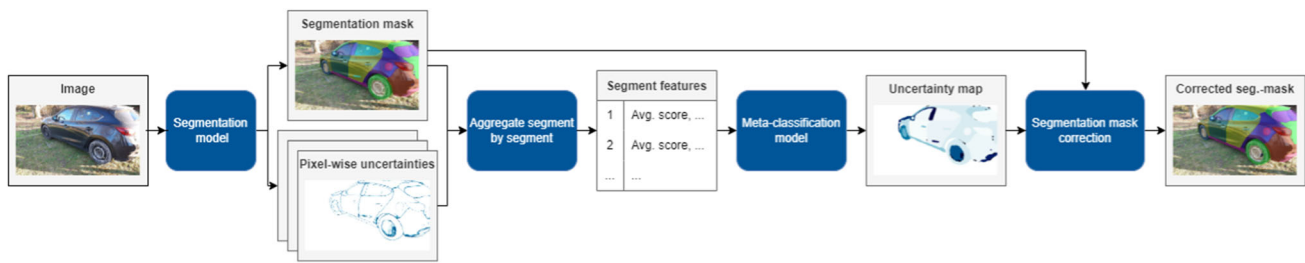
Fig. 2 Schematic diagram of the explored method. An input image is processed by a semantic segmentation model and the resulting segmentation mask and softmax probabilities are aggregated to segment wise features. These are processed by a meta-classification model in order to produce a segment-wise uncertainty map. Finally, the segment uncertainties are used to correct the segmentation mask

- the entropy

$$E_i = \frac{1}{\log K} \sum_{k=1}^{N} p_i^k \log(p_i^k),$$

which is maximized when the model result sees all classes as equally likely,
- as well as the difference between the two largest softmax values,

$$D_i = \hat{p}_i - \max_{k \neq \hat{c}_i} p_i^k,$$

which targets cases where the network predicts a similar probability for the two most likely classes.

In [20], a gradient-based approach for uncertainty quantification in semantic segmentation is introduced. The gradient of a categorical cross entropy loss with respect to the last convolutional layer of the segmentation network can be computed efficiently. When taking the predicted class $\hat{c}_i$ as the one-hot label per pixel, these gradients quantify how similar the result is to the examples in the training data set. Intuitively, larger gradients mean that the weights of the convolutional layer need to be changed more strongly to accommodate the input, therefore indicating an uncertain result. The norm of the pixel-wise gradients is taken as an additional measure of the uncertainty, which can be efficiently computed [20] as $G_i = \left\| p_i^k (1 - \delta_{k\hat{c}_i}) \psi_i \right\|_2$, with ψ_i denoting the features before the last convolution layer.

Figure 3 shows qualitative heat-maps of the pixel-wise uncertainty measures for the example image of Fig. 1. Due to the labeling accuracy, the boundaries between segments of different classes are uncertain and highlighted in the heat-maps. The wrongly predicted segments at the door and at the rim of the rear left wheel are also indicated by high pixel-wise uncertainties. On the other hand, the uncertainties vary strongly in these segments. The pixel-wise uncertainties are aggregated to segment-wise measures, in order to build features for the classification of high- and low-quality segments.

The aggregation of uncertainty estimates from pixel to segment level has been shown to improve the performance for the detection of anomalies by accounting for the correlation between neighboring pixels [21].

The predicted semantic segmentation mask for an image is split into a set $\hat{\mathcal{K}}$ of segments, i.e. connected areas of the same class. Segment by segment, the pixel-wise uncertainty measures are averaged over all pixels of the segment, e.g. the mean entropy $E(\hat{k})$ of a segment $\hat{k} \in \hat{\mathcal{K}}$ is $E(\hat{k}) = 1/|\hat{k}| \sum_{i \in \hat{k}} E_i$ and analogously for the other uncertainty measures. The values are also averaged separately over the boundary and the inner part of the segment, as defined by [18], because the boundaries typically exhibit higher uncertainties. Additionally, the standard deviation of the pixel-wise uncertainty distributions on the boundary, inner and full segment is used as an input to the meta-classification model.

The quality of segments is defined with respect to the ground truth using the measure of intersection over union [22]. The ground truth segmentation mask is split into a set $\mathcal{K}$ of segments, analogously to the prediction. Predicted segments are then compared to all ground truth segments with a matching class label and a non-trivial intersection, denoted as $\mathcal{K}|_{\hat{k}}$. For a predicted segment $\hat{k} \in \hat{\mathcal{K}}$ and the union of the matching and intersecting ground truth segments $K = \bigcup_{k \in \mathcal{K}|_{\hat{k}}} k$, the segment-wise intersection over union is defined as

$$IoU(\hat{k}) = \frac{\left| \hat{k} \cap K \right|}{\left| \hat{k} \cup K \right|}.$$

Figure 4 shows a sketch to clarify the definition of the IoU and further quality metrics, which are defined and motivated below.

The IoU penalizes scenarios in which, for example, one ground truth segment is covered by two disjoint predicted segments, which are split by a small, wrongly predicted area. Intuitively, both predicted segments describe a fraction of the ground truth segment well, even though, in the original
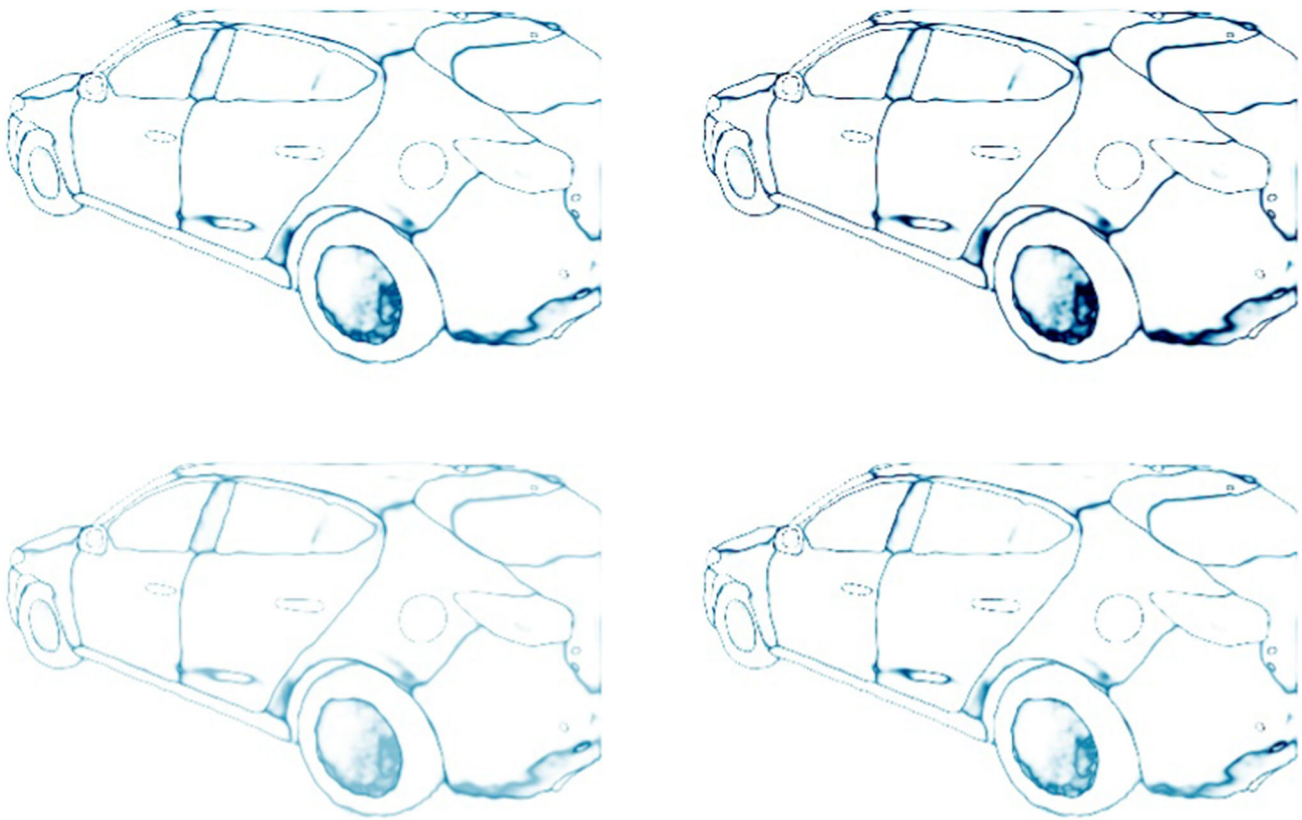
Fig. 3 Qualitative heat-maps of $1 - \hat{p}_i$ (top left), $1 - D_i$ (top right), the entropy E_i (bottom left) and the gradient uncertainty G_i (bottom right), for the example image shown in Fig. 1. Darker shades indicate higher pixel-wise uncertainties (color figure online)

definition, the IoU is small. To address this, the adjusted intersection over union, $IoU_{\mathrm{adj.}}$, is defined in [18] by restricting the denominator to the union of the predicted segment with the area of the matching ground truth segments which is not covered by other predicted segments of the same class.

In a similar fashion, we assess the quality of predicted segments by their precision,

$$p(\hat{k}) = \frac{\left| \hat{k} \cap K \right|}{\left| \hat{k} \right|},$$

i.e., the fraction of pixels in the predicted segment which overlap with a matching ground truth segment. For completely wrong predictions, i.e. without overlap of the predicted segment and the ground truth, $p = IoU = IoU_{\mathrm{adj.}} = 0$. Only for at least partially correct segments, the behavior of the metrics differ and $p \geq IoU_{\mathrm{adj.}} \geq IoU$. By choosing the precision instead of the IoU, we intentionally neglect to quantify how much of the ground truth segment is covered. For some downstream tasks using the segmentation information of a partial, but precise segment can still be valuable. As an example, a damage detected on a precise but incomplete

segment of a car body part is, in many cases, sufficient to provide a correct cost calculation.

3 Segment quality classification

The aforementioned metrics are used to train a segment meta-classifier for a semantic segmentation model for car body parts. The segmentation model is a fully convolutional DNN, distinguishing between 70 car body parts. Segment metrics and ground truth information are collected for about 3000 labeled images, which were used as a validation data set for the training of the segmentation model. An independent set of about 1000 labeled images, which was not used for the training of the segmentation model, provides a test set of segments with ground truth information.

Segments with $p > 0.5$ are labeled as correctly predicted. The threshold, τ_p, is determined from the distribution of the segment precision, c.f. Fig. 5, visual investigation of segments with varying precision and in consideration of downstream tasks. The performance of the meta-classification model does not strongly depend on the chosen precision threshold, as will be detailed below.
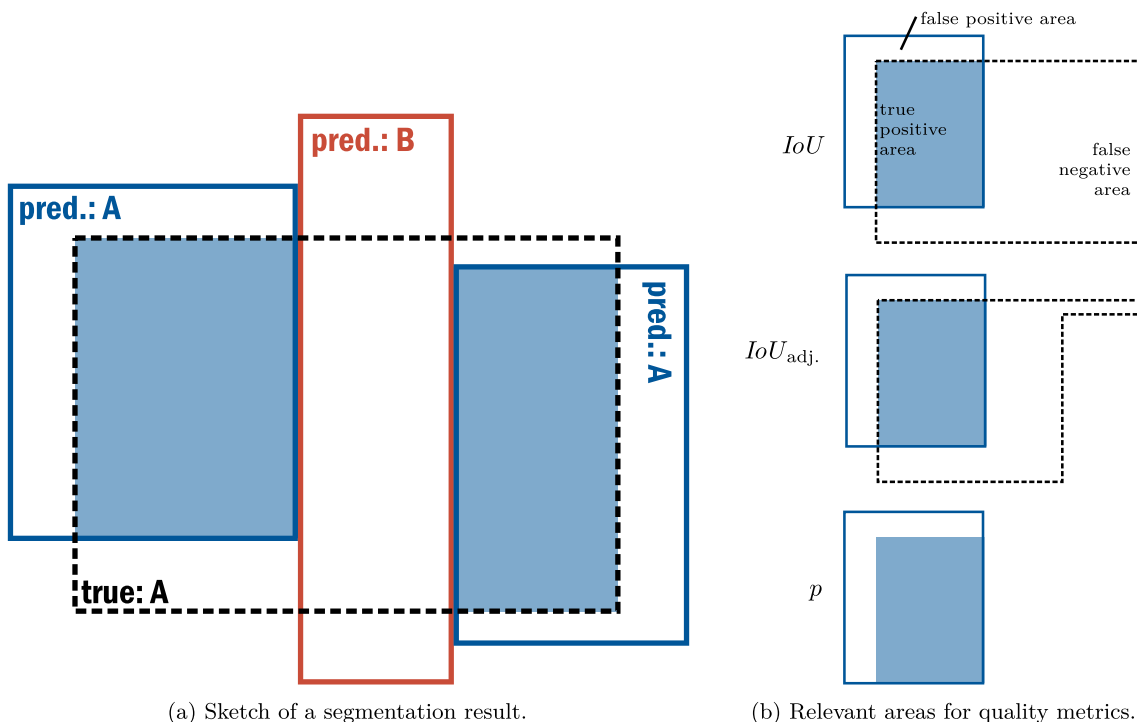
(a) Sketch of a segmentation result.

(b) Relevant areas for quality metrics.

Fig. 4 Sketch of a segmentation result and the quality metrics for one of the segments. **a** A ground truth segment of class *A* (black dashed rectangle) is covered by three predicted segments: two of class *A* (blue), divided by a segment of a different class *B* (red). The correctly segmented area is indicated by the two blue shaded rectangles. **b** The *IoU* of the left-most predicted segment is small, as it is calculated by dividing the blue shaded area by the intersection of the ground truth and the predicted segment, respectively. In contrast, for the $IoU_{adj.}$ the area covered by the other segment of class *A* is disregarded. For the precision, p, the correctly predicted area is compared only to the full predicted segment (color figure online)
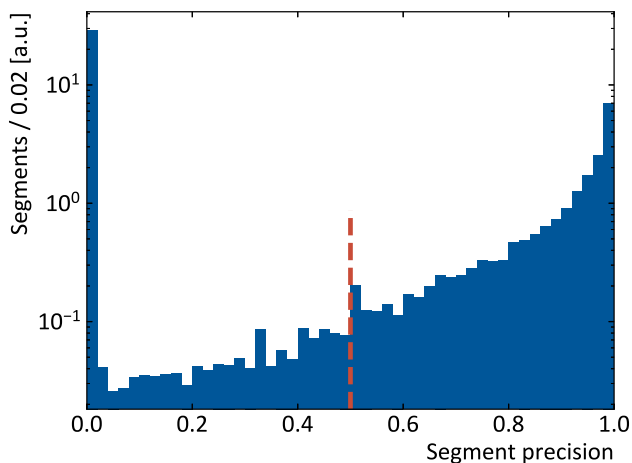

Fig. 5 Distribution of the segment-wise precision. Segments with $p > 0.5$ are selected as correct predictions. The population of segments at very low precision consists mostly of small, wrongly predicted segments

Table 1 List of segment-wise features included in the three feature sets: 'all', 'reduced', and 'uncertainty only'

Features	All	Reduced	Uncertainty only
Averages of pixel-wise uncertainties	✓	✓	✓
Relative segment size	✓	✓	
Predicted class	✓	✓	
Standard deviation of pixel-wise uncertainties	✓		
Boundary/inner pixel information	✓		

Various classification models are trained to predict the binary segment quality, i.e. classify $p > 0.5$ versus $p \leq 0.5$, and the resulting performance is compared. Different sets of features are tested, as listed in Table 1.

Two types of classifiers are tested: a gradient boosted decision tree, based on the XGBoost library [23] as a high performance method [24], as well as a linear regression classifier [25], as a simpler baseline. The XGBoost hyperparameters are optimized in a grid search employing 5-fold cross validation on the training data set.

Table 2 lists the area under the receiver operator characteristic curves (AUROC, [26]) obtained for all combinations of classifier types and segment feature sets. The precision-recall curves are displayed in Fig. 6. The XGBoost model trained

Table 2 AUROC scores for all evaluated combinations of classifier types and feature sets, with statistical uncertainties due to the size of the test data set

Feature set	XGBoost	Log. reg
All	0.916 ± 0.002	0.891 ± 0.003
Reduced	0.915 ± 0.002	0.885 ± 0.003
Uncertainty only	0.890 ± 0.003	0.882 ± 0.003
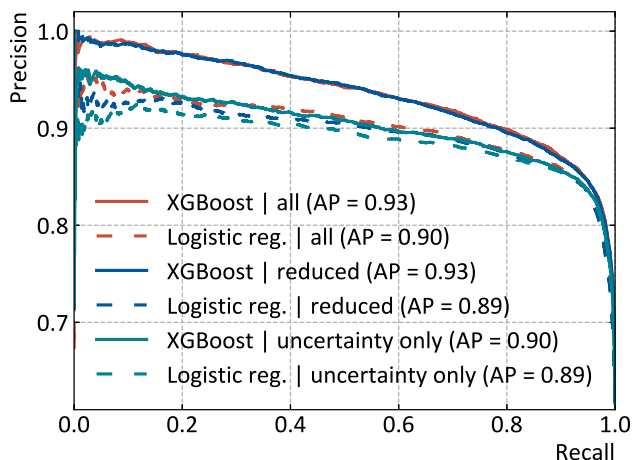
Fig. 6 Precision as a function of the recall obtainable for selecting low-quality segments for all evaluated combinations of classifier types and feature sets. The legend states the average precision AP

Fig. 7 Average segment quality in bins of the meta-classifier output. Shown are p (black), IoU (blue) and $IoU_{adj.}$ (red) for the meta-classifier trained with the nominal precision threshold $\tau_p = 0.5$, as well as p for meta-classifiers trained with $\tau_p = 0.2$ (gray dotted) and $\tau_p = 0.8$ (gray dashed). The legend lists the correlation coefficient ρ for each case (color figure online)

using all features performs best, achieving an AUROC score of $91.6\% \pm 0.2\%$ and an average precision of $93.4\% \pm 0.2\%$. Reducing the feature set by excluding the standard deviation of the uncertainty distributions and split of segment features into boundaries and inner areas only entails a minor decrease in performance. The achieved AUROC score is $91.5\% \pm 0.2\%$ with an average precision of $93.3\% \pm 0.2\%$. The results are comparable to the classification results achieved in [18] for a different model and data set.

The predicted class and the segment size are important for the performance of the XGBoost classifier. Without them, the AUROC score is reduced to $89.0\% \pm 0.3\%$ and is on par with the results obtained using the simpler logistic regression of the input features.

For further studies, the XGBoost model trained with the reduced feature set is used. The output of this meta-classification model is scaled to a range of $[0, 1]$ with higher values for segments with a low predicted quality and is used as a measure of the uncertainty for a segment. As can be seen in Fig. 7, the classifier score is strongly correlated with the segment precision ($\rho = 0.74$), and the two variants of IoU ($\rho \geq 0.90$). This correlation prevails even when choosing a different segment precision threshold to define the binary target for meta-classification.
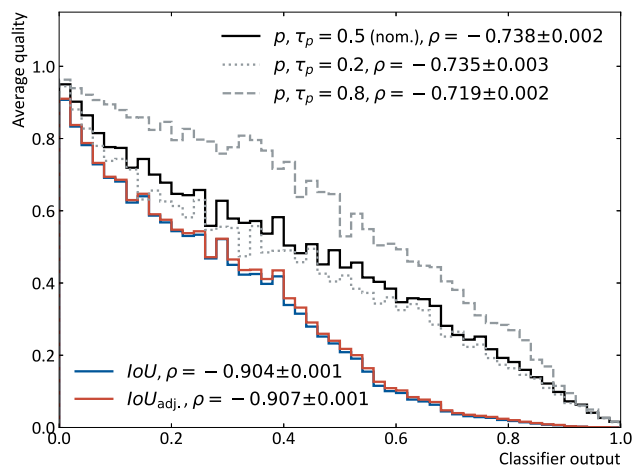
The uncertainty measure can be used to remove low-quality segments from the predicted mask. This prevents downstream tasks from including wrong predictions, which can lead to false positive results for car body parts that are not at the predicted location or not even displayed in an image. The failure modes of the segmentation model include small, wrongly predicted segments within larger areas of correct predictions. This can be caused by reflections or dirt on the surface of the car. Segments with an uncertainty larger than a specific threshold are removed from the segmentation mask, as detailed in Listing 1. If such a segment is fully enclosed by just one other segment, i.e. if all neighboring pixels have the same predicted class in the original prediction, it is replaced by the enclosing class. Otherwise, the segment is set to the "background" class, thus preventing downstream tasks from using the pixels for further results. Figure 8 shows an example of the segment-wise uncertainties and the corrected segmentation mask for the image shown in Fig. 1. The wrongly detected air intake segment at the rim is removed, preventing wrong input to subsequent processes. The wrongly predicted molding segment on the door is removed, and replaced by the surrounding door class. Figure 9 shows additional examples. Comparing the uncertainty map with the segmentation mask and the original image, it can be seen that well segmented parts have low uncertainties, while challenging areas, e.g. due to bad lighting or being in the background of the image, lead to higher segment-wise uncertainties. The mask correction procedure is able to remove many of the erroneously predicted segments.

The segment-wise uncertainty map provides comprehensive and easy-to-use information about the reliability of each segment for further applications. For example, if damages are
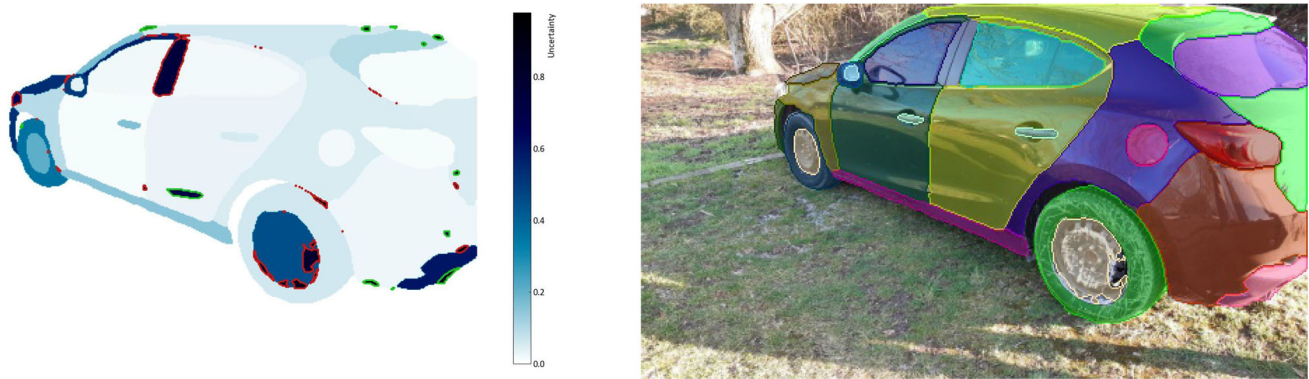
Fig. 8 Heat-map of the segment-wise uncertainties (left) and corrected segmentation mask (right) for the example image shown in Fig. 1. Within the heat-map, the colored contours show segments with an uncertainty above the threshold which are either removed and set to the background class (red), or replaced by the unambiguous surrounding class (green), as decided by the algorithm described in the text (color figure online)

Fig. 9 Additional examples, showing (from left to right) the original image, the segmentation mask, the heat-map of the segment-wise uncertainties and the corrected segmentation mask. Several mistakes, for example on the rear bumper in the upper image and on the trunk in the lower image, are removed

---

Algorithm 1 Segmentation mask correction

```
for all segments s do
    if UNCERTAINTY(s) > τ then
        | Collect neighbor segments, for example using
        | a dilation operation on the pixel mask
        n ← NEIGHBORSEGMENTS(s)
        if LEN(n) = 1 then
            CLASS(s) ← CLASS(n₀)
        else
            CLASS(s) ← BACKGROUND
        end if
    end if
end for
```

---

found only on segments with a low uncertainty, the claims handling process can be automated with high confidence in the end result. Individual high uncertainty segments can be removed from the segmentation mask, in order to improve the quality of the result.

The quality of a segmentation mask for an image can be characterized by the mean (i.e., class averaged) *IoU*. Given the sets of predicted classes, $\hat{\mathcal{C}}$, and of the classes in the ground truth labels, $\mathcal{C}$, for an image, this metric is defined as

$$mIoU = \frac{1}{\left|\hat{\mathcal{C}} \cup \mathcal{C}\right|} \sum_{c \in \hat{\mathcal{C}} \cup \mathcal{C}} \frac{tp_c}{tp_c + fp_c + fn_c},$$

where tp_c, fp_c, and fn_c are the numbers of true positive, false positive and false negative predicted pixels of class c, respectively. Notably, any class which is neither in the prediction nor in the labels does not affect the *mIoU*, while classes which are in the predicted segments but not in the ground truth labels (and vice versa) reduce the *mIoU* of a segmentation mask.

The *mIoU* is computed image by image for the original segmentation mask, as well as for the corrected mask, to quantify the impact of removing segments with a high uncertainty. Figure 10 shows the distribution of the differ-
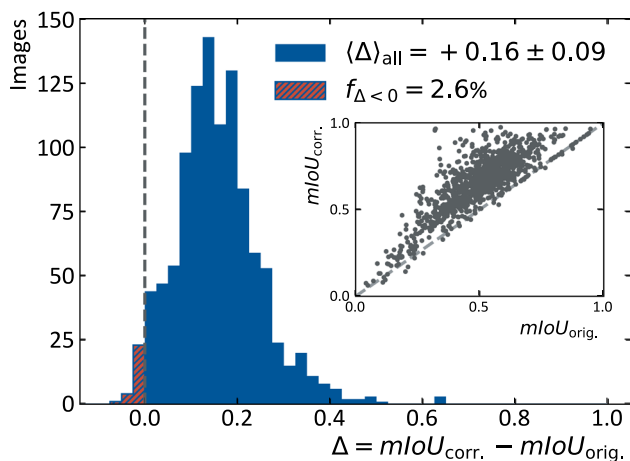
Fig. 10 Distribution of the $mIoU$ difference, Δ, between the corrected and the original mask. The red, hatched area marks entries with $\Delta < 0$, indicating a quality degradation, occurring only for $f_{\Delta<0} = 2.6\%$ of the images. The inset shows a scatter plot of the $mIoU$ of the corrected prediction in dependence on the $mIoU$ of the original mask (color figure online)

Table 3 Average $\Delta mIoU$ for images in different categories of image perspective, exposure and resolution

Perspective		
Full car		Zoom
0.16 ± 0.11		0.17 ± 0.07
Exposure		
Underexposed	Balanced exposure	Overexposed
0.14 ± 0.09	0.17 ± 0.10	0.17 ± 0.10
Resolution		
$< 1\,\mathrm{MP}$	$1\text{–}4\,\mathrm{MP}$	$> 4\,\mathrm{MP}$
0.17 ± 0.11	0.17 ± 0.10	0.18 ± 0.07

ations of 4.0 and 1.5, respectively. At the same time, a small decrease in the number of correctly predicted classes is observed as well, reducing the number from 11.2 to 10.6, with standard deviations of 7.3 and 6.9. Crucially, this reduction prevents false positive detections in downstream tasks.

4 Conclusion

In this work, the development and application of a meta-classification model is presented, which is used to assess the quality of the output of a semantic segmentation model for car body parts. Pixel-wise uncertainties are derived from the softmax probabilities and gradients, and are combined to segment-wise features. A gradient boosted decision tree classifier based on the average uncertainty features per segment has been trained to distinguish between precise and imprecise segments. The resulting meta-model achieves an AUROC score of 0.915 ± 0.002. The outputs of this classifier provide a comprehensive uncertainty measure for each segment.

In a production setting, the meta-classification model runs as a post-processing step after evaluating the car body part segmentation model. The resulting uncertainty scores are then used to remove low-quality segments from the predictions. This removal prevents false positive detections in downstream tasks and improves the segmentation mask quality for this use-case by $\overline{\Delta mIoU} = 0.16$.

The proposed method can improve the reliability of a segmentation model output. In the context of motor claims handling, this has been proven to be a valuable tool for the automation of damage assessment tasks.

Author Contributions The work presented in this paper was a collaborative effort. Dr. Jan Küchler, as the corresponding author, played a leading role in the conceptualization and overall direction of the research. Daniel

ence between these two values. On average over all images, the $mIoU$ is improved by $\overline{\Delta mIoU} = 0.16$, corresponding to an increase of the average $mIoU$ from 0.50 to 0.66. The standard deviation of the distribution of $\Delta mIoU$ is 0.09 on the test set. For $> 97\%$ of the images in the test set an improvement of the result is observed. In the rare cases that the correction procedure results in a $mIoU$ decrease, usually small, irregularly formed but precise segments within the larger area of a misidentified car body part are removed. Figure 10 also shows the $mIoU$ values for corrected masks in dependence on the uncorrected result. The method yields improvements over a large range of $mIoU$.

In order to study the robustness of the correction procedure, images are grouped into different categories. Different image perspectives bring different challenges to the model: images showing the full car have smaller relative segment sizes, while zoom images can lack helpful context. The exposure of the image could have an impact on the procedure, as under- or over-exposed areas effectively hide information. Lastly, the image resolution is an important factor for the overall image quality. Table 3 lists the average improvement $\Delta mIoU$ due to the correction procedure for images in different categories. The individual results agree well with the overall average, showing that the method is robust under the tested effects.

A major factor of the improvement is the removal of small segments, in turn leading to a wrongly predicted class being removed from the mask entirely. Even though only a small fraction of pixels in the image is affected, the effect on the $mIoU$ is significant because every class has the same weight. The number of wrongly predicted classes per image is reduced from 6.3 to 1.4, on average, with standard devi-

Data Availability The authors do not have the permission to make the datasets generated and analysed during the study publicly available.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Shotton, J., Winn, J., Rother, C., Criminisi, A.: TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) Computer Vision—ECCV 2006, pp. 1–15. Springer, Berlin (2006). https://doi.org/10.1007/11744023_1

2. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3349–3364 (2021). https://doi.org/10.1109/TPAMI.2020.2983686

3. Yu, Y., Wang, C., Fu, Q., Kou, R., Huang, F., Yang, B., Yang, T., Gao, M.: Techniques and challenges of image segmentation: a review. Electronics (2023). https://doi.org/10.3390/electronics12051199

4. Jorge, P., Volpi, R., Torr, P.H.S., Rogez, G.: Reliability in semantic segmentation: are we on the right track? In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7173–7182 (2023). https://doi.org/10.1109/CVPR52729.2023.00693

5. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans. Med. Imaging **39**(12), 3868–3878 (2020). https://doi.org/10.1109/tmi.2020.3006437

6. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 41–50 (2019). https://doi.org/10.1109/CVPR.2019.00013

7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning—Volume 70. ICML'17, pp. 1321–1330. JMLR.org, Sydney, NSW, Australia (2017)

8. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S.: A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf. Fusion **76**, 243–297 (2021). https://doi.org/10.1016/j.inffus.2021.05.008

9. Dechesne, C., Lassalle, P., Lefèvre, S.: Bayesian U-Net: estimating uncertainty in semantic segmentation of earth observation images. Remote Sens. (2021). https://doi.org/10.3390/rs13193836

10. Kendall, A., Badrinarayananm, V., Cipolla, R.: Bayesian SegNet: model Uncertainty in deep convolutional encoder–decoder architectures for scene understanding. In: Kim, T.-K., Zafeiriou, S., Brostow, G., Mikolajczyk, K. (eds.) Proceedings of the British Machine Vision Conference (BMVC), pp. 1–12. BMVA Press, London (2017). https://doi.org/10.5244/C.31.57

11. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019). https://doi.org/10.1016/j.neucom.2019.01.103

12. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 1613–1622. PMLR, Lille (2015)

13. Lee, H.J., Kim, S.T., Lee, H., Navab, N., Ro, Y.M.: Efficient ensemble model generation for uncertainty estimation with bayesian approximation in segmentation (2020)

14. Ng, M., Guo, F., Biswas, L., Petersen, S.E., Piechnik, S.K., Neubauer, S., Wright, G.: Estimating uncertainty in neural networks for cardiac MRI segmentation: a benchmark study. IEEE Trans. Biomed. Eng. **70**(6), 1955–1966 (2023). https://doi.org/10.1109/TBME.2022.3232730

15. Landgraf, S., Wursthorn, K., Hillemann, M., Ulrich, M.: DUDES: Deep uncertainty distillation using ensembles for semantic segmentation (2023). arXiv:2303.09843

16. Holder, C.J., Shafique, M.: Efficient uncertainty estimation in semantic segmentation via distillation. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3080–3087 (2021). https://doi.org/10.1109/ICCVW54120.2021.00343

17. Lin, W.-H., Hauptmann, A.: Meta-classification: combining multimodal classifiers. In: Zaïane, O.R., Simoff, S.J., Djeraba, C. (eds.) Mining Multimedia and Complex Data, pp. 217–231. Springer, Berlin (2003). https://doi.org/10.1007/978-3-540-39666-6_14

18. Rottmann, M., Colling, P., Paul Hack, T., Chan, R., Huger, F., Schlicht, P., Gottschalk, H.: Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, Glasgow (2020). https://doi.org/10.1109/IJCNN48605.2020.9206659

19. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: 5th International Conference on Learning Representations, ICLR 2017. OpenReview.net, Toulon (2017). https://openreview.net/forum?id=Hkg4TI9xl

20. Maag, K., Riedlinger, T.: Pixel-wise gradient uncertainty for convolutional neural networks applied to out-of-distribution segmentation (2023). arXiv:2303.06920

21. Grcić, M., Šarić, J., Šegvić, S.: On advantages of mask-level recognition for outlier-aware segmentation (2023). arXiv:2301.03407

22. Jaccard, P.: The distribution of the flora in the alpine zone. New Phytol. **11**(2), 37–50 (1912). https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

23. Chen, T., Guestrin, C.: XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York (2016). https://doi.org/10.1145/2939672.2939785

24. Shwartz-Ziv, R., Armon, A.: Tabular data: deep learning is not all you need. Inf. Fusion **81**, 84–90 (2022). https://doi.org/10.1016/j.inffus.2021.11.011

25. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011). arxiv:1201.0490

26. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. ICML'06, pp. 233–240. Association for Computing Machinery, New York (2006). https://doi.org/10.1145/1143844.1143874

Jan Küchler studied physics at the University of Wuppertal and continued his research in particle physics as a member of the ATLAS collaboration during a fellowship at the DESY research institute. Afterwards, he joined the research and development department at ControlExpert GmbH. The focus of his work lies in the development of computer vision models and their application to business needs.

Daniel Króll studied physics at the University of Bonn and continued his research as a PhD student in astrophysics at the technical University of Munich. Afterwards, he joined the research and development department at ControlExpert GmbH as a data scientist. Later he advanced to the team lead position of the computer vision team at ControlExpert GmbH focussing on supervising and managing the development of computer vision models and their application in products.

Sebastian Schoenen Director of Innovation & Technology at ControlExpert, is responsible for the Research & Development department with more than 30 data scientists, where he drives the development of artificial intelligence for motor claims management. After completing his doctorate in physics at RWTH Aachen University with a specialization in big data analytics and machine learning, he expanded his expertise as a data scientist and team leader in automatic image recognition and predictive modeling.

Andreas Witte was with ControlExpert until the end of 2023 and, as Managing Director, responsible for all topics related to product and technology development. Before joining ControlExpert, he worked for the Fraunhofer Institute and RWTH Aachen University. Andreas Witte studied mechanical engineering at RWTH, where he completed his doctorate in production technology.