Medical Image Segmentation via Morphological Conformal Prediction and Fractional Iterations*

Bruce Cyusa Mukama[1 [0009−0001−2251−3937]], Soundouss Messoudi[1 [0009−0008−8750−3829]], Sylvain Rousseau[1 [0000−0002−2212−5950]], and Sébastien Destercke[1 [0000−0003−2026−468X]]

Université de technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France
{bruce.cyusa-mukama or firstname.lastname}@hds.utc.fr
https://www.hds.utc.fr/

Abstract. To provide statistical coverage guarantees when the predicted probabilities of an image segmentation model are not accessible, Morphological Conformal Image Segmentation (MCIS) can infer the number of consecutive morphological dilation(s) that shall stretch the predicted binary segmentation mask and ensure that the obtained (conformalized) binary mask shall cover the ground truth segmentation mask, with a confidence level $1 - \alpha$ that is specified by the user. Therefore, to post-process any predicted binary mask, the number of consecutive dilation(s) that are applied by MCIS is always a positive integer: the whole mask is post-processed. However, to reduce the area of the conformalized mask, it can be interesting to only dilate some parts of the predicted mask. For this purpose, we sort the pixels in the predicted mask according to their distances w.r.t. the borders of the regions that are formed by the foreground pixels. Afterwards, for the last dilation, we only modify a portion of the sorted pixels instead of the whole mask. On multiple medical image segmentation datasets, the experiments show that the empirical coverage gap of MCIS and the areas of the conformalized masks are reduced by our novel approach, on average.

Keywords: Semantic Image Segmentation · Uncertainty Quantification · Normalized · Conformal Prediction · Morphological Filtering.

1 Introduction

To thoroughly analyze an image, a semantic class can be assigned to each pixel in the image to identify the category of the structure that is represent by the pixel(s). Among other purposes, this semantic image segmentation process is applied in safety-critical applications such as medical imaging [34] and autonomous driving [23]. However, many (off-the-shelf) segmentation algorithms do not come with statistical guarantees. Therefore, their deployment can necessitate rigorous Uncertainty Quantification (UQ) when major risks are involved.

* Supported by the UTC Foundation, the Safe AI Chair, and Hauts-de-France.

To segment an image, the (trained) model can output a continuous score $\widetilde{Y}_i(u, v) \in (0, 1)$ which can be interpreted as the predicted confidence that the pixel (u, v) in the image X_i represents a certain structure or object, e.g., a blood cell. However, in some cases, these continuous outputs can be unavailable. Instead, the prediction can be a binary segmentation mask, where the values $\widehat{Y}_i(u, v) \in \{0, 1\}$ indicate a non-probabilistic membership. For example, this can be the case if the segmentation model is encapsulated into an application programming interface (a remote service) that prevents any distillation of the underlying segmentation model. This can also happen if the segmentation model has already been distilled into a digital circuit (for hardware acceleration purposes) or if the segmentation model was not designed to readily output continuous scores, e.g., some of the image segmentation methods that are based on level-set methods [15, 31].

In the absence of the continuous outputs of the segmentation model, morphological operations [5] can be combined with Conformal Prediction (CP) [33] to post-process the predicted binary segmentation masks. In Morphological Conformal Image Segmentation (MCIS) [21], the goal is to provide statistical coverage guarantees. In this article we shall only speak of full coverage and it is achieved when there are no false negative pixels in a segmentation mask. This is formalized in Section 3. For this purpose, CP is leveraged to infer how many times a predicted binary mask ought to be dilated to ensure that the conformalized mask (the mask that is obtained) shall cover the ground truth mask with a confidence level of at least $1 - \alpha$, as specified the user.

As a consequence, the number of consecutive dilation(s) that are applied by MCIS is always a whole number: all the parts of the mask are post-processed. Regardless, to reduce the areas of the (obtained) conformalized masks, it can be interesting to only dilate some parts of the predicted mask. With that end in view, we sort the pixels in the predicted mask according to their distances w.r.t. the borders of the regions that are formed by the foreground pixels. Afterwards, for the last iteration of the consecutive dilation(s), we only modify a portion of the sorted pixels instead of the whole mask. This pixel-sorting criteria is motivated by the observation [13, 35] that many segmentation neural networks tend to perform worse on the boundaries of the structure(s) or object(s) that are represented by the image than in their interiors.

To share our findings, we have organized this article as follows. In Section 2, we present the preexisting research works that are related to our contributions. In Section 3, we formally introduce MCIS and detail how we modified it. In Section 4, we test the proposed method. Finally, in Section 5, we outline this study and recommend future research directions.

2 Related works

UQ in Image Segmentation. For high performance, deep learning is often used in semantic image segmentation [19, 23]. To quantify the uncertainty in the outputs of such segmentation algorithms, various UQ frameworks [14, 12] can been ap-

plied. For instance, to calibrate the continuous outputs of segmentation models, [30] applied some of the traditional post-hoc probability calibration methods, i.e., Platt scaling and Beta calibration. To directly output accurate uncertainty estimates, [28] transformed preexisting architectures into Bayesian neural networks and [18] applied ensembles of models to detect out-of-distribution test examples. Furthermore, [37] leveraged subjective logic theory and evidential learning to explicitly model uncertainty in medical image segmentation. To aggregate multiple multi-class classifiers into a more reliable credal classifier, [26] proposed a framework which remains to be generalized to image segmentation, to the best of our knowledge.

CP in Image Segmentation. Conformal Prediction (CP) is a straightforward UQ framework [33] that is often preferred due to its practical effectiveness and its theoretical soundness. CP considers a finite-size calibration sample, a black box predictor, and an unspecified (i.e., unknown) distribution of the data. More specifically, the statistical exchangeability of the data is sufficient in CP [3]. CP provides non-asymptotic statistical guarantees in semantic image segmentation [20, 7] and in instance segmentation [9, 16]. More importantly, to consider uses cases where the (only) accessible prediction is a binary mask, CP has been combined with morphological filters [5] through MCIS [21, 22]. MCIS is a major alternative to most CP methods [1, 20, 9, 4] because they assume the availability of predicted (pixel-wise) probabilities. Nonetheless, to achieve (full) coverage, the conformalized masks that are produced by MCIS can be larger than necessary, i.e., overconservative and less informative, due to the fact that MCIS dilates the whole (predicted) binary mask. To the best of our knowledge, MCIS has not been modified in order to only dilate some portions of the predicted binary mask. Hence, we investigate this approach to reduce the areas of the conformalized masks, in this article. That being said, we have also investigated a different approach to the same end in [25]. In [25], we have tried to reduce the area of the conformalized segmentation masks by changing the shape of the structuring element from one pixel to another, to dilate each predicted mask according the continuous outputs of the segmentation model (which are assumed to be accessible in [25]). Hence, the two articles target different use cases and propose two solutions that are completely different.

3 Methods

Since the proposed method is based on MCIS [21], let's begin by presenting the preexisting MCIS approach in Section 3.1 and then describe the novel approach in Section 3.2.

3.1 Morphological conformal image segmentation

For convenience purposes, let's represent the image X_i as a function $X_i : \Omega \to \mathbb{R}^3$ that assigns a color code to a pixel (u, v) in a grid $\Omega \subset \mathbb{N}^2$. Let $h_i \times w_i$ be the

Algorithm 1 MCIS with whole non-conformity scores

Require: a finite size dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=-k}^{n}$ and a segmentation model f_θ.

1: split the dataset into $\mathcal{D}$ two disjoint subsets $\mathcal{D}_{\text{Tra}}$ and $\mathcal{D}_{\text{Cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$,
2: train the segmentation algorithm f_θ on the proper training set $\mathcal{D}_{\text{Tra}}$,
3: **for** $i \in \{1, \ldots, n\}$ **do** `/* Calibration */`
4: apply the pretrained algorithm f_θ on X_i to predict the binary mask $\widehat{Y}_i$,
5: evaluate the non-conformity score:

$$\lambda_i = \inf\{\lambda \in \mathbb{N} : Y_i \subseteq (\widehat{Y}_i \oplus \delta)^{\lambda}\} \tag{1}$$

6: **end for**
7: under the assumption that the scores $\{\lambda_i\}_{i=1}^{n} \cup \lambda_{n+1}$ are statistically exchangeable

$$\lambda_\alpha \text{ is the } (\lceil (1-\alpha)(n+1) \rceil / n) - \text{quantile of the scores } \{\lambda_i\}_{i=1}^{n}, \tag{2}$$

8: At test-time (when $i > n$) `/* Deployment */`

$$Y_i^{\alpha} = (\widehat{Y}_i \oplus \delta)^{\lambda_\alpha}, \tag{3}$$

number of pixels in the image X_i. Furthermore, let's define another function that identifies the pixels in the image X_i representing a particular category of structure, e.g., an organ in the human body. This latter function is the ground truth segmentation mask $Y_i : \Omega \to \{0, 1\}$, where $Y_i(u, v)$ is 1 if the pixel (u, v) represents the particular category of structure(s), the foreground, otherwise 0.

By dilating a predicted segmentation mask $\widehat{Y}_i : \Omega \to \{0, 1\}$ morphologically, MCIS provides the following statistical guarantees [21]:

$$P(Y_{n+1} \subseteq Y_{n+1}^{\alpha} \mid \mathcal{D}_{\text{Cal}}) \geq 1 - \alpha, \tag{4}$$

where $1 - \alpha$ denotes the user-specified confidence level, $\mathcal{D}_{\text{Cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$ denotes the calibration dataset, and $Y_i \subseteq Y_i^{\alpha}$ means that the conformalized mask Y_i^{α} covers the ground truth segmentation mask Y_i:

$$Y_i \subseteq Y_i^{\alpha} \iff \forall (u, v) \in \Omega,\ Y_i(u, v) = 1 \implies Y_i^{\alpha}(u, v) = 1. \tag{5}$$

In other words, coverage[1] is achieved when the dilated mask no longer contains any false negative pixel. Note that the dilated mask can still contain false positive pixels. Since the calibration set $\mathcal{D}_{\text{Cal}}$ in Equation (5) does not contain the test example (X_{n+1}, Y_{n+1}), the goal of MCIS is to infer a conformalized mask Y_{n+1}^{α} containing (with a minimal probability $1 - \alpha$) all the foreground pixels in the unknown ground truth mask Y_{n+1}. This requires a pretrained segmentation model, the calibration dataset $\mathcal{D}_{\text{Cal}}$, and the unlabeled image X_{n+1}.

The conformalized binary mask Y_i^{α} is obtained by applying $\lambda_\alpha \in \mathbb{N}$ consecutive morphological dilation(s) to the predicted mask $\widehat{Y}_i$:

$$Y_i^{\alpha} = (\widehat{Y}_i \oplus \delta)^{\lambda_\alpha} = (\widehat{Y}_i \oplus \delta) \oplus \delta \oplus \ldots \oplus \delta, \tag{6}$$

[1] Again, we shall only speak of full coverage in this article.

where δ is a symetrical structuring element. In this article, we shall only use the 3×3 cross-shaped structuring element  . Algorithm 1 describes how the sufficient number of consecutive dilation(s) λ_α is determined. We refer the readers to [21] for the formal proofs.

3.2 Fractional iterations

Having described the preexisting MCIS approach in the previous section, let's now detail our contributions. The use of whole numbers as non-conformity scores $\lambda_i \in \mathbb{N}$ induces multiple adverse effects. First, as shown by Equation (4), the guarantees of MCIS are defined by an inequality. In fact, the upper bound of the inequality is

$$1 - \alpha + \frac{1}{n' + 1} \geq P(Y_{n+1} \subseteq Y_{n+1}^{\alpha} \mid \mathcal{D}_{\text{Cal}}), \tag{7}$$

where n' is the number of distinct values in the sequence of scores $(\lambda_1, \lambda_2, \ldots, \lambda_n)$. When the non-conformity scores $\{\lambda_i\}_{i=1}^{n}$ are natural numbers, their sequence is likely to contain some duplicates because discrete values vary less than continuous values, naturally. These duplicate values artificially shrink the size of the calibration dataset and thereby increase the gap between the empirical coverage rate and the user-specified confidence level $1 - \alpha$. Here, the areas of the conformalized masks that are obtained with λ_α dilation(s) can be larger than necessary, i.e., overconservative w.r.t. the user-specified confidence level $1 - \alpha$.

Second, the non-conformity scores in Equation (1) are not compatible with normalized CP. In normalized CP [27], dividing the non-conformity scores by a quantity $\sigma_i \in \mathbb{R}$ can suffice to obtain prediction sets with adaptive sizes. For instance, in conformal object detection, [24] normalized the non-conformity scores by the distance between the ego-vehicle and the obstacles and [32] divided them by the sizes of the predicted bounding boxes. However, given the definition of the number of necessary dilation(s) λ_α in Equation (2), a fractional number of iterations $\lambda_i / \sigma_i \in \mathbb{R}$ makes no senses in MCIS, by default.

The solution that we propose for the aforementioned issues was motivated by the following observation. If the pretrained segmentation model is excellent, then the number of false positive pixels in predicted mask $\widehat{Y}_i$ can be very small. In such cases, one iteration of morphological dilation might be enough to achieve coverage. However, a single dilation could also add a non-negligible number of inaccurate pixels to the foreground. In such cases, the conformalized segmentation mask Y_i^{α} would have been enlarged more than necessary.

Under the assumption that the false negative pixels are closer to the boundary of the predicted binary mask than most of the background pixels, we can sort the pixels by using their distance to the closest predicted foreground pixel(s). Afterwards, by starting from the closest pixel, we can modify a portion $\tau_i \in [0, 1]$ in the conformalized mask Y_i^{α} according to the results of the dilation $(\widehat{Y}_i \oplus \delta)$[1]. Here, as we assume that $\lambda_\alpha = 1$, $\tau_i = 1$ means that we update 100% of the pixels in the conformalized mask Y_i^{α}. The portion of the pixels that are skipped during the update is $1 - \tau_i$. This provides an interpretable definition of fractional

Algorithm 2 MCIS with fractional non-conformity scores

Require: a finite size dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=-k}^{n}$, the normalization scores σ_i, and a segmentation model f_θ.

1: split the dataset into $\mathcal{D}$ two disjoint subsets $\mathcal{D}_{\text{Tra}}$ and $\mathcal{D}_{\text{Cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$,
2: train the segmentation algorithm f_θ on the proper training set $\mathcal{D}_{\text{Tra}}$,
3: **for** $i \in \{1, \ldots, n\}$ **do** `/* Calibration */`
4: apply the pretrained algorithm f_θ on X_i to predict the binary mask $\widehat{Y}_i$,
5: initialize a counter $\tau_i' \leftarrow 0$ and a segmentation mask $\widehat{Y}_i' \leftarrow \widehat{Y}_i$,
6: **while** $Y_i \nsubseteq (\widehat{Y}_i' \oplus \delta)$ **do**
7: $\widehat{Y}_i' \leftarrow (\widehat{Y}_i' \oplus \delta)$,
8: $\tau_i' \leftarrow (\tau_i' + h_i \times w_i)$,
9: **end while**
10: increasingly sort the pixels in $\widehat{Y}_i$ by their distance to the closest foreground pixel(s),
11: $\tau_i'' \leftarrow$ the rank of the false negative pixel in $\widehat{Y}_i'$ that is the furthest to the foreground,
12: evaluate the non-conformity score:

$$\tau_i = \frac{(\tau_i' + \tau_i'')}{\sigma_i}, \tag{8}$$

13: **end for**
14: under the assumption that the scores $\{\tau_i\}_{i=1}^{n} \cup \tau_{n+1}$ are statistically exchangeable

$$\tau_\alpha \text{ is the } (\lceil (1-\alpha)(n+1) \rceil / n) - \text{quantile of the scores } \{\tau_i\}_{i=1}^{n}, \tag{9}$$

15: **if** $i > n$ (at test-time) **then** `/* Deployment */`
16: initialize the counter $\tau_i' \leftarrow 0$, the mask $\widehat{Y}_i' \leftarrow \widehat{Y}_i$, and the mask $Y_i^\alpha \leftarrow \widehat{Y}_i$,
17: **while** $\tau_i' < \lceil \tau_\alpha \times \sigma_i \rceil$ **do**
18: $\widehat{Y}_i' \leftarrow (\widehat{Y}_i' \oplus \delta)$,
19: **if** $\lceil \tau_\alpha \times \sigma_i \rceil - \tau_i' \geq h_i \times w_i$ **then**
20: $Y_i^\alpha \leftarrow \widehat{Y}_i'$,
21: $\tau_i' \leftarrow (\tau_i' + h_i \times w_i)$,
22: **else**
23: $\tau_i'' \leftarrow \lceil \tau_\alpha \times \sigma_i \rceil - \tau_i'$,
24: increasingly sort the pixels in $\widehat{Y}_i'$ by their distance to the foreground pixel(s),
25: use the values of the pixels in $\widehat{Y}_i'$ to update the corresponding pixels in Y_i^α, but skip all the pixels with a rank greater than τ_i'' in the sorted sequence (obtained in Step 24),
26: **break**
27: **end if**
28: **end while**
29: **end if**

non-conformity scores for MCIS and Algorithm 2 generalizes it to cases where coverage can only be obtained after more than one morphological dilation. In Algorithm 2, to determine the distances in Step 10 and Step 24, the Euclidean Distance Transform (EDT) [29, 6] is applied to the (pixel-wise) logical inverse of

the binary mask $\widehat{Y}_i$, i.e., the (binary) complement of the mask $\widehat{Y}_i$. Unfortunately, our definition limits the idea of prioritized updates to the last iteration of the consecutive dilation(s). Figure 1 shows examples of the EDT.

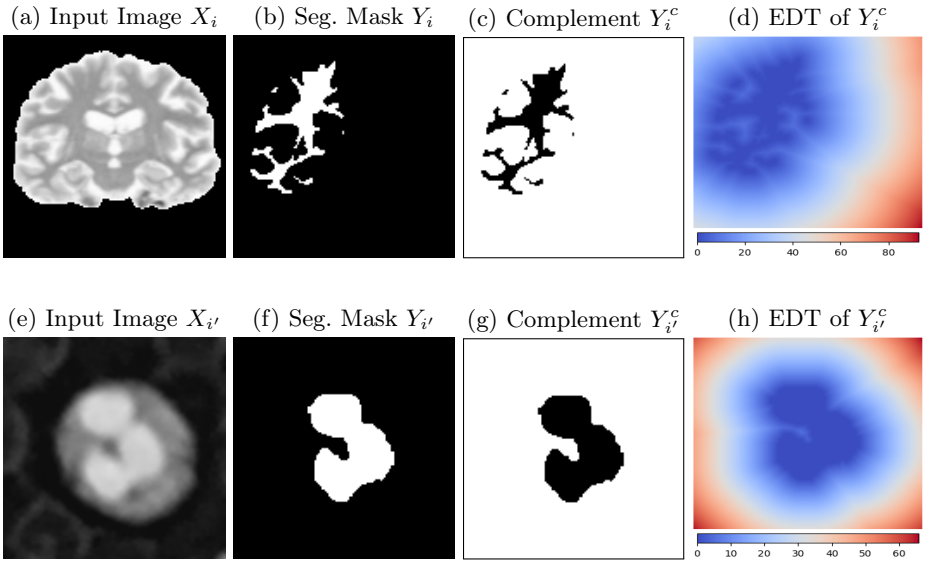

Fig. 1: Examples of the euclidean distance transform (EDT), as computed from the complement or logical inverse Y_i^c of a segmentation mask Y_i. On the top row, the segmented structure is the cerebral left white matter [11, 17]. On the bottom row, the segmented structure is a the nucleus of a white blood cell [36].

4 Experiments

4.1 Implementation details

Datasets. We evaluate the proposed method on two publicly available datasets that are composed of medical images: OASIS Brain [11, 17] and WBC [36]. The OASIS dataset contains cerebral MRI scans of diverse patients and the WBC dataset contains microscopic images of blood cells. We only segment white blood cells (nuclei) on WBC and cerebral left white matter(s) on OASIS. On both datasets, we use the official source code[2] of [21] to prepare the data and to apply a (few-shot) segmentation model: UniverSeg [8][3]. On both datasets, we use 40% of the images in the official validation set(s) for calibration and the

[2] https://github.com/deel-ai-papers/consema.git
[3] https://github.com/JJGO/UniverSeg.git

other 60% for testing. We only use the finely annotated images in the official training set(s) to (pre)train the (deep) segmentation models that are (post-hoc) calibrated afterwards.

Baselines. As in [21], we use the UniverSeg [8] few-shot segmentation model to predict the binary segmentation masks. The applied segmentation threshold is 0.5. We compare 3 post-hoc calibration methods (the wrappers): "MCIS", "Ours^{1}", and "Ours^{2}". We designate by "MCIS" the preexisting method [21] and by "Ours^{1}" our method in Algorithm 2 without any normalization score: here, we set $\sigma_i = 1$, for each prediction. Furthermore, we use the area, i.e., number of foreground pixels in the predicted mask as the normalization scores $\sigma_i = |\widehat{Y}_i|$, in Algorithm 2, and we designate this baseline by "Ours^{2}". The baseline "Ours^{2}" is inspired by [32] because the authors normalized the non-conformity scores by the sizes of the predicted bounding boxes, in conformal object detection.

Evaluation metrics. On the test sets, we evaluate the empirical coverage rate (and gap) of the conformalized masks Y^{α}_{n+1} and the image-level mean intersection over union (mIoU) as follows:

$$\text{mIoU} = \frac{1}{m}\sum_{j=1}^{m}\frac{|Y_j \cap Y_j^{\alpha}|}{|Y_j \cup Y_j^{\alpha}|}, \tag{10}$$

$$\text{Coverage rate} = \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\{Y_j \subseteq Y_j^{\alpha}\}, \tag{11}$$

$$\text{Coverage gap} = |(1-\alpha) - \text{Coverage rate}|, \tag{12}$$

where $\mathbb{1}\{C\}$ is 1 if the condition C is verified, otherwise 0, and $j \in \{1, \ldots, m\}$ is the index of a test example. $|Y_j \cap Y_j^{\alpha}|$ is the number of foreground pixels that are common to the ground truth mask Y_j and the conformalized mask Y_j^{α}. Similarly, $|Y_j \cup Y_j^{\alpha}|$ is the number of pixels that are non-zero in Y_j or in Y_j^{α}. Hence, when $Y_j \subseteq Y_j^{\alpha}$, the ratio in Equation (10) is the area of the ground truth mask Y_i over the area of the dilated mask Y_j^{α}.

In the presented tables, we highlight (in boldface) the best results for each dataset and for each value of the user-specified confidence level $1-\alpha$. We use the format "average $\pm$ standard deviation" to report the observed metrics. For this purpose, the experiments are repeated 100 times, by randomly re-splitting the set of images that are left-out for testing and calibration: the underlying (base) segmentation models are only trained once.

4.2 Results

As shown on the left side of Figure 2 and detailed by Table 1, the proposed approaches "Ours^{1}" and "Ours^{2}" yield smaller coverage gaps than the preexisting approach "MCIS". This is more noticeable on the left side of Figure 2, it shows that the calibration curves of "Ours^{1}" and "Ours^{2}" are closer to the perfect calibration line. However, the performance of MCIS gets closer to the performances

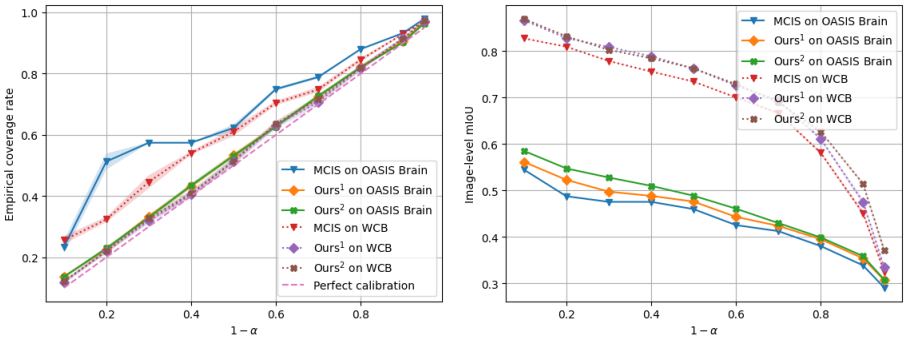

Fig. 2: On the left, the empirical coverage rates on the two datasets. On the right, the corresponding values of the mIoU metric. The error bands indicate the square of the standard deviation(s).

of the proposed approaches as the user-specified confidence level $1-\alpha$ gets closer to 1. On the WBC dataset, as detailed by the lower part of Table 1, the coverage gaps of the baseline Our[1] are generally smaller than the ones of Ours[2]. On this dataset, normalizing the non-conformity scores by the area of the predicted mask does not yield a clear advantage. Note that empirical coverage rates of all the tested methods stay greater or equal to the corresponding user-specified confidence level(s) $1-\alpha$, on both datasets, as required by Equation (4).

Table 1: the empirical coverage rates on the OASIS Brain dataset (top) and on the WBC dataset (bottom), for different values of $1-\alpha$. On both datasets, the empirical coverage rate of the base model (UniverSeg) is 0.0 ± 0.0.

$1-\alpha$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
MCIS	0.234±0.148	0.513±0.162	0.574±0.034	0.574±0.034	0.623±0.105	0.749±0.072	0.789±0.068	0.880±0.041	0.933±0.041	0.979±0.015
Ours[1]	0.137±0.057	**0.226±0.073**	0.334±0.082	**0.432±0.086**	0.534±0.084	0.628±0.077	**0.726±0.077**	**0.821±0.068**	0.907±0.055	0.968±0.027
Ours[2]	**0.136±0.062**	0.229±0.075	**0.330±0.081**	0.435±0.087	**0.531±0.088**	**0.627±0.081**	0.727±0.080	0.821±0.070	**0.902±0.055**	**0.964±0.030**
MCIS	0.256±0.113	0.324±0.093	0.446±0.147	0.541±0.074	0.610±0.116	0.704±0.085	0.749±0.092	0.845±0.079	0.930±0.048	**0.971±0.033**
Ours[1]	**0.118±0.075**	0.221±0.089	**0.319±0.095**	**0.405±0.103**	**0.513±0.103**	**0.631±0.107**	**0.707±0.105**	**0.818±0.092**	0.916±0.060	**0.971±0.033**
Ours[2]	0.122±0.067	**0.220±0.085**	0.326±0.105	0.408±0.100	0.514±0.103	0.636±0.097	0.717±0.102	0.819±0.094	**0.913±0.068**	0.973±0.036

Regarding the areas of the conformalized masks, as shown by the right side of Figure 2, the proposed approaches "Ours[1]" and "Ours[2]" yield higher values of the mIoU than the preexisting approach "MCIS". However, as shown by the same figure and detailed by Table 2, as $1-\alpha$ gets closer to 1, the mIoU of MCIS also gets closer to the performances of the proposed approaches. As detailed by the upper part of Table 2, on the OASIS dataset, the mIoU of the baseline Ours[2] (which normalizes the non-conformity scores by the area of the predicted mask) is higher than the mIoU of the baseline Ours[1]. However, on the WBC dataset, the lower part of Table 2, the advantage of Ours[2] is not as clear. Furthermore, the values of the mIoU are higher on the WBC dataset than on the OASIS Brain

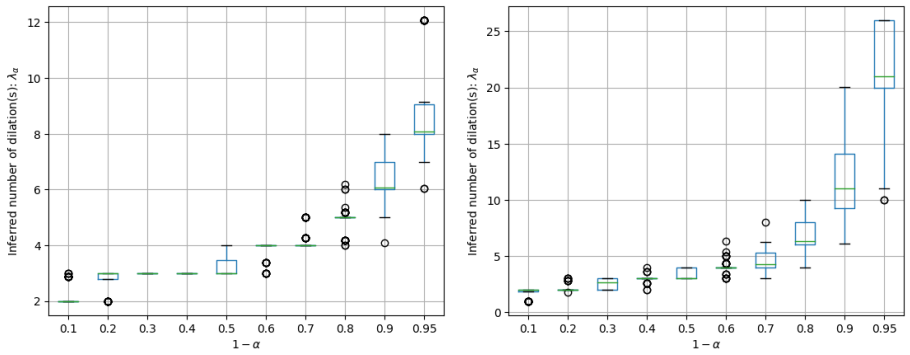

Fig. 3: The number of consecutive dilation(s) λ_α that are applied by the MCIS baseline on OASIS (left) and on WBC (right).

Table 2: the mIoU on OASIS (top) and on WBC (bottom). For the base model (Universeg), the mIoU is 0.843 ± 0.002 on OASIS and 0.932 ± 0.006 on WBC.

$1-\alpha$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
MCIS	0.545±0.028	0.488±0.029	0.476±0.002	0.476±0.002	0.460±0.024	0.426±0.014	0.413±0.016	0.381±0.012	0.339±0.018	0.291±0.022
Ours[1]	0.562±0.010	0.523±0.022	0.498±0.007	0.489±0.005	0.476±0.014	0.443±0.016	0.424±0.010	0.395±0.013	0.354±0.020	0.307±0.024
Ours[2]	**0.585±0.025**	**0.548±0.010**	**0.528±0.010**	**0.510±0.012**	**0.489±0.014**	**0.461±0.015**	**0.430±0.016**	**0.399±0.013**	**0.359±0.023**	**0.308±0.026**
MCIS	0.828±0.028	0.810±0.020	0.779±0.031	0.756±0.017	0.735±0.029	0.701±0.028	0.667±0.045	0.582±0.051	0.451±0.065	0.326±0.068
Ours[1]	0.867±0.022	0.828±0.016	**0.809±0.016**	**0.789±0.017**	0.763±0.020	0.726±0.026	**0.694±0.037**	0.612±0.055	0.476±0.070	0.334±0.071
Ours[2]	**0.870±0.020**	**0.832±0.020**	0.803±0.019	0.785±0.016	**0.763±0.016**	**0.729±0.023**	0.690±0.036	**0.627±0.039**	**0.515±0.058**	**0.371±0.084**

dataset. This can be explained by the fact that the mIoU of the base model is higher on the WBC dataset than on OASIS.

Thus, the advantages of the proposed methods decrease as the user-specified confidence level $1-\alpha$ gets closer to 1, on both datasets, and the advantages of the baseline Ours[2] are not clear on the WBC dataset. As shown by Figure 3, the number of consecutive dilation(s) λ_α (see Equation (2)) that is required to achieve coverage increases as the user-specified confidence level $1-\alpha$ increases. Moreover, the post-processed predictions require a higher number of consecutive dilation(s) on the WBC dataset (the right side of Figure 3) than on the OASIS dataset (the left side of Figure 3). This suggests that the advantages of the proposed methods decrease as the number of applied dilation(s) increases. In other words, the proposed methods are only suited to predictions such that coverage can be achieved with a small number of consecutive dilation(s). Essentially, the effects of the proposed approaches get negligible when a high number of consecutive dilation(s) is applied, since the proposed approaches only differ from the preexisting MCIS approach w.r.t. the last of the (applied) consecutive dilation, as shown by Algorithm 2. Nonetheless, this sensitivity to a high number of consecutive dilation(s) may not be a significant limitation in practice. For example, in real-time systems, the number of sufficient consecutive dilation(s) should be very low (ideally) because the predictions must be post-processed as fast as possible: low latency matters in any most interractive systems.

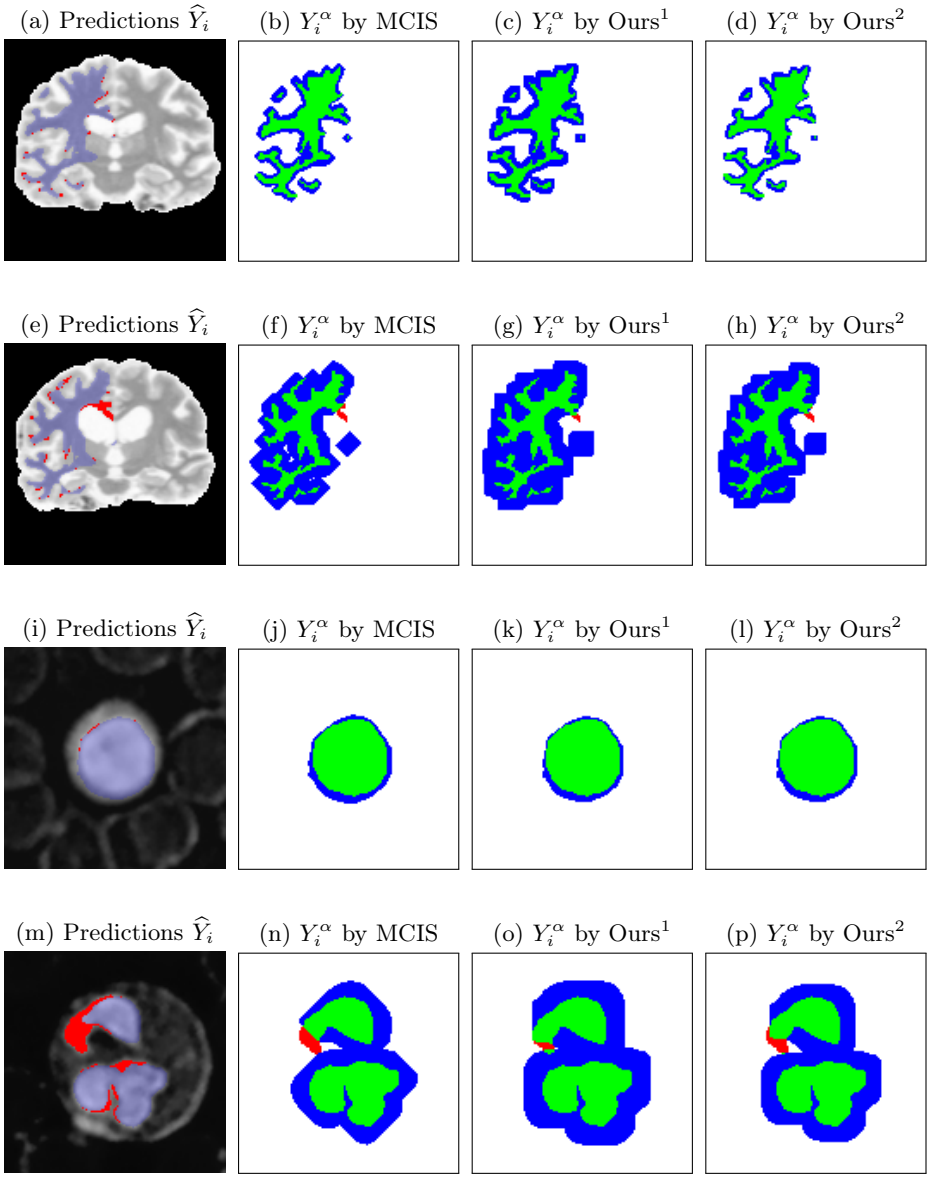

Fig. 4: Qualitative results. The first and third rows show examples where the compared methods succeed, with $(1-\alpha) = 0.1$. The second and fourth rows show examples where the methods fail, with $(1-\alpha) = 0.9$. The predictions of the base model (UniverSeg) are shown in purple, the false negative pixels are in red, the true positive pixels are in green, and the dilated masks are in blue.

Regarding the qualitative results on the first and third rows of Figure 4, coverage is successfully achieved: for these examples, as required by Equation (5), all the false negative pixels (in red) are corrected by the dilation(s). The conformalized masks of our normalized approach Ours^2 are tighter, but the differences are barely noticeable. On the other rows of the Figure 4, coverage is not achieved and the conformalized masks of our approach are larger than the conformalized masks of MCIS. These examples show the limitations of the mIoU metric. First, the mIoU metric is an average. Therefore, in some cases, a method with a high mIoU can still yield conformalized masks that are larger than the conformalized masks of a method with a low mIoU. Second, the mIoU characterizes the global overlap between two segmentation masks: the errors on the boundary of a predicted mask are not more penalized than the error in the interior. The mIoU can be enough for our definition of coverage in Equation (5), but it does not evaluate all the possible desiderata. In fact, our overlap-based definition of coverage may not be right when predicting an accurate contour of the segmented structure is more important than eliminating all the false negative pixels. For such cases, coverage definitions that are based on boundary-aware metrics such as the Hausdorff distance [10] may be more appropriate.

5 Conclusion

In this article, we have proposed a novel MCIS approach that relies on fractional non-conformity scores to reduce the areas of the conformalized masks. The experiments results show that the novel approach reduces the empirical coverage gap and the areas of the conformalized masks, on average. However, the advantages of this novel approach decrease as the number of consecutive dilation(s) that are applied increases. Nonetheless, this may not be a significant limitation because the number of necessary dilation(s) should be small, ideally, unless the applied structuring element δ is too small for the segmented structure or the predictions of the segmentation model are very poor. Future works may explore GPU-accelerated algorithms [2] to reduce the additional computational load that is introduced by the Euclidean distance transform, in Algorithm 2, in Step 10 and Step 24. They may also redefine the coverage criterion in Equation (5), to obtain accurate boundaries for example, or use heuristic estimates of the quality of the (unlabeled) predictions as the normalization scores σ_i in Algorithm 2.

References

1. Angelopoulos, A., Bates, S., Fisch, A., Lei, L., Schuster, T.: Conformal risk control. In: International conference on learning representations. vol. 2024, pp. 55198–55218 (2024)
2. de Assis Zampirolli, F., Filipe, L.: A fast cuda-based implementation for the euclidean distance transform. In: 2017 International Conference on High Performance Computing & Simulation (HPCS). pp. 815–818. IEEE (2017)
3. Barber, R.F., Tibshirani, R.J.: Unifying different theories of conformal prediction. Electronic Journal of Statistics **20**(1), 1428–1474 (2026)

4. Bereska, J.I., Karimi, H., Samavi, R.: Sacp: Spatially-adaptive conformal prediction in uncertainty quantification of medical image segmentation. In: Medical Imaging with Deep Learning (2025)
5. Bloch, I., Heijmans, H., Ronse, C.: Mathematical Morphology, pp. 857–944. Springer Netherlands (2007). https://doi.org/10.1007/978-1-4020-5587-4_14
6. Borgefors, G.: Distance transformations in digital images. Computer vision, graphics, and image processing **34**(3), 344–371 (1986)
7. Brunekreef, J., Marcus, E., Sheombarsing, R., Sonke, J.J., Teuwen, J.: Kandinsky conformal prediction: efficient calibration of image segmentation algorithms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4135–4143 (2024)
8. Butoi, V.I., Ortiz, J.J.G., Ma, T., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Universeg: Universal medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21438–21451 (2023)
9. Dai, M., Luo, W., Li, T.: Robust quality control framework for medical instance segmentation tasks based on conformal prediction theory. In: 2025 International Conference on Computers, Information Processing and Advanced Education (CIPAE). pp. 490–496. IEEE (2025)
10. Dubuisson, M.P., Jain, A.K.: A modified hausdorff distance for object matching. In: Proceedings of 12th international conference on pattern recognition. vol. 1, pp. 566–568. IEEE (1994)
11. Hoopes, A., Hoffmann, M., Greve, D.N., Fischl, B., Guttag, J., Dalca, A.V.: Learning the effect of registration hyperparameters with hypermorph. The journal of machine learning for biomedical imaging **1**, 003 (2022)
12. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine learning **110**(3), 457–506 (2021)
13. Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B.: Boundary loss for highly unbalanced segmentation. In: International conference on medical imaging with deep learning. pp. 285–296. PMLR (2019)
14. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical ai solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. Artif. Intell. Medicine **150**, 102830 (2024)
15. Lie, J., Lysaker, M., Tai, X.C.: A binary level set model and some applications to mumford-shah image segmentation. IEEE transactions on image processing **15**(5), 1171–1181 (2006)
16. Lu, K., Kluger, D.M., Bates, S., Wang, S.: Conformal prediction sets for instance segmentation. arXiv preprint arXiv:2602.10045 (2026)
17. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience **19**(9), 1498–1507 (2007)
18. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE transactions on medical imaging **39**(12), 3868–3878 (2020)
19. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(7), 3523–3542 (2021)
20. Mossina, L., Dalmau, J., Andéol, L.: Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3574–3584 (2024)

21. Mossina, L., Friedrich, C.: Conformal prediction for image segmentation using morphological prediction sets. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 78–88. Springer (2025)
22. Mossina, L., Friedrich, C.: Controlling false positives in image segmentation via conformal prediction. arXiv preprint arXiv:2511.15406 (2025)
23. Muhammad, K., Hussain, T., Ullah, H., Del Ser, J., Rezaei, M., Kumar, N., Hijji, M., Bellavista, P., De Albuquerque, V.H.C.: Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. IEEE Transactions on Intelligent Transportation Systems **23**(12), 22694–22715 (2022)
24. Mukama, B.C., Messoudi, S., Destercke, S., Rousseau, S.: Copula-based conformal prediction for prioritized heterogeneous multi-task learning. Pattern Recognition p. 112347 (2025)
25. Mukama, B.C., Messoudi, S., Destercke, S., Rousseau, S.: Morphological conformal image segmentation with adaptive structuring elements. Congrès Reconnaissance des Formes, Image, Apprentissage et Perception, RFIAP (2026)
26. Nguyen, V.L., Zhang, H., Destercke, S.: Credal ensembling in multi-class classification. Machine Learning **114**(1), 19 (2025)
27. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive confidence machines for regression. In: European conference on machine learning. pp. 345–356. Springer (2002)
28. Ramos, D.L., Hortua, H.J.: Deep bayesian segmentation for colon polyps: Well-calibrated predictions in medical imaging. Biomedical Signal Processing and Control **104**, 107383 (2025)
29. Rosenfeld, A., Pfaltz, J.L.: Distance functions on digital pictures. Pattern recognition **1**(1), 33–61 (1968)
30. Rousseau, A.J., Becker, T., Appeltans, S., Blaschko, M., Valkenborg, D.: Post hoc calibration of medical segmentation models. Discover Applied Sciences **7**(3), 180 (2025)
31. Song, B., Chan, T.: A fast algorithm for level set based optimization. UCLA Cam Report **2**(68) (2002)
32. Timans, A., Straehle, C.N., Sakmann, K., Nalisnick, E.: Adaptive bounding box uncertainties via two-step conformal prediction. In: European Conference on Computer Vision. pp. 363–398. Springer (2024)
33. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer Nature (2022)
34. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: A survey. IET image processing **16**(5), 1243–1267 (2022)
35. Wu, D., Guo, Z., Li, A., Yu, C., Gao, C., Sang, N.: Conditional boundary loss for semantic segmentation. IEEE Transactions on Image Processing **32**, 3717–3731 (2023)
36. Zheng, X., Wang, Y., Wang, G., Liu, J.: Fast and robust segmentation of white blood cell images by self-supervised learning. Micron **107**, 55–71 (2018)
37. Zou, K., Chen, Y., Huang, L., Zhou, N., Yuan, X., Shen, X., Wang, M., Goh, R.S.M., Liu, Y., Tham, Y.C., et al.: Toward reliable medical image segmentation by modeling evidential calibrated uncertainty. IEEE Transactions on Cybernetics (2025)