Conformal Semantic Image Segmentation: Post-hoc Quantification of Predictive Uncertainty

Luca Mossina[1,†] Joseba Dalmau[1] Léo Andéol[2,3]
[1] IRT Saint Exupéry, Toulouse, France
[2] Institut de Mathématiques de Toulouse, Toulouse, France
[3] SNCF, Saint-Denis, France
[†] luca.mossina@irt-saintexupery.com

Abstract

We propose a post-hoc, computationally lightweight method to quantify predictive uncertainty in semantic image segmentation. Our approach uses conformal prediction to generate statistically valid prediction sets that are guaranteed to include the ground-truth segmentation mask at a predefined confidence level. We introduce a novel visualization technique of conformalized predictions based on heatmaps, and provide metrics to assess their empirical validity. We demonstrate the effectiveness of our approach on well-known benchmark datasets and image segmentation prediction models, and conclude with practical insights.

1. Introduction

Despite the success of Machine Learning (ML) and Deep Learning (DL) models in challenging computer vision tasks such as object detection [12, 49] or image segmentation [43, 51], the complexity of the models makes them akin to black boxes. It is difficult to define and assess their trustworthiness, which hinders their adoption in safety-critical industrial applications [1, 29, 39], and complicates their certification processes [21, 37]. In assessing a models trustworthiness, the lack of rigorous uncertainty estimates for ML predictions can be a major drawback, notably in the case of Semantic Image Segmentation (SIS) [43]. Most segmentation models provide softmax scores (i.e., probability-like scores) for every pixel of an input image; at inference, one builds a segmentation mask by taking the class whose score is the highest, pixel-wise. However, softmax scores are known to be overly confident and ill-calibrated [22, 24]; they tend to yield scores very close to one for the maximum softmax value, sometimes even for ambiguous inputs. For this reason, softmax values, even if useful for classification purposes, cannot be directly used as measures of uncertainty.

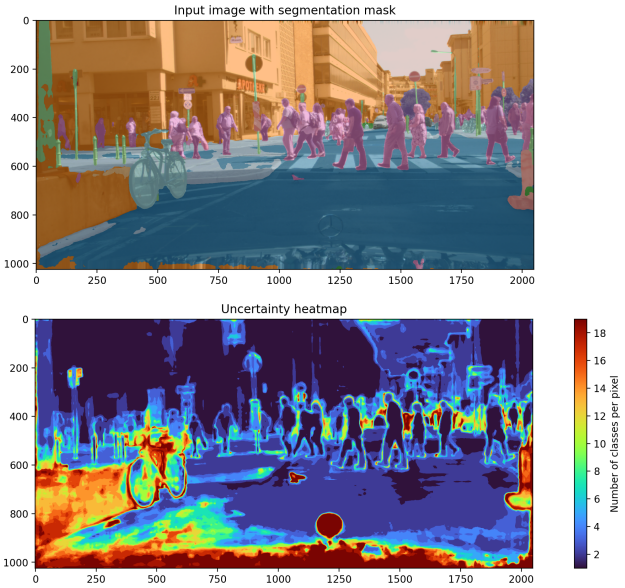

Figure 1. Top: A predicted semantic segmentation mask, overlayed on the input image, for the dataset *Cityscapes* [18].
Bottom: A *varisco* uncertainty heatmap, for a user-defined risk $\alpha = 0.01$ and a minimum coverage ratio τ of 99%; it is defined in Eq. (1) and statistically valid as in Eq. (2) of Conformalized Risk Control (CRC): every pixel is a prediction set that contains the highest scoring label (top-1) but potentially also the second, third, etc., highest scoring labels.

Contributions We introduce a method based on Conformal Prediction (CP) [45, 56] to assess the predictive uncertainty of a pre-trained segmentation predictor $\widehat{f}$. Our procedure works with any model $\widehat{f}$ (provided that it outputs softmax scores for each pixel) regardless of its architecture and the distribution of the training data; notably, this covers the case of $\widehat{f}$ being only accessible via an API or being prohibitively expensive to retrain. Our method quantifies the uncertainty of the predictor $\widehat{f}$ in the form of *segmentation*

multi-labeled masks, that is, segmentation masks that can take multiple labels per pixel. Following the conformal algorithm of [42], we build multi-labeled masks as follows: given a *coverage parameter* $\lambda \in [0,1]$, the mask $\mathcal{C}_\lambda(X)$ consists, for each pixel in the image, of the labels c having a softmax value higher than $1-\lambda$. That is, $\forall$ pixel ij,

$$\mathcal{C}_\lambda(X)_{ij} = \left\{\text{classes } k : \widehat{f}_{ijk}(X) \geq 1-\lambda\right\}. \quad (1)$$

As it can be seen in Figure 2, larger values of the coverage parameter λ produce multi-labeled masks with more classes per pixel, while smaller values of λ produce multi-labeled masks with less classes per pixel. In order to choose the right value for the coverage parameter λ, the user predefines a notion of "risk" (or "error") via a loss function ℓ and a maximum tolerable risk α. With the sole requirement of procuring held-out calibration data, one estimates $\hat{\lambda}$ from the calibration data that give rise to the finite-sample, model-agnostic and marginal guarantee of conformal prediction [1]

$$\mathbb{E}\left[\ell(\mathcal{C}_{\hat{\lambda}}(X_{\text{test}}), Y_{\text{test}})\right] \leq \alpha. \quad (2)$$

The probabilistic guarantee in Eq. (2) holds under a minimal assumption on the data generation process: calibration and test data are i.i.d. and statistically independent of the training data. We also show how these multi-labeled masks can be visualized by uncertainty *varisco* (visual assessment of risk control) heatmaps, which are computed *post-hoc* with the information of softmax scores. The code to test our methods can be found at https://github.com/deel-ai-papers/conformal-segmentation

2. Background

Semantic Image Segmentation. Semantic Image Segmentation (SIS) is the task of assigning labels to pixels in an image. Let $\mathcal{X}$ be the set of pixel values (typically $\mathcal{X} = [0,1]$ for grey-scale images and $\mathcal{X} = [0,1]^3$ for color images). An image X of H pixels of height and W pixels of width is encoded as the tensor $X = \{x_{ij} \in \mathcal{X} : ij \in \mathcal{I}_{HW}\}$, where $\mathcal{I}_{HW} := \{1,\dots,H\} \times \{1,\dots,W\}$ represents the set of indices of the pixels in the image.

Let $\mathcal{L} = \{1,2,\dots,K\}$ be a set of labels (or "classes"); each pixel x_{ij} is associated to one label $y_{ij} \in \mathcal{L}$. The set $Y = \{y_{ij} \in \mathcal{L} : ij \in \mathcal{I}_{HW}\}$ is commonly referred to as the *segmentation mask* of the image X (see Fig. 1), and the goal of the SIS task is to infer the segmentation mask Y given the image X. This is typically done by training a predictor $\widehat{f}$ that outputs softmax values for each pixel.

Conformal Prediction. Conformal Prediction (CP) [2, 56] is a family of uncertainty quantification techniques that provide model-agnostic, finite-sample guarantees on the predictions of ML models. The most common CP technique, split CP [45], is applied post-hoc on a trained model $\widehat{f}$. It requires a calibration dataset $\{(X_1,Y_1),...,(X_nY_n)\}$ independent of the training data, and an acceptable error rate $\alpha \in (0,1)$ set by the user. Split CP uses nonconformity scores (to be understood as a form of measure of prediction error) computed on the calibration dataset in order to build a prediction set $\mathcal{C}_\alpha(X_{n+1})$ for a new test sample X_{n+1}. The guarantee achieved by using split CP is

$$P\big(Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1})\big) \geq 1-\alpha. \quad (3)$$

The only assumption is that the calibration and test data form an exchangeable sequence (a condition weaker than forming an i.i.d. sequence) and that they are independent of the training data. The main limitation of CP is that the guarantee in Eq. (3) is marginal, i.e. it holds on average over both the choice of the calibration dataset and the test sample.

Conformal Risk Control. In many applications, errors of different nature may have a different severity, a false negative vs a false positive in a tumor detection application. The notion of severity of an error can be captured via a risk or an error function. CRC [5] generalizes the ideas of Conformal Prediction to this setting: prediction sets are guaranteed to keep the expected risk below a user pre-defined level α. We show how to adapt this approach to SIS in Section 4.2. Note that when using binary losses, the guarantee of CRC is the same as that of CP in Eq. (3).

3. Related works

State-of-the-art ML predictors, based on deep learning, are so complex that they are commonly approached as black boxes: the users provide some input data (an image) and they retrieve a prediction. How accurate are these models? The study of this subject is known as Uncertainty Quantification (UQ) and is a key element towards building trustworthiness in systems powered by ML models [40, 41].

Uncertainty is commonly conceptualized [28, 35] as having an *aleatoric* and *epistemic* component. Aleatoric uncertainty is inherent to the modeled phenomenon and non reducible. Epistemic uncertainty, on the contrary, stems from the fact that the models we use do not capture the phenomenon being modeled faithfully enough, and can usually be reduced by taking into account new observations or by enriching the model family being used. CP provides an estimation of the global uncertainty in the model's predictions, since it is post-hoc and with minimal hypotheses.

[1] More precisely, this is the guarantee provided by CRC [5] which has CP as a special case.

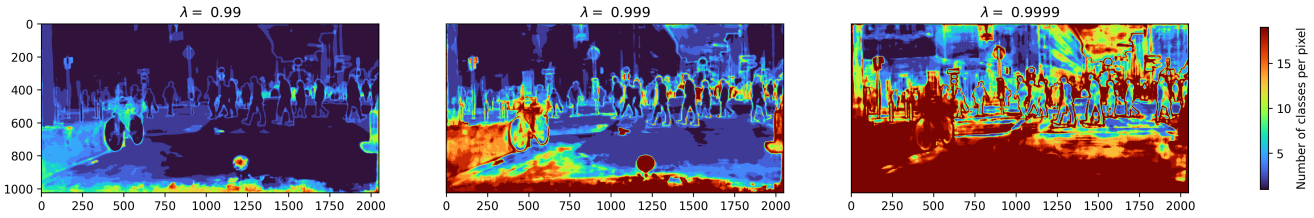

Figure 2. For three (arbitrary) values $\lambda \in \{0.99, 0.999, 0.9999\}$, we apply Eq. (7) to every pixel and obtain *varisco* heatmaps, for the dataset *Cityscapes* [18]. The CRC algorithm described in Sec. 4.2.1 searches for the optimal λ such that, for a given conformalization loss and a risk level α, the guarantee in Eq. (2) is attainable.

3.1. Uncertainty quantification methods for semantic image segmentation

Some approaches to UQ leverage model architectures that provide not only a point prediction but also the associated uncertainty, notable examples include Bayesian neural networks, networks based on Monte Carlo techniques and Deep Ensembles; others fit auxiliary models. [33] use Monte Carlo dropout neural network as well as Bayesian neural networks; [44] combine Gaussian Discriminant Analysis to estimate the density of the feature-space with the entropy measure of the softmax predictions in order to disentangle epistemic and aleatoric uncertainty and estimate each of them separately; in a similar vein, [31, 32] train a bayesian neural network in order to estimate point-predictions, aleatoric uncertainty and epistemic uncertainty simultaneously. [53] work on reducing the uncertainty of SIS models within the framework of model adaptation, that is, when domain adaptation is to be performed in the absence of source data. We also refer to [17], who work on failure prediction, a subject related to UQ. They train an auxiliary model to score the confidence of a prediction. They detect when a predictor has made a wrong prediction and assess this via a confidence measure. With respect to these approaches, for our contribution we restrict the scope to post-hoc methods: we suppose to be given a pretrained predictor that we cannot modify and that the training data are not accessible, while providing a theoretical guarantee on the UQ.

3.1.1 Calibration of image segmentation

A well-established approach to UQ is that of the *calibration* of ML models that output (pseudo) probabilities for the labels, where either auxiliary models or empirical adjustments to models are employed. It is known [22] that these scores do not admit a valid probabilistic interpretation, notably for deep-learning models based on the minimization of cross-entropy. [22], among others, brought this concept to the attention of the ML community, studying how calibration methods such as Platt scaling [48] may yield a probabilistically valid interpretation. This notion of uncertainty is applicable [20, 30, 38, 52, 63] to semantic image segmentation, where each pixel embeds a multiclass classification problem. [19, 52], for instance, give some empirical results on the effect of several methods on calibration errors. Similary, [10] have used calibration methods to address the issue of domain shift. Recently, [57] proposed selective scaling as a means to calibrate the segmentation softmax scores. These methods could be used as a complement to CP [61], at the cost of training an additional model. Finally, some literature [23, 24] does provide theoretically-founded, distribution-free algorithms for calibration and study their connection to CP [25]. Our work is based on Conformal Prediction, which is not a calibration method, but rather a different technique of UQ. As such, it is complementary to calibration, and can be used both with a model's original output as well as with an output that has been previously calibrated.

3.2. Applications of CP to segmentation

[5] use their CRC to control the false negative rate in tumor segmentation. Also based on risk control, albeit using different mathematical frameworks, the contributions of [3, 8, 46] extend the concept of tolerance regions to ML problems. They offer stronger guarantees at the cost of inferior sample efficiency. We refer to [2] for an introduction. Of these, [8] apply their methods to binary segmentation of medical images.

As for existing work using CP based on nonconformity scores, [60] apply CP to medical imaging, building pixel-wise confidence scores based on nonconformity scores and p-values [56]. [55] compute the nonconformity scores in the feature space and present an application to image segmentation. For the case of CP in imaging, we also point out to the literature on image-to-image regression (image reconstruction) [4, 9, 34, 54] which builds intervals for each output pixel. Previous work using CRC for semantic image segmentation focuses on the binary segmentation case. To the best of our knowledge, our work is the first that addresses the multi-class segmentation task with the theoretical guarantee of conformal risk control.

4. Conformal Semantic Segmentation

The goal of conformal semantic segmentation is to produce prediction sets that remain below a user-specified risk. The prediction sets are then used to assess the behaviour of the underlying predictor $\widehat{f}$ together with the problem data.

4.1. Multi-labeled masks.

A prediction set will take the form of a *multi-labeled mask*, that is, a tensor

$$Z = \big\{ z_{ijk} \,:\, ij \in \mathcal{I}_{HW}, k \in \mathcal{L} \big\}, \tag{4}$$

where $(z_{ijk})_{k=1}^{K} \in \{0,1\}^K$ encodes the subset of labels corresponding to the pixel ij; Note that this tensor has as many channels as the number of classes, where each channel is a binary segmentation mask (class k vs others).

For a multiclass segmentation mask Y, its one-hot encoding $h(Y)$ is a particular instance of a multi-labeled mask: every pixel has exactly one channel (out of K) with value one. We say that a multi-labeled mask Z contains a multi-labeled mask Z' and we write $Z \geq Z'$, if $z_{ij} \geq z'_{ij}$ for each pixel (i,j).

4.1.1 Nested multi-labeled masks.

Let $\widehat{f}$ be any semantic segmentation predictor that produces pixel-wise softmax scores, that is, for an image X, we have

$$\widehat{f}(X) := \big\{ \widehat{f}_{ijk}(X) \,:\, ij \in \mathcal{I}_{HW}, k \in \mathcal{L} \big\}, \tag{5}$$

with $\widehat{f}_{ijk}(X) \in [0,1]$ and $\sum_{k=1}^{K} \widehat{f}_{ijk}(X) = 1$. Our baseline conformal segmentation method builds prediction multi-labeled masks based on the point-predictor $\widehat{f}$, via the Least Ambiguous Set-Valued Classifiers (LAC) [42]. Given $\lambda \in [0,1]$ and a probability $p \in [0,1]$, we define the thresholding $T_\lambda(p)$ by setting:

$$T_\lambda(p) = \begin{cases} 1 & \text{if } p \geq 1-\lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The LAC mapping on the whole image X is defined by applying the mapping T_λ to the tensor $\widehat{f}(X)$

$$\mathcal{C}_\lambda^{\text{LAC}}(X) = \big\{ T_\lambda\big(\widehat{f}_{ijk}(X)\big) \,:\, ij \in \mathcal{I}_{HW}, k \in \mathcal{L} \big\}. \tag{7}$$

The multi-labeled masks generated via the LAC procedure are nested, i.e.

$$\lambda \leq \lambda' \quad \Rightarrow \quad \mathcal{C}_\lambda^{\text{LAC}} \leq \mathcal{C}_{\lambda'}^{\text{LAC}}. \tag{8}$$

Note that following Eq. (11), for high values of α, one can get a valid $\hat{\lambda}$ that can produce some empty pixels (i.e. there is no class with a high-enough score) when plugged into Eq. (7). To prevent this, we always include the most highly scored class in the multi-labeled mask (or "top-1" class).

4.2. Conformal Risk Control for multi-labeled mask

Instead of working with loss functions that compare ground-truth values to point-predictions, Conformalized Risk Control (CRC) considers loss functions that compare ground-truth values to set-predictions. For the particular application of semantic segmentation, we consider a loss function $\ell(Z,Y)$ taking as arguments a multi-labeled mask Z and a one-hot encoded mask Y. We assume that ℓ takes values in the bounded interval $(-\infty, B]$ for some $B \in \mathbb{R}$, and that it is non-increasing in Z:

$$\forall Y,\ \forall Z \leq Z' \quad \Rightarrow \quad \ell(Z,Y) \geq \ell(Z',Y), \tag{9}$$

i.e. larger multi-labeled masks produce smaller loss values. This assumption, together with the nestedness of the LAC masks imply that the mapping $\lambda \mapsto \ell\big(\mathcal{C}_\lambda^{LAC}(X), Y\big)$ is non-increasing in λ. The loss ℓ is not used as a training loss, that is, applying CRC does not affect the pre-trained model nor the predictive algorithm directly. The loss ℓ rather allows the user to *encode the notion of error* [5] in their predictions. In order to be clear, the loss ℓ will be referred to as *conformalization loss*.

Let us consider a sequence $(X_i, Y_i)_{i=1}^{n+1}$ of images with their corresponding ground-truth segmentation masks. The first n examples constitute our calibration set $\mathcal{D}_{\text{cal}}$ and the example $n+1$ is taken to be the test example. We denote $L_i(\lambda) := \ell\big(\mathcal{C}_\lambda(X_i), Y_i\big)$ the loss on the i-th example, one can then compute the *empirical risk* of the prediction sets on calibration data $\mathcal{D}_{\text{cal}}$ as

$$\widehat{R}_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} L_i(\lambda). \tag{10}$$

The purpose of the calibration set is to estimate the right value $\hat{\lambda}$ with the guarantee that the risk will remain below the maximum tolerated risk level. Given a maximum tolerated risk level $\alpha \in \mathbb{R}$, we define

$$\hat{\lambda} := \inf \left\{ \lambda \in [0,1] : \frac{n}{n+1} \widehat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha \right\} \tag{11}$$

Theorem 4.1 (Theorem 1 in [5].)**.** *Assume that the $L_i(\lambda)$ are non-increasing, right-continuous and bounded by $B < +\infty$. Assume that there exists $\lambda_{max} \in [0,1]$ such that $L_i(\lambda_{max}) \leq \alpha$. Assume further that $L_1(\lambda), \ldots, L_{n+1}(\lambda)$ form an exchangeable sequence. Let $\hat{\lambda}$ be computed as in Equation (11). Then it holds that*

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda}) \right] \leq \alpha. \tag{12}$$

Computing the optimal $\hat{\lambda}$. [5] do not provide an explicit optimization algorithm to find the optimal parameter in Eq. (11). Because of the hypothesis of monotonicity of $L_i(\lambda)$ with respect to λ, this can be achieved, for instance, running a dichotomic search over the parameter λ, up to any user-defined error ϵ.

4.2.1 Conformalization algorithm

When we say we "conformalize" a ML predictor, we mean computing the losses and the optimal $\hat{\lambda}$ on some calibration data. More specifically:

Algorithm 1: Conformalization of Semantic Image Segmentation, setup and estimation of $\hat{\lambda}$.

Data: Predictor $\widehat{f} \in [0,1]^{K\times H\times W}$. Prediction set parametrization $\mathcal{C}_\lambda(\cdot)$.
Result: $\hat{\lambda}$
1 Collect calibration data $\mathcal{D}_{\text{cal}} = (X_i, Y_i)_{i=1}^n$ from the same distribution as the test data;
2 Choose a conformalization loss $\ell(\mathcal{C}_\lambda(X), Y) \in [0,1]$ (see Sec. 4.3);
3 Set an acceptable risk level $\alpha \in (0,1)$;
4 Compute $\hat{\lambda}$ as in Eq. (11): since the empirical risk is monotonic w.r.t. λ, dichotomic search is a fast option;

Note that to ensure the statistical validity, one must pick a value α before observing the calibration data: like in statistical hypotheses testing, one cannot adjust their significance level α after computing the p-values. In practice, one could use two calibration datasets, the first to explore CP on the use case and the second reserved to the estimation of the $\hat{\lambda}$ to be deployed in production.

We say we have a "conformalized prediction" when we build the prediction set with the $\hat{\lambda}$ as computed above, applying Eq. (7).

Algorithm 2: Conformalization of Semantic Image Segmentation, inference.

Data: Input image $X \in [0,1]^{3\times H\times W}$, Predictor $\widehat{f} \in [0,1]^{K\times H\times W}$. Prediction set parametrization $\mathcal{C}_\lambda(\cdot)$
Result: $Z = \mathcal{C}^{\text{LAC}}(X)$
1 Compute $\mathcal{C}^{\text{LAC}}(X)$: apply Eq. (7) to X.

4.3. Choosing the loss function.

In Fig. 3 we show how, for the same risk level $\alpha = 0.1$, different losses generate different *varisco* heatmaps. These losses encode different notions of error: from the left to right-hand side, we see a shift from stricter to less strict. When implementing a CP method, the users need to choose a conformalization loss ℓ suitable to their problem. For the guarantee of CRC to hold, one needs to ensure that the losses $L_i(\lambda)$ are non-increasing with respect to λ; since the LAC procedure in Eq. (7) produces nested prediction sets [26], it is enough to ensure that $\ell(Z, Y)$ is non-decreasing with respect to the first argument Z. In this section we give three examples of natural choices for losses that respect this condition.

The first is a **binary loss**, which yields a guarantee equivalent to that of CP based on nonconformity scores, whose underlying loss would be $\ell(\mathcal{C}(X), Y) = \mathbb{1}\{Y \notin \mathcal{C}_\alpha(X)\}$. It takes value one whenever the prediction set does not contain the true value Y. We write as

$$\ell_{\text{bin}}(Z, Y) = \begin{cases} 0 & \text{if} \quad Z \geq Y \\ 1 & \text{otherwise.} \end{cases} \tag{13}$$

In the case of conformalized SIS, that happens when the multi-labeled mask does not cover every pixel in the image. Empirically, the conformalization of segmentation produces very small values $\hat{\lambda}$ that result in multi-labeled masks corresponding to (almost) the whole target space $\mathcal{Y} = \{1, 1, \ldots, 1\}^{K\times H\times W}$ for each inference (*e.g.* left-hand side in Fig. 3).

One can however set an acceptable trade-off in coverage, with a *minimum coverage ratio* τ: the user specifies a priori the minimal proportion of pixels that need to be covered for a prediction to be considered successful. We thus define the **binary loss with threshold** as

$$\ell_\tau(Z, Y) = \begin{cases} 1 & \text{if} \quad \frac{\sum_{ijk} Z_{ijk} Y_{ijk}}{\sum_{ijk} Y_{ijk}} < \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

where τ is the minimum acceptable coverage ratio. For $\tau = 1$ we recover the binary loss in Eq. (13).

Binary losses constitute a strict criterion: during conformalization, for $\tau = 90\%$, an empirical coverage of 89.9% will be considered a failure. A less strict notion of error is given by directly controlling the coverage via the **miscoverage loss**

$$\ell(Z, Y) = 1 - \frac{\sum_{ijk} Z_{ijk} Y_{ijk}}{\sum_{ijk} Y_{ijk}} \tag{15}$$

The miscoverage loss is directly related to the concept of accuracy [15, 27, 64] and can be easily extended to follow the balanced accuracy known in the medical literature [6, 7] or even a weighted version (*e.g.* lower importance to background pixels), inspired for instance by [11, 36]. [2]

4.4. Varisco heatmaps

In CP, the size a prediction set (*e.g.* a prediction interval) is taken as a signal of uncertainty: for a risk set by the user, it

[2] Example of weighted miscoverage loss:

$$\ell_w(Z, Y) = 1 - \frac{1}{\sum_k w_k} \sum_k w_k \frac{\sum_{ij} Z_{ijk} Y_{ijk}}{\sum_{ij} Y_{ijk}} \tag{16}$$

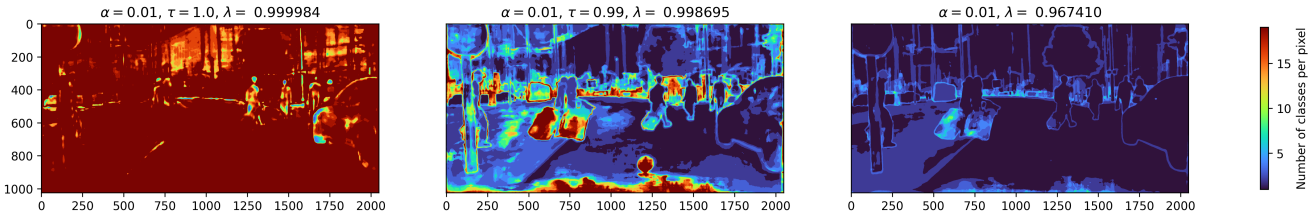

Figure 3. For the same risk level $\alpha = 0.01$, different losses yield different heatmaps: (left) binary loss ℓ_{bin}, (center) binary loss with threshold ℓ_τ, (right) miscoverage loss ℓ. If the notion of risk is too restrictive, the prediction set will be theoretically valid but not very informative. In this example, the figure on the left (binary loss, $\tau = 1.0$) has most of the pixels of color red, indicating that K (out of K) classes are in the prediction set. Dataset: *Cityscapes* [18].

corresponds to the "typical" error measured in the calibration dataset. In our case of image segmentation, we look at every pixel in the output: how many classes there are, whose softmax score is above the threshold $\hat{\lambda}$. If we count the labels in each pixel and normalize by K, we can generate an image which has, for each pixel, a scalar value in $[0, 1]$. Mapping these scalars to a gradient of colors we obtain a **heatmap** corresponding to the underlying prediction multi-labeled mask.

In Figure 2 are three examples: for the same predicted softmax, we apply thresholds $\lambda \in \{0.99, 0.999, 0.9999\}$ and obtain three different heatmaps. When λ is computed with a CRC procedure on calibration data, these heatmaps provide a qualitative visualization of the model's uncertainty obtained from the risk control procedure, hence the name *varisco* (visual assessment of risk control). Furthermore, for a better visualization in datasets with many classes (*e.g. LoveDA*, see Section 6), scaling the class count in every pixel by the maximum count observed in the multi-labeled mask (often $\ll K$) is also helpful.

The use of heatmaps is not new in semantic image segmentation, and one can find recent examples in [13, 14], where they are used for Out-of-Distribution (OOD) detection or in some of the UQ literature cited in Sections 3.1 and 3.2. To the best of our knowledge, however, this is the first time that this kind of visualization based on prediction sets is mentioned in the context of UQ and CP, with their underlying theoretical guarantee.

4.4.1 Characteristics of heatmaps

Our *varisco* heatmaps are monotone in the parameter λ: as λ grows, the set of pixels for each class is non-decreasing in size. Note that the heatmaps contain information about the aleatoric and epistemic uncertainty. As a general rule of thumb, for semantic segmentation tasks the aleatoric uncertainty should be maximal around the edges of the ground-truth figures, so that a heatmap with warm regions away of the contours should warn the user that the epistemic uncertainty of the model is high, and better models might be available for the data at hand.

The parameter λ encapsulates a notion of conservativeness, i.e. the higher the parameter λ, the more activation we will get in our multi-labeled mask, and therefore in our heatmap. The calibration of λ corresponds to the user setting an acceptable risk α and finding the least conservative λ such that their need is met. Note that for an arbitrary λ, the associated heatmap provides little information about the epistemic uncertainty of the model $\widehat{f}$, meaning that, given two different point predictors $\widehat{f}_1$ and $\widehat{f}_2$, plotting the heatmaps $H_1(\lambda)$ for the first model and $H_2(\lambda)$ will give us no information about which of the two models performs best. This is because the heatmaps $H_i(\lambda)$ carry no information about the errors of the models, but rather about the entropy of the softmax-es in each model. However, given a pre-set risk level α, once the appropriately values $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ are chosen through the CRC calibration procedure, we can safely compare the heatmaps $H_1(\widehat{\lambda}_1)$ and $H_2(\widehat{\lambda}_2)$, because both heatmaps guarantee the same risk level for both models. A warmer heatmap for the model $\widehat{f}_1$ means that for the chosen risk level α, the model $\widehat{f}_1$ carries more epistemic uncertainty than the model $\widehat{f}_2$.

5. Metrics and UQ diagnostics

To the best of our knowledge, ours is the first paper that uses prediction sets via multi-labeled masks to provably quantify the uncertainty in multiclass semantic segmentation; it is not possible to compare our results directly to existing approaches (see Section 3), as they are *essentially* different. However, as it is common in the CP literature, one can test different nonconformity scores or, as in our case, the coupling of set parametrization (*e.g.* $\mathcal{C}^{\text{LAC}}$) and conformalization loss. Throughout the paper we restrict our exposition to the nested-set parametrization derived from the algorithm of [42], which we refer to as LAC. Although out-of-scope for this paper, our code repository (see Sec. 1) contains some examples using the Adaptive Prediction Sets (APS) algorithm of [50], which employs a threshold on the sum of the softmax scores sorted in decreasing order.

To assess experimentally the validity of CRC, we compute the **empirical risk** as in Eq. (10) on the test data. In CP, a standard metric of the efficacy of the method is the average size of the prediction sets on the test data. For classification, this boils down to counting the number of classes in the prediction set, averaged over the test set. With semantic segmentation, this can be thought of as computing the average number of "activated" classes (*i.e.*, whose softmax is above the threshold $\hat{\lambda}$) over all pixels in the input image. For a one-hot-encoded multi-labeled masks Z defined as in Eq. (4), the *prediction set size* of a pixel (i, j) is $\sum_{k=1}^{K} z_{kij}$. Extending this to the whole image and normalizing by the number of valid pixels n_{pixels} (*e.g.* excluding void pixels, artefacts, etc. common in computer vision), we have, for one multi-labeled mask Z, the **activation ratio**

$$\text{AR}(Z) = \frac{1}{n_{\text{pixels}}} \sum_{i,j,k} z_{ijk}, \tag{17}$$

for all pixels x_{ij} that were labeled in X.

6. Experiments

Since CP is model-agnostic, we test architectures and datasets of varying size and complexity, knowing that regardless of the predictor chosen, our method will be statistically valid. For all the models tested, we used the PyTorch [47] implementation provided by the open-source Python library *MMSegmentation* [16], which includes code to run inferences as well as pretrained weights for many datasets.

We run our experiments on *Cityscapes* [18] (19 classes, automotive vision), *ADE20K* [66, 67] (150 generic common classes) and *LoveDA* [58, 59] (aerial images, 7 classes). As for the architectures of the neural networks, we selected the best performing ones within our computational budget: PSPNet [65] for *Cityscapes* and *LoveDA*, SegFormer for *ADE20K* [62].

For conformalization, we split the validation data into two partitions, one for calibration and one for testing. We tested both the binary loss with threshold of Eq. (14) and the miscoverage loss of Eq. (15). For the risk α, we tested values that would have made sense in the real world: for *Cityscapes*, our pretrained model has a very good performance (*e.g.* mIoU) and the user can aim for small risks. For the other cases, such small α could yield hardly informative prediction sets. As seen for instance in Figure 3, a combination of a restrictive loss ($\tau = 1.0$) and a small $\alpha = 0.01$ would entail selecting (almost) all classes for every pixel. This can be taken as a diagnostic signal: the model is not good enough for our notion of risk, we need to either augment our tolerance for errors or revise the prediction model, for instance.

Dataset	α	τ	**Empirical Risk**	**AR**
Cityscapes	0.1	0.99	0.106 ± (0.019)	1.028
	0.1	0.95	0.100 ± (0.021)	1.274
	0.01	0.95	0.011 ± (0.008)	1.208
	0.01	0.99	0.011 ± (0.014)	2.557
ADE20K	0.1	0.75	0.082 (± 0.021)	1.440
	0.1	0.90	0.076 (± 0.021)	3.483
	0.01	0.75	0.004 (± 0.005)	9.349
LoveDA	0.10	0.50	0.097 (± 0.018)	1.231
	0.10	0.75	0.103 (± 0.013)	2.672
	0.10	0.90	0.092 (± 0.012)	3.946
	0.01	0.50	0.010 (± 0.008)	3.607
	0.01	0.75	0.010 (± 0.005)	4.956
	0.01	0.90	0.010 (± 0.006)	5.761

Table 1. Metrics on $\mathcal{D}_{\text{test}}$: empirical risk and activation ratio (AR) for **binary loss** ℓ_τ. Empirical risk should be as close as possible to α to show validity. For each line, we repeat several times this procedure: (1) shuffle the dataset, (2) split validation data into $\mathcal{D}_{\text{cal}}$ & $\mathcal{D}_{\text{test}}$, (3) run calibration on $\mathcal{D}_{\text{cal}}$, (4) Compute metrics on $\mathcal{D}_{\text{test}}$. We finally average the metrics over the multiple runs (standard deviation in the parentheses).

Dataset	α	**Empirical Risk**	**AR**
Cityscapes	0.05	0.041 (± 0.001)	1.000†
	0.01	0.006 (± 0.001)	1.230
	0.005	0.001 (± 0.0003)	1.998
ADE20K	0.2	0.179 (± 0.005)	1.000†
	0.1	0.098 (± 0.008)	1.362
	0.05	0.048 (± 0.006)	2.474
	0.01	0.008 (± 0.002)	15.285
LoveDA	0.2	0.199 ± (0.009)	1.388
	0.1	0.100 ± (0.006)	2.650
	0.05	0.049 ± (0.004)	4.069
	0.01	0.008 ± (0.0008)	6.350
	0.005	0.003 ± (0.0004)	6.796

Table 2. Metrics on $\mathcal{D}_{\text{test}}$: empirical risk and activation ratio (AR) for **miscoverage loss** ℓ. †: the underlying predictor attains the risk level without adding any class to the prediction set, that is, the output semantic mask with one class per pixel already satisfies this risk level.

7. Results

In Table 1 and Table 2, we report the results of our experiments. For both the empirical risk and the activation ratio (AR), we average the metrics over multiple runs (ten) of each loss configuration.

As it is customary in CP, one first ensures that the theoretical guarantees holds also in practice for a given dataset

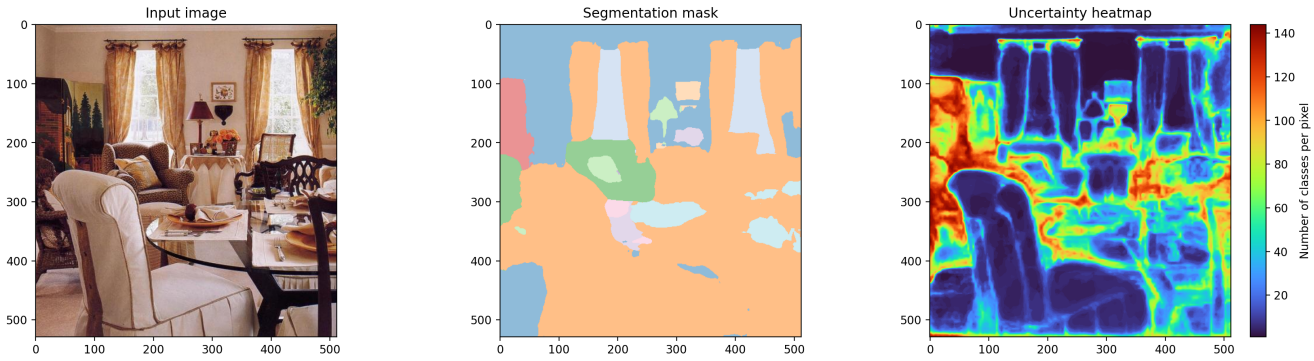

Figure 4. Visualization of a *varisco* heatmaps (miscoverage loss, $\alpha = 0.01$) for the *ADE20K* dataset [66, 67]: (left) input image, (center) predicted segmentation mask, (right) *varisco* heatmap.

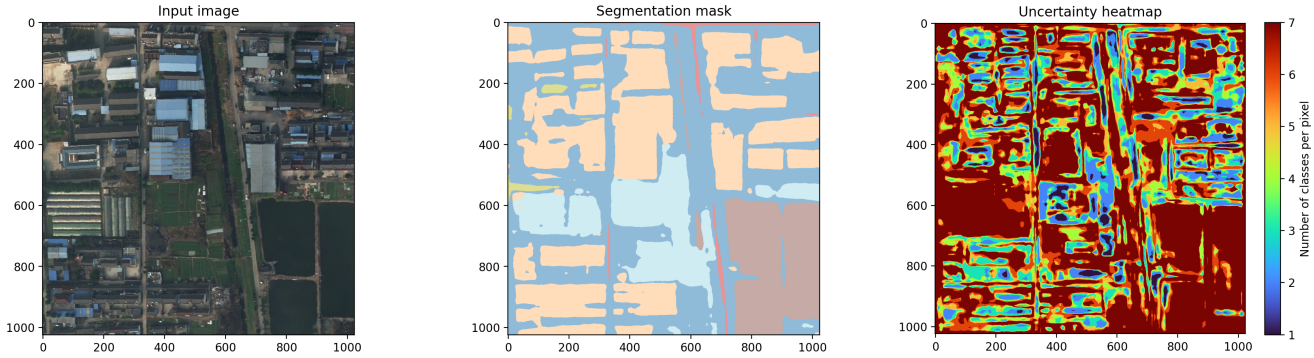

Figure 5. Visualization of a *varisco* heatmaps (miscoverage loss, $\alpha = 0.01$) for the *LoveDA* dataset [58, 59]: (left) input image, (center) predicted segmentation mask, (right) *varisco* heatmap.

and predictive model. As expected, the empirical risks are a very close approximation of the nominal risk values α. As for the activation ratios, they follow a reasonable pattern: they increase as the notion of error gets stricter. In Tab. 2, we see how the AR increases when the risk diminishes from $\alpha = 0.05$ to $\alpha = 0.01$.

In Fig. 1 and Fig. 5 we give examples of conformalized predictions for the three datasets. In all cases, the borders of the masks often appear to be highlighted as more uncertain. For the other patches in darker shades of red, the signal of uncertainty produced by the conformalized predictors go towards more ambiguous areas, smaller and farther objects.

8. Conclusion

Thanks to its light computational footprint, CP can be applied wherever ML is deployed in sensitive or critical applications, with limited knowledge of the underlying model (e.g. black box). We have shown how to extend the theoretical guarantees of CP to any predictor for semantic image segmentation, using a post-hoc a procedure that only requires access to softmax scores.

In the future, it could be interesting to study the interactions of our method with existing UQ predictors (*e.g.* bayesian). Another promising direction would also be to extend this work to panoptic segmentation, which would allow to extend the theoretical guarantees to instances in the input images. From the point of view of safety in AI, one could also profit from experimenting with class-conditional conformal guarantees: one can restrict CRC to a subset of priority classes, such as pedestrians or bicycle riders in automonous driving, ignoring the others. Finally, a major methodological challenge would be to work towards providing theoretical guarantees to sequences of data, notably for real-time video.

Acknowledgments

This work has benefited from the support of the DEEL project,[3] with fundings from the Agence Nationale de la Recherche, and which is part of the ANITI AI cluster.

[3] https://www.deel.ai/

References

[1] Lucian Alecu, Hugues Bonnin, Thomas Fel, Laurent Gardes, Sébastien Gerchinovitz, Ludovic Ponsolle, Franck Mamalet, Éric Jenn, Vincent Mussot, Cyril Cappi, Kevin Delmas, and Baptiste Lefevre. Can we reconcile safety objectives with machine learning performances? In *ERTS*, 2022. 1

[2] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. 2, 3

[3] Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021. 3

[4] Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *Proceedings of the 39th International Conference on Machine Learning*, pages 717–730. PMLR, 2022. 3

[5] Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 4

[6] Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Thomas Geiser, Andreas Christe, and Stavroula Mougiakakou. Semantic segmentation of pathological lung tissue with dilated fully convolutional networks. *IEEE journal of biomedical and health informatics*, 23(2):714–722, 2018. 5

[7] Mohamed Attia, Mohammed Hossny, Saeid Nahavandi, and Hamed Asadi. Surgical tool segmentation using a hybrid deep cnn-rnn auto encoder-decoder. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3373–3378. IEEE, 2017. 5

[8] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *J. ACM*, 68(6), 2021. 3

[9] Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, and Michael Elad. Volume-oriented uncertainty for inverse problems. In *NeurIPS 2023 Workshop on Deep Learning and Inverse Problems*, 2023. 3

[10] Ondrej Bohdal, Da Li, and Timothy Hospedales. Label calibration for semantic segmentation under domain shift. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023. 3

[11] Gülcan Can, Dario Mantegazza, Gabriele Abbate, Sébastien Chappuis, and Alessandro Giusti. Semantic segmentation on swiss3dcities: A benchmark study on aerial photogrammetric 3d pointcloud dataset. *Pattern Recognition Letters*, 150: 108–114, 2021. 5

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[13] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 6

[14] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5128–5137, 2021. 6

[15] Pei Chen, Yangkang Zhang, Zejian Li, and Lingyun Sun. Few-shot incremental learning for label-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3697–3707, 2022. 5

[16] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 7

[17] Charles Corbière, Nicolas THOME, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3, 6, 7

[19] Sebastian Cygert, Bartłomiej Wróblewski, Karol Woźniak, Radosław Słowiński, and Andrzej Czyżewski. Closer look at the uncertainty estimation in semantic segmentation under distributional shift. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021. 3

[20] Steffen Czolbe, Kasra Arnavaz, Oswin Krause, and Aasa Feragen. Is segmentation uncertainty useful? In *Information Processing in Medical Imaging*, pages 715–726, Cham, 2021. Springer International Publishing. 3

[21] European Union Aviation Safety Agency EASA. Artificial intelligence roadmap 2.0. Technical report, European Union Aviation Safety Agency, 2023. 1

[22] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 1, 3

[23] Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Representations*, 2022. 3

[24] Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. In *Advances in Neural Information Processing Systems*, pages 3711–3723. Curran Associates, Inc., 2020. 1, 3

[25] Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. In *Advances in Neural Infor-*

mation Processing Systems, pages 3711–3723. Curran Associates, Inc., 2020. 3

[26] Chirag Gupta, Arun K. Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022. 5

[27] Hexiang Hu, Zhiwei Deng, Guang-Tong Zhou, Fei Sha, and Greg Mori. Recalling holistic information for semantic segmentation. *arXiv preprint arXiv:1611.08061*, 2016. 5

[28] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021. 2

[29] Eric Jenn, Alexandre Albore, Franck Mamalet, Grégory Flandin, Christophe Gabreau, Hervé Delseny, Adrien Gauffriau, Hugues Bonnin, Lucian Alecu, Jérémy Pirard, et al. Identifying challenges to the certification of machine learning for safety critical systems. In *European congress on embedded real time systems (ERTS 2020)*, 2020. 1

[30] Thierry Judge, Olivier Bernard, Mihaela Porumb, Agisilaos Chartsias, Arian Beqiri, and Pierre-Marc Jodoin. Crisp-reliable uncertainty estimation for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 492–502. Springer, 2022. 3

[31] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 3

[32] A Kendall, V Badrinarayanan, and R Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *British Machine Vision Conference 2017, BMVC 2017*. British Machine Vision Association, 2017. 3

[33] Michael C Krygier, Tyler LaBonte, Carianne Martinez, Chance Norris, Krish Sharma, Lincoln N Collins, Partha P Mukherjee, and Scott A Roberts. Quantifying the unknown impact of segmentation uncertainty on image-based simulations. *Nature communications*, 12(1):5414, 2021. 3

[34] Gilad Kutiel, Regev Cohen, Michael Elad, and Daniel Freedman. What's behind the mask: Estimating uncertainty in image-to-image problems, 2022. 3

[35] Preethi Lahoti, Krishna Gummadi, and Gerhard Weikum. Responsible model deployment via model-agnostic uncertainty learning. *Machine Learning*, 112(3):939–970, 2022. 2

[36] Yating Ling, Yuling Wang, Wenli Dai, Jie Yu, Ping Liang, and Dexing Kong. Mtanet: Multi-task attention network for automatic medical image segmentation and classification. *IEEE Transactions on Medical Imaging*, 2023. 5

[37] David Fernández Llorca and Emilia Gómez. *Trustworthy Autonomous Vehicles: Assessment Criteria for Trustworthy AI in the Autonomous Driving Domain*. Publications Office of the European Union, 2021. 1

[38] João Lourenço-Silva and Arlindo L. Oliveira. Using soft labels to model uncertainty in medical image segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 585–596, Cham, 2022. Springer International Publishing. 3

[39] Franck Mamalet, Eric Jenn, Gregory Flandin, Hervé Delseny, Christophe Gabreau, Adrien Gauffriau, Bernard Beaudouin, Ludovic Ponsolle, Lucian Alecu, Hugues Bonnin, Brice Beltran, Didier Duchel, Jean-Brice Ginestet, Alexandre Hervieu, Sylvain Pasquet, Kevin Delmas, Claire Pagetti, Jean-Marc Gabriel, Camille Chapdelaine, Sylvaine Picard, Mathieu Damour, Cyril Cappi, Laurent Gardès, Florence De Grancey, Baptiste Lefevre, Sébastien Gerchinovitz, and Alexandre Albore. White Paper Machine Learning in Certified Systems. Research report, IRT Saint Exupéry ; ANITI, 2021. 1

[40] Juliette Mattioli, Henri Sohier, Agnès Delaborde, Kahina Amokrane, Afef Awadid, Zakaria Chihani, Souhaiel Khalfaoui, and Gabriel Pedroza. An overview of key trustworthiness attributes and kpis for trusted ml-based systems engineering. In *Workshop AITA AI Trustworthiness Assessment-AAAI Spring Symposium*, 2023. 2

[41] Juliette Mattioli, Henri Sohier, Agnès Delaborde, Gabriel Pedroza, Kahina Amokrane-Ferka, Afef Awadid, Zakaria Chihani, and Souhaiel Khalfaoui. Towards a holistic approach for ai trustworthiness assessment based upon aids for multi-criteria aggregation. In *SafeAI 2023-The AAAI's Workshop on Artificial Intelligence Safety*, 2023. 2

[42] Jing Lei Mauricio Sadinle and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234, 2019. 2, 4, 6

[43] Yujian Mo, Yan Wu, Xinneng Yang, Feilin Liu, and Yujun Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493: 626–646, 2022. 1

[44] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H.S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24384–24394, 2023. 3

[45] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, 2002. 1, 2

[46] Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. Pac confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*, 2020. 3

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 7

[48] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 3

[49] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object de-

tection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[50] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, pages 3581–3591. Curran Associates, Inc., 2020. 6

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 1

[52] Axel-Jan Rousseau, Thijs Becker, Jeroen Bertels, Matthew B. Blaschko, and Dirk Valkenborg. Post training uncertainty calibration of deep networks for medical image segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1052–1056, 2021. 3

[53] Prabhu Teja S and Francois Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9613–9623, 2021. 3

[54] Jacopo Teneggi, Matthew Tivnan, Web Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *Proceedings of the 40th International Conference on Machine Learning*, pages 33940–33960. PMLR, 2023. 3

[55] Jiaye Teng, Chuan Wen, Dinghuai Zhang, Yoshua Bengio, Yang Gao, and Yang Yuan. Predictive inference with feature conformal prediction. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[56] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005. 1, 2, 3

[57] Dongdong Wang, Boqing Gong, and Liqiang Wang. On calibrating semantic segmentation models: Analyses and an algorithm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23652–23662, 2023. 3

[58] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran Associates, Inc., 2021. 7, 8

[59] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. https://doi.org/10.5281/zenodo.5706578, 2021. 7, 8

[60] Håkan Wieslander, Philip J. Harrison, Gabriel Skogberg, Sonya Jackson, Markus Fridén, Johan Karlsson, Ola Spjuth, and Carolina Wählby. Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images. *IEEE Journal of Biomedical and Health Informatics*, 25(2):371–380, 2021. 3

[61] Huajun Xi, Jianguo Huang, Lei Feng, and Hongxin Wei. Does confidence calibration help conformal prediction?, 2024. 3

[62] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, pages 12077–12090. Curran Associates, Inc., 2021. 7

[63] Hongzheng Yang, Cheng Chen, Yueyao CHEN, , Hon Chi Yip, and DOU QI. Uncertainty estimation for safety-critical scene segmentation via fine-grained reward maximization. In *Advances in Neural Information Processing Systems*, pages 36238–36249. Curran Associates, Inc., 2023. 3

[64] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2736–2746, 2022. 5

[65] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 7

[66] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 7, 8

[67] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 7, 8