Pattern Recognition 179 (2026) 113753

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Segmetron: Sample-efficient model-agnostic semantic segmentation with a trustworthy reject option via PQ learning☆

T.A. Bohoran [a], K.S. Parke [b], A. Cowley [b], G.S. Gulsin [b], J. Yeo [b], A. Dattani [b], G.P. McCann [b], A. Giannakidis [a,c,*]

[a] School of Science and Technology, Nottingham Trent University, Nottingham, NG11 8NS, UK
[b] Department of Cardiovascular Sciences, University of Leicester and the NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, LE3 9QP, UK
[c] Archimedes Unit in Artificial Intelligence, Data Science and Algorithms, Athena Research Center, Marousi, 15125, Greece

ARTICLE INFO

Keywords:
Covariate shift
Detection
Semantic segmentation
Hypothesis test
Arbitrary test distribution
Dataset shift
PQ learning

ABSTRACT

Semantic segmentation can help gain a deeper understanding of a depicted scene and deliver a variety of transformative technologies. However, its application is limited by the covariate shift and the lack of reliable detection techniques. In this study, we introduce a trustworthy, sample-efficient, distribution-free and model-agnostic hypothesis test, named Segmetron, to detect image-level covariate shift in semantic segmentation. To assess an unlabelled target domain, Segmetron relies on an existing (but random) pre-trained semantic segmentation model and the labelled samples used to train it. The test statistic is based on the sample disagreement rate of two ensemble models trained to disagree with the baseline segmenter on unseen samples from the training and deployment sets, respectively. To obtain theoretical guarantees on unknown arbitrary test distributions, we build on recent work on the PQ learning setting of selective classification and extend it to a different discriminative model (i.e. segmenters). To train the enforced disagreement segmenters of each ensemble, we innovatively propose loss functions (to agree) which are more apropos to the semantic segmentation task and comply with the training of the baseline segmenter. We demonstrate that Segmetron outperforms other state-of-the-art techniques in terms of statistical power on two challenging real-world tasks from the cardiovascular magnetic resonance imaging field, concerned with two or more semantic classes, given access to only one image. This work aligns with "Responsible AI" principles, supporting reliable deployment of AI by enhancing robustness. It can potentially enable the widespread adoption of deep learning semantic segmentation technologies across various fields.

1. Introduction

1.1. Background

Semantic segmentation is a central task in computer vision serving as the cornerstone of downstream analysis in autonomous driving [1], medical decision-making (diagnosis and treatment) [2], vision-enabled robots [3], underwater scenarios [4,5] etc. Even though state-of-the-art (SOTA) deep learning (DL) semantic segmentation models shine out within the training-data distribution, they completely flop outside of it [6]. The disparity in the distribution of the input samples between model training and testing (deployment), also known as covariate shift, is the rule (rather than the exception) and a cause of serious safety and reliability concerns with respect to real-life operation. The covariate shift is exacerbated in the medical imaging field due to the diverse image acquisition protocols, patient population heterogeneity, and wide variety of medical conditions among other factors [7]. Therefore, automated covariate shift detection techniques for semantic segmentation whose results are rigorously guaranteed in the absence of labelled target data would be of enormous value for delivering DL transformative technologies. Nevertheless, current literature has paid limited attention to the development of pertinent techniques [8].

1.2. Related work

There are relatively few papers directly combining semantic segmentation with an image-level reject/abstention option. The authors

☆ Tuan Aqeel Bohoran is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 801604. This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

* Corresponding author at: School of Science and Technology, Nottingham Trent University, Nottingham, NG11 8NS, UK.
E-mail address: archontis.giannakidis@ntu.ac.uk (A. Giannakidis).

https://doi.org/10.1016/j.patcog.2026.113753
Received 13 May 2025; Received in revised form 26 March 2026; Accepted 11 April 2026
Available online 22 April 2026

of [9] formalised the theoretical problem of selective prediction for semantic segmentation and addressed the design of an optimal confidence estimator when a specific model is given. The scope of their study is nevertheless limited in binary segmentation tasks and single segmenter models. However, it is known that the selection functions which are based on the statistics of ensembles of multiple models are more robust to distribution shifts and are able to produce more accurate results. Recently, approaches for detecting (as well as generalising) covariate shift in semantic segmentation by leveraging generative models have been proposed [10,11]. However, such methods face the fundamental limitation that they cannot guarantee detection of arbitrary covariate shifts outside the generative prior's capacity. In practice, the model's ability to detect unseen shifts is tied to the diversity of synthesised examples provided during training and the learned invariances. There are several interesting works that explore uncertainty quantification in semantic segmentation [12–14]. This is a closely related task which could be used to implement a kind of "reject when uncertain" logic. However, such approaches depend on a rejection threshold that is typically determined in an ad-hoc manner [15–17]. Selective semantic segmentation works that rely on trainable rejection gates were put forward in [18–20]. However, covariate shift detection is fundamentally a statistical inference problem, not a prediction problem. Prediction methods offer no theoretical guarantees outside the training distribution, and will fail on novel covariate shifts. A more principled solution is needed.

A fit-for-purpose approach for detecting covariate shift in image classification has been to cast the problem as a two-sample statistical hypothesis test. The first sample comprises the training data, whereas the second sample is the latest deployment data. Then, the null hypothesis, $\mathcal{H}_0$, is that the samples were drawn from the same probability distribution, as opposed to the alternative hypothesis $\mathcal{H}_1$ that the two distributions are different. A non-parametric deep kernel-based two-sample hypothesis test was proposed in [21]. The test statistic was based on the maximum mean discrepancy (MMD), which measures the differences between the two kernel mean embeddings. The kernels were parameterised by deep neural networks trained to optimise the test power, rendering this test particularly suited for high-dimensional (such as image) data. The authors of [22] designated H-divergence as a test statistic for two-sample tests. H-divergence is based on the generalised entropy defined by the maximum log-likelihood of readily available deep generative models. It allows to take advantage of inductive biases for each type of data, leading to improved test power. A baseline alternative for detecting covariate shift is to perform a non-parametric Kolmogorov–Smirnoff (KS) test directly on the distribution of relative Mahalanobis distance (RMD) confidence scores obtained for the two samples [23]. The RMD metric was initially introduced to improve out-of-distribution (OOD) detection [23]. Lastly, another way to conduct a two-sample test for recognising covariate shift is to use a classifier-based method. According to this approach, a binary classifier is trained to differentiate between source and target samples, and the test statistic could be based on the classifier's accuracy in a held-out test sample [24].

All the above hypothesis testing studies looked into domain shift complications suffered by DL classification models. Covariate shift in the richer semantic segmentation is far less investigated [8]. Even though semantic segmentation and classification are related tasks, task-specific studies are valuable given that learning algorithms may behave inconsistently across different tasks. The findings obtained in classification studies might not be valid for semantic segmentation. As an illustration, the calibration methods, that have been proposed in the literature to deal with the overconfidence issue in DL classification models, behave differently in semantic segmentation [8]. In addition, semantic segmentation is a task of increased complexity when compared to classification, as the dense individual pixel predictions must ensure that they are spatially consistent and that they pick up adjacent pixel relationships (local context) [25]. Moreover, semantic segmentation models have to effectively deal with occlusions [25].

1.3. Our contribution

In this study we make the following pivotal contributions:

- We develop a reliable, sample-efficient, distribution-free and model-agnostic hypothesis test, named the Segmetron (Fig. 1), to detect image-level covariate shift in semantic segmentation. To assess an unlabelled target domain, Segmetron relies on an existing (but random) pre-trained semantic segmentation model and the labelled samples (pixels) used to train it. The test statistic of the one-sided hypothesis test is based on the rate of sample disagreement of two ensemble models trained to disagree with the baseline segmenter on unseen samples from the training and deployment sets, respectively.
- To obtain strong performance theoretical guarantees on unknown arbitrary test distributions, we build on recent work on the PQ learning setting of selective classification (SC) and extend it to a different discriminative model (i.e. segmenters).
- To train the enforced disagreement segmenters (EDSs) of each ensemble model to learn the same generalisation region as the pre-trained semantic segmentation model, we innovatively propose loss functions (to agree) which are more apropos to the semantic segmentation task and comply with the training of the baseline segmenter.
- We examine real-world covariate shifts that arise naturally (i.e. without human intervention), as opposed to previous studies on semantic segmentation robustness which relied on synthetic domain shifts, obtained by injecting noise/blur or crafting adversaries [26]. In particular, we analyse two covariate shifts from the cardiovascular magnetic resonance imaging (CMR) field, concerned with both binary (aorta, background) and multi-class (left ventricle, right ventricle, myocardium, background) semantic segmentation tasks, ensuring diversity of set-ups.
- We demonstrate that Segmetron outperforms other SOTA techniques in terms of statistical power on the two semantic segmentation tasks, given access to only one image.

2. Materials and methods

2.1. Covariate shift

Definition 1: Let X and Y be the input and output (label) spaces, respectively, defining the semantic segmentation task. Then, the input and output data are simply random variables. Labels are discrete (C classes). Assume that the marginal distributions of X and Y in the training (source) and testing/deployment (target) domains are denoted by $P_s(X)$, $P_s(Y)$ and $P_t(X)$, $P_t(Y)$, respectively. Similarly, the conditional distributions of the output variables given the input variables in the two domains are denoted by $P_s(Y|X)$ and $P_t(Y|X)$. Covariate shift refers to the situation where the marginal distribution of the input variables varies across the source and target domains (i.e. $(P_s(X) \neq P_t(X))$), whereas the conditional distribution of the output variables given the inputs remains unaltered (i.e. $P_s(Y|X) = P_t(Y|X)$) [27]. To put it simply, this type of shift means that the input data distribution is different between the source and target domains, but the relationship between the input and output variables is the same.

2.2. The PQ learning setting of selective classification

The goal of SC (a.k.a. classification with a reject option) is to learn a classifier model which is allowed to abstain from making predictions when it is not adequately confident [28]. Unlike standard classifiers, which are forced to provide a prediction for every input, a selective classifier can choose to not classify specific examples if it deems the predictions unreliable. This allows SC models to achieve higher accuracy

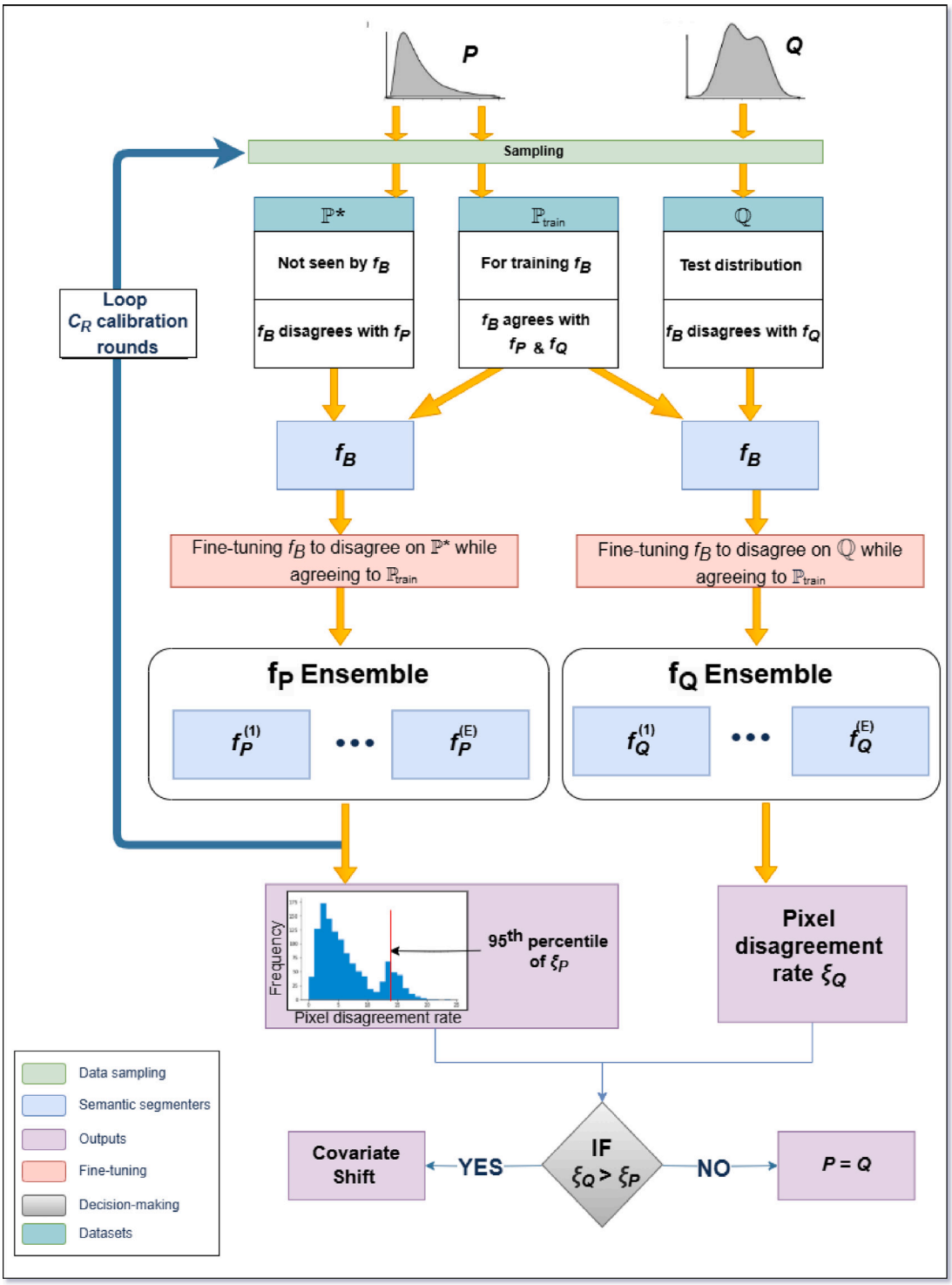


Fig. 1. The Segmetron hypothesis test.

by reducing the number of misclassifications at the expense of coverage (i.e. the fraction of inputs on which predictions are made).

The PQ learning setting of SC has permitted to obtain strong theoretical guarantees on learning with arbitrary and potentially adversarial future test examples [28]. Their work represented a great leap forward, since the specific problem was considered intractable up until then. The authors showed that both the finite-sample error on the random and unknown test distribution Q and the rejection rate on the training distribution P can jointly remain bounded within an acceptable limit ϵ with high probability 1-δ.

Formally, in the PQ learning setting of SC, we are given: (i) a training set of n samples $(x_1, x_2, \ldots, x_n)$, drawn i.i.d. from P over the input space X, (ii) the labels $(f(x_1), f(x_2), \ldots, f(x_n))$ for some unknown target function $f \in F$ of Vapnik–Chervonenkis (VC) dimension d, (iii) an unlabelled test set of n samples $(\dot{x}_1, \dot{x}_2, \ldots, \dot{x}_n)$ (i.i.d. from Q), and (iv) the bound parameter ϵ, which also controls the trade-off between errors

and rejections. The output is a selective classifier $h|_S$, which is allowed to only predict on certain examples in a subset $S \subset X$, and otherwise abstain from predicting (S is also an output). The theoretical bound is given by the below definition [28].

Definition 2: Learner $L(\epsilon, \delta, n)$ PQ-learns a function class F if, for any distributions P and Q over the input space X and any target function $f \in F$, the output $h|_S, = L(P, f(P), Q)$ satisfies:

$$\Pr_{x\sim P^n,\ \tilde{x}\sim Q^n}\left[\texttt{Reject}_P + \texttt{Err}_Q \leq \epsilon\right] \geq 1-\delta, \tag{1}$$

where $\texttt{Reject}_P$ is the rejection rate of the classifier on the training distribution P, and $\texttt{Err}_Q$ is the error rate on the test distribution Q. Learner L PQ-learns F if it runs in non-exponential time and there is a polynomial p such that $L(\epsilon, \delta, n)$ PQ-learns F for every $\epsilon, \delta > 0, n \geq p(1/\epsilon, 1/\delta)$. A recent implementation of the PQ learning setting of SC is Rejectron [28]. It is shown that Rejectron achieves an error bound of $\widetilde{\mathcal{O}}(\sqrt{d/n})$, for any class F of functions with a bounded VC dimension d, where the $\widetilde{\mathcal{O}}$ notation hides logarithmic factors including the dependence on the failure probability δ.

2.3. Problem set-up: Semantic segmentation with a trustworthy reject option

Problem 1: Let $f_B : X \to Y$ be a semantic segmentation model from a function class F that maps from space X to a discrete set of classes $Y = \{1, \ldots, C\}$. Suppose f_B was trained on a dataset of labelled pixel samples (x_i, y_i) for $i = 1, \ldots, n$ where each x_i is drawn identically from a distribution P over X. During testing/deployment, f_B is asked to predict on new unlabelled samples from an arbitrary distribution Q over X.

We define as "semantic segmentation with a trustworthy reject option" the problem of building an automated segmenter that abstains from predicting on those samples of Q for which there is covariate shift and predicts otherwise. The goal is that the algorithm does so with strong guarantees. Even though samples are individual pixels, the decision whether to predict or not (or, equivalently, whether there is covariate shift or not) is taken at image-level.

2.4. Enforced disagreement segmenter

The EDS is a modification of the standard segmenter, designed to enhance the model's sensitivity to shifts in data distribution. Unlike typical segmenters that aim for accurate predictions across all examples, an EDS is tailored to maximise disagreement on specific out-of-distribution data while maintaining consistent predictions on in-distribution image-level data. This approach leverages the inherent variability in the model's response to different data distributions to detect shifts.

An EDS is characterised by the following properties: **(i) Model Consistency:** It belongs to the same model class as the base segmenter and is trained using the same algorithm, ensuring that it does not deviate in fundamental learning capabilities. **(ii) In-Distribution Performance:** It achieves similar performance on unseen samples that follow the in-distribution, verifying that the model's utility is retained for familiar data. **(iii) Maximal Disagreement:** On elements of a dataset $\mathbb{Q}$, representing a potential out-of-distribution set, the EDS is trained to disagree maximally with the predictions of the base segmenter, without compromising its performance on in-distribution data.

The operational mechanism of an EDS involves training the model to identify and emphasise discrepancies between the predicted and actual semantic segmentations on new unseen datasets. This is achieved by: **(i)** Training the base segmenter on a labelled dataset from distribution P. **(ii)** Developing the EDS by further training on a subset from distribution Q, tweaking it to maximise prediction disagreement specifically on Q while ensuring it agrees with the base segmenter's predictions on P. **(iii)** Applying early stopping in the EDS training, if validation performance drops by a certain amount to avoid catastrophic overfitting in small sample regimes.

The primary utility of an EDS lies in its ability to act as a diagnostic tool for flagging up shifts in data distribution that might affect the model's performance, providing an early warning system for potential degradation in model accuracy due to distribution changes. The development of the EDS represents a strategic shift in handling data distribution changes in semantic segmentation. By focusing on the disagreement in predictions between known and new data distributions, EDS offers a robust mechanism for enhancing the reliability of deployed semantic segmentation systems.

2.5. Learning to agree and disagree

The goal of an EDS is to agree on elements of a dataset $\mathbb{P}^*$ from distribution P. In addition, the training of the EDS should not deviate from the learning of the baseline segmenter. To this end, and unlike [29], we innovatively propose to use loss functions (Focal Tversky loss or Dice loss) that are better suited for semantic segmentation tasks with highly unbalanced data.

The Focal Tversky loss is an enhancement of the Tversky loss, aimed at addressing class imbalances in image segmentation tasks. It is defined as:

$$\text{Focal Tversky Loss} = (1 - \text{Tversky Loss})^{\gamma} \tag{2}$$

Where γ is a focusing parameter that controls the contribution of hard-to-segment pixels, and the Tversky loss is given by:

$$\text{Tversky Loss} = \frac{\sum_i^N p_{0i} g_{0i}}{\sum_i^N p_{0i} g_{0i} + \tilde{\alpha} \sum_i^N p_{0i} g_{1i} + \beta \sum_i^N p_{1i} g_{0i}}. \tag{3}$$

In this formula, N is the total number of pixels, p_{0i} is the predicted probability that pixel i belongs to the target class, and p_{1i} is the probability that pixel i is part of the background. Similarly, g_{0i} is 1 if pixel i belongs to the target class, and 0 otherwise, while g_{1i} represents the opposite. The parameters $\tilde{\alpha}$ and β balance the penalties for false positives and false negatives, respectively. The Focal Tversky loss helps control class imbalances and focuses on difficult cases.

The Dice loss is another popular loss function for image segmentation tasks. It is based on the Dice coefficient, which measures the overlap between the predicted segmentation and the ground truth. The Dice loss is defined as:

$$\text{Dice Loss} = 1 - \frac{2 \sum_i^N p_{0i} g_{0i}}{\sum_i^N p_{0i} + \sum_i^N g_{0i}}. \tag{4}$$

Here, p_{0i} is the predicted probability that pixel i belongs to the target class, and g_{0i} is 1 if the pixel is correctly segmented as part of the target class, and 0 otherwise.

The Dice loss directly optimises for the overlap between predicted and actual segments. Maximising this overlap helps reduce false positives and false negatives, leading to improved segmentation performance.

Inspired by the work of [29], we relied on the disagreement cross entropy (DCE) to train segmenters that maximise disagreement on out-of-distribution data. This loss function extends the classical cross entropy loss by encouraging pixel predictions to diverge from the true class label on new, unseen data distributions, thereby effectively identifying distribution shifts.

The DCE loss for a segmenter predicting a distribution over C classes is defined as:

$$L_{DCE}(\widetilde{y}, f_B(x_i)) = \frac{1}{1-C} \sum_{c=1}^{C} \mathbb{1}_{f_B(x_i)\neq c} \log p(c|x_i), \tag{5}$$

where $\widetilde{y}$ is the predictive distribution by the ensemble segmenter over C classes, $f_B(x_i)$ is the predicted pixel label by the baseline segmenter, and $p(c|x_i)$ is the probability that the ensemble segmenter predicts class c on pixel i. The indicator function $\mathbb{1}_{f_B(x_i)\neq c}$ equals 1 when $f_B(x_i)$ is not equal to c, pushing the ensemble segmenter to assign higher probabilities to incorrect classes.

The DCE has many attractive properties such as being very stable to optimise using gradient descent methods, having a bounded global minimum, and satisfying

$$\forall p \in P \quad \min_{q \in Q} L_{DCE}(q; y) \leq L_{DCE}(p; y) \tag{6}$$

meaning that for each probability vector in P, there is a corresponding probability vector in Q that attains a score that is at least as low [29].

Then, the overall training objective L_{EDS} can be obtained by combining one of the two losses proposed above that enforce agreement on P, and the L_{DCE} for samples from the new distribution Q:

$$L_{EDS}(P, Q) = \sum_{(x_i, y_i) \in P} L_{agr}(\widetilde{y}, y_i) + \lambda \sum_{x_i \in Q} L_{DCE}(\widetilde{y}, f_B(x_i)), \tag{7}$$

where L_{agr} is either the Focal Tversky or Dice loss and λ is a tuning parameter that balances fitting to P and learning to disagree on Q.

2.6. Segmetron

To detect covariate shift in semantic segmentation, Segmetron expands on [29], which in turn had built on earlier work [28]. Segmetron conducts a statistical hypothesis test between the distributions of a potentially shifted new dataset $\mathbb{Q}$ and a dataset $\mathbb{P}^\star$ which was not seen during training but it is known to be from the same distribution as the training data. It adopts a transductive approach in the sense that it is constructed by: (i) training segmenters using L_{EDS} on observed training (and test) cases, (ii) performing reasoning to the specific test data.

Let f_B be a baseline segmenter trained on dataset $\mathbb{P}_{train}$ comprising input data (pixels) sampled from P and the corresponding masks. Assume f_Q is a segmenter which agrees (i.e. segments alike) with f_B on $\mathbb{P}_{train}$ and disagrees on a dataset $\mathbb{Q}$ sampled from an unknown arbitrary Q. We use ξ_Q to denote the rate at which f_Q disagrees with f_B on n unseen pixels from Q, and ξ_P to denote the rate at which f_Q disagrees with f_B on n unseen pixels from P. Then, by viewing semantic segmentation as a pixel-wise classification problem outputting a dense mask with a predicted class for every pixel, we argue that ξ_Q being greater than ξ_P, implies a covariate shift. The proof for the above is based on the fact that under the null hypothesis ($P = Q$), the upper bound of the probability that f_Q is more likely to disagree on Q than P is 0.5 [29]:

$$P = Q \quad \Rightarrow \quad \mathbb{P}(\xi_Q > \xi_P) \leq \frac{1}{2}\left(1 - 4^{-n}\binom{2n}{n}\right) < \frac{1}{2}. \tag{8}$$

Segmetron trains two EDS ensembles f_P and f_Q. f_P is trained to disagree on unseen $\mathbb{P}^\star$ from P, and f_Q is trained to disagree on $\mathbb{Q}$. After the training is completed, ξ_P and ξ_Q are calculated. If $\xi_Q > \xi_P$, then there is a covariate shift in the data (alternative hypothesis). If $\xi_P \geq \xi_Q$, then $\mathbb{Q}$ dataset is in P distribution (null hypothesis). Segmetron is distribution-free which means no assumptions about P and Q are made.

Two-sample hypothesis tests are associated with Type I errors (i.e. rejecting the true $\mathcal{H}_0$) and Type II errors (i.e. failing to reject a false $\mathcal{H}_0$). The upper bound of the probability of Type I error is controlled by choosing an appropriate significance level. The probability of not making a Type II error is called test power, and is regarded the main efficacy measure in null hypothesis testing. In this study, in order to test for shift on a set $\mathbb{Q}$, we follow the typical setup [30] by: (i) performing a permutation test to guarantee a significance level (or, else, a bounded Type I error), and (ii) empirically measuring the test power. To obtain the Segmetron result at a significance level α (=0.05), we train Segmetron for C_R (=100) calibration rounds with random $\mathbb{P}^\star$. Then, the Segmetron result is significant at the 5% level if ξ_Q is greater than the (1-α) percentile of ξ_P. The pseudocode for Segmetron algorithm is given below (**Algorithm 1**). The flowchart is illustrated in Fig. 1.

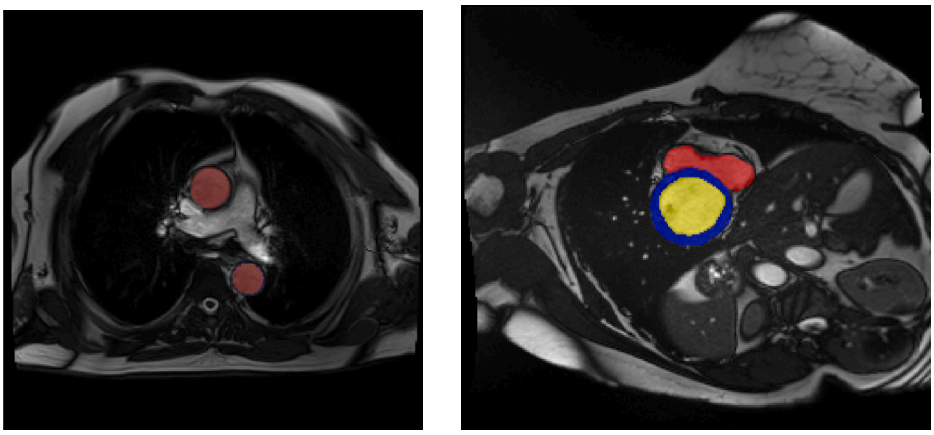
Algorithm 1 The Segmetron algorithm

Input: $\mathbb{P}$: labelled dataset, (x_i, y_i), $\mathbb{Q}$: unlabelled dataset, (x_i), L_A: learning algorithm, C_R: calibration rounds = 100, E: ensemble size = 5, α: significance level = 0.05, M_e: evaluation metric (Dice accuracy), ϵ: tolerance (=0.05), e_m: max epochs (=4).
Output: test result for covariate shift at significance level α.
Partition $\mathbb{P}$ into $\mathbb{P}_{\text{train}}, \mathbb{P}_{\text{val}}, \mathbb{P}^\star$
$N_S \leftarrow |Q|$, $\xi_P \leftarrow [\ ]$
$f_B \leftarrow L_A(P_{\text{train}}, P_{\text{val}})$
for $C_R \leq 100$ **do**
  $\mathbb{P}^\star \leftarrow \text{RandomSampling}(\mathbb{P}^\star, N_S)$
  while $n > 0$ and epochs $\leq N_S$ **do** // Train ensembles of EDSs on $\mathbb{P}^\star$
    $\hat{\mathbb{P}^*} \leftarrow \{(x, f_B(x)) \mid x \in \mathbb{P}^*\}$ // Infer pseudo labels on $\mathbb{P}^*$ using f_B
    //Dataloader using $\mathbb{P}_{train}$ and $\mathbb{P}^*$
    $\mathbb{P}, \mathbb{P}^* \leftarrow \text{Batched}(\{(x, y) \mid (x, y) \in \mathbb{P}_{train} \wedge (x, f_B(x) \in \mathbb{P}^*\})$
    Initialise $f_P \leftarrow f_B$
    $m_0 \leftarrow M_e(f_B, \mathbb{P}_{\text{val}})$ // Compute the validation performance of f_B
    while $M_e(f_B, \mathbb{P}_{\text{val}}) > m_0 - \epsilon$ and iterations $< e_m$ **do**
      for batch in $\mathbb{P}, \mathbb{P}^*$ **do**
        $x_P, y_P \leftarrow \{(x, y) \mid (x, y) \in \text{batch and } (x, y) \in \mathbb{P}_{\text{train}}\}$
        $x_{P^*}, y_{P^*} \leftarrow \{(x, f_B(x)) \mid x \in \text{batch} \wedge x \in \hat{\mathbb{P}^*}\}$
        Update f_P using L_A for (x_P, y_P) and disagreement update for (x_{P^*}, y_{P^*})
      end for
    end while
    return f_P
    Filter out agreed pixels $\mathbb{P}^\star \leftarrow \{x \mid x \in \mathbb{P}^\star \text{ and } f_B(x) = f_P(x)\}$
    Update disagreement rate: $\xi_P \leftarrow 1 - \frac{|\mathbb{P}^\star|}{N_S}$
  end while
  Append ξ_P to $[\xi_P]$ list
end for
while $n > 0$ and epochs $\leq N_S$ **do**
  $\hat{\mathbb{Q}} \leftarrow \{(x, f_B(x)) \mid x \in \mathbb{Q}\}$ // Infer pseudo labels on $\mathbb{Q}$ using f_B
  $\mathbb{P}, \mathbb{Q} \leftarrow \text{Batched}(\{(x, y) \mid (x, y) \in \mathbb{P}_{train} \wedge (x, y) \in \mathbb{Q}\})$ //Dataloader using $\mathbb{P}_{train}$ and $\mathbb{Q}$
  Initialise $f_Q \leftarrow f_B$
  $m_0 \leftarrow M_e(f_B, \mathbb{P}_{\text{val}})$ // Compute the validation performance of f_B
  while $M_e(f_B, \mathbb{P}_{\text{val}}) > m_0 - \epsilon$ and iterations $< e_m$ **do**
    for batch in $\mathbb{P}, \mathbb{Q}$ **do**
      $x_P, y_P \leftarrow \{(x, y) \mid (x, y) \in \text{batch and } (x, y) \in \mathbb{P}_{\text{train}}\}$
      $x_Q, y_Q \leftarrow \{(x, f_B(x)) \mid x \in \text{batch} \wedge x \in Q\}$
      Update f_Q with L_A for (x_P, y_P) and disagreement update for (x_Q, y_Q)
    end for
  end while
  return f_Q
  Filter out agreed pixels $\mathbb{Q} \leftarrow \{x \mid x \in \mathbb{Q} \text{ and } f_B(x) = f_Q(x)\}$
  Update disagreement rate: $\xi_Q \leftarrow 1 - \frac{|Q|}{N_S}$
end while
return $\xi_Q > (1 - \alpha)$ quantile of ξ_P

2.7. Experiments

2.7.1. Datasets

Segmetron was validated in binary and multi-class semantic segmentation tasks, both from the CMR medical imaging field. The former task involved segmenting aorta (both ascending and descending) from steady-state free precession (SSFP) cine CMR images. The initial (unshifted) dataset (Fig. 2(a)) is described in [31]. It consists of 340 3D (2D space + time) training datasets and 84 testing datasets from the same distribution. Each patient dataset comprised 30 time points (2D



(a) Binary Segmentation (b) Multi-class Segmentation

Fig. 2. The binary (a) and multi-class (b) semantic segmentation tasks analysed in this study. The two classes of the former task are (ascending and descending) aorta (brown) and background. The latter task involves four semantic classes: left ventricle (yellow), myocardium (blue), right ventricle (red), and background. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

images). All patients had aortic stiffness-related diseases. To assess covariate shift detection, 280 additional (unlabelled) patient datasets were acquired by the same clinical institution, for which the base segmenter failed due to covariate shift caused by acquisition hardware variability, elevated flow velocity in the aortic valve, and different patient demographics.

The multi-class segmentation task involved four classes, namely left ventricle, right ventricle, myocardium, and background. The dataset (Fig. 2(b)) from the Automated Cardiac Diagnosis Challenge (ACDC) served as the initial dataset sampled from P [32]. It consists of 100 training 3D (spatial) datasets (namely, 20 healthy volunteers, 20 patients with previous myocardial infarction, 20 patients with dilated cardiomyopathy, 20 patients with hypertrophic cardiomyopathy, 20 patients with abnormal right ventricle) and 50 testing 3D datasets. These are all 3D CMR data obtained using the SSFP sequence. The shifted dataset, acquired using different CMR sequences (namely late gadolinium enhancement (LGE) or T2), comes from the 2019 Multi-sequence Cardiac MR Segmentation Challenge (MS-CMRSeg 2019) [33], and involves 45 3D datasets from cardiomyopathy patients.

2.7.2. Base segmenters

The binary base semantic segmentation model is described in [31]. It is a UNet-based model, equipped with 2D Bi-directional ConvLSTM layers with densely connected convolutions, dropout and batch normalisation layers. This architecture carries out concurrent spatio-temporal learning of the video inputs, and employs non-linear functions to couple the feature maps from the encoder.

The multi-class base segmentation model was by [34], and it achieved the overall 3rd place winner award in the ACDC challenge in 2019. It is a modified version of the 2D UNet by setting the total of feature maps in the upsampling path transpose convolutions to be equal to the number of classes (=4).

2.7.3. State-of-the-art approaches and evaluation

Segmetron was juxtaposed with SOTA two-sample tests that are based on MMD [21], H-divergence [22], RMD [23], as well as other mainstream uncertainty quantification and OOD detection criteria. To report performance, the True Positive Rate at a 5% Significance Level (TPR@5) was employed. This metric indicates the frequency with which a method correctly detects covariate shift ($P \neq Q$), while maintaining a false positive rate of only 5%. It is equivalent to the statistical power of a test with a significance level (α) of 5%.

2.8. Implementation

For detecting covariate shift in the binary semantic segmentation task, $\mathbb{P}_{train}$, $\mathbb{P}_{val}$, $\mathbb{P}^*$, and $\mathbb{Q}$ were randomly selected. $\mathbb{P}^*$ was obtained from the test set of the respective study. Each of the four datasets involved 3D (i.e. 2D+time) data from a single patient, comprising 1966080 (=30 × 256 × 256) pixels (samples). Each ensemble comprised five EDSs. f_P and f_Q were both initialised using the weights of f_B. Each EDS was trained for five epochs, with early stopping if validation performance dropped by 5%. The Focal Tversky loss was employed to enforce agreement on P, and also to mirror the base training. We set $\tilde{\alpha} = 0.8, \beta = 0.8$ and $\gamma = 1$ to improve model convergence and recall [31]. The recommended learning rates and batch sizes by the authors of the original study [31] were utilised. The L_{EDS} tuning parameter λ was chosen to be equal to $\frac{1}{|\mathbb{Q}|+1}$, as suggested by [28]. $\mathbb{P}^*$ datasets from P were permuted across 100 random calibration runs to generate 100 f_P ensemble models, from which we obtained 100 pixel disagreement rate values. ξ_P was calculated as the 95th percentile of these rates. One test run was used to produce f_Q which gave the pixel disagreement rate value ξ_Q. The permutation test allowed to deliver strong statistical guarantees. To enhance the presentation of the results, the number of disagreement pixels (rather than the rate values) is purposefully illustrated. The above experiments were executed for 100 randomly selected $\mathbb{Q}$ sets, and the TPR@5 is reported. To validate the usefulness of the chosen loss function, the pixel disagreement and in-distribution accuracy are also plotted, both as functions of the ensemble size.

For the multi-class cardiac semantic segmentation task, all the details were the same as above, except that each dataset involved 3D spatial image data, comprising 719,104 (=16 × 212 × 212) pixels. In addition, the Dice loss function was chosen to agree on P, following the training of the base segmenter. The learning rates and batch sizes that were utilised in the experiments were those recommended by the authors of the baseline study [34].

The penultimate layer of the pre-trained base models was used to test for covariate shift using the RMD approach [23]. The Kolmogorov–Smirnov (KS) test [35] was performed on the distribution of RMD confidence scores derived from $\mathbb{P}^*$ and $\mathbb{Q}$ for both binary and multi-class semantic segmentation tasks. The `scipy.stats.ks_2samp` implementation of the KS test was employed, providing directly the p-values.

For the MMD test, the feature extraction network ϕ_ω was a five-layer fully-connected neural network using soft-plus activations. The

number of neurons in hidden and output layers was set to 50. The Adam optimiser was used and the dropout rate was set to 0 during training. The original source code provided by the authors at https://github.com/fengliu90/DK-for-TST was employed.

The H-Divergence implementation involved training variational autoencoder models on both $\mathbb{P}^*$ and $\mathbb{Q}$ datasets individually, and on their uniform mixture $\frac{(\mathbb{P}^*+\mathbb{Q})}{2}$. The variational autoencoder loss, comprising reconstruction loss and a Kullback–Leibler divergence term, was used to compute a test statistic that quantifies the difference between the distributions by comparing the entropy of the mixture to the individual datasets. To realise the H-Divergence, the following general class of continuous functions was chosen

$$\phi(\theta, \lambda) = \frac{(\theta_s + \lambda_s)^{\frac{1}{s}}}{2} \tag{9}$$

for $s > 1$, which generalises the H-Jensen Shannon divergence for s=1 and the H-Min divergence for $s = \infty$. The loss function $l(x, a)$ was chosen as the negative log-likelihood of x under a distribution a, where a belongs to a model family A. The implementation was validated through a permutation testing scheme to compute the test power, which conducted 100 permutations for each experiment while maintaining an overall significance level at α=0.05. The source code at https://github.com/a7b23/H-Divergence [22] was employed.

In addition, more mainstream baselines, which are based on uncertainty quantification and OOD detection, were adjusted to semantic segmentation with an image-level reject option. To produce pixel-wise uncertainty maps using single deterministic segmentation models, the Monte-Carlo Dropout strategy was exploited by adding a dropout layer at the end of each convolution block. Ten feed-forward stochastic passes through the network were run with dropout active. In addition, given their widely known sensitivity to input distribution shifts, post-hoc OOD detectors were also implemented. The aggregated pixel-level uncertainty score (predictive entropy, mutual information, variance) or confidence score (maximum softmax probability (MSP), energy-based [36], soft Dice confidence (SDC) [9]) was used to compute the test statistic for the two-sample hypothesis testing. The KS test was performed on the distribution of uncertainty scores and confidence scores derived from $\mathbb{P}^*$ and $\mathbb{Q}$ for both binary and semantic segmentation tasks.

We also performed experiments that evaluate the effect of different choices in the Segmetron rule. The analysis included study of: (i) the impact of the ensemble size, (ii) the agreement loss function (by comparing the chosen loss functions with the classical cross-entropy) to show why segmentation-specific losses yield superior calibration and maintain in-distribution performance, (iii) the disagreement loss weight λ to justify our choice, (iv) the test sample size to show robustness, and (v) using pseudo labels vs. ground truth for training the EDSs to prove that pseudo-labels suffice.

Finally, we carried out a detailed quantitative comparison on inference times, GPU/CPU and RAM memory usage, as well as the associated energy usage and carbon emissions for Segmetron and each baseline under identical hardware to objectively evaluate deployment feasibility.

All the experiments in this study were conducted on an Intel(R) Core(TM) i9-10900K CPU and an NVIDIA RTX A6000 48 GB GPU. Segmetron and the SOTA methods were compiled in a Python 3.8.5 development environment.[1] All model training associated with Segmetron was performed using the TensorFlow 2.4.0 framework.

[1] https://github.com/tuanaqeelbohoran/Segmetron

Table 1
True positive rates at a 5% significance level for detecting natural covariate shifts for binary and multi-class semantic segmentation tasks from the CMR medical imaging field. Boldface indicates best performance. CMR: Cardiovascular Magnetic Resonance Imaging, TPR@5: True Positive Rate at 5% significance, RMD: Relative Mahalanobis Distance, MMD: Maximum Mean Discrepancy .

Method	TPR@5	
	Binary	Multi-class
Segmetron	**1.0**	**1.0**
RMD	0.95	0.95
MMD	0.24	0.88
H-Divergence	0.46	0.49
Predictive Entropy	0	0.17
Variance	0	0.62
Mutual Information	0	0.59
Maximum Softmax Probability	0	0.11
Energy Score	0	1.0
Soft Dice Confidence	0.89	N/A

3. Results

3.1. Statistical power comparison

Table 1 presents the TPR@5 comparison between Segmetron and the baseline methods for both the binary and multi-class cardiac semantic segmentation tasks. This metric essentially gives how often each method correctly detects covariate shift while maintaining a false alarm rate of only 5%. Out of 100 test cases with actual covariate shift, Segmetron detected 100. This perfect detection rate occurred for both the binary and multi-class semantic segmentation tasks.

Meanwhile, competing methods exhibited varied performance. RMD achieved 95% detection rate missing only 5 out of 100 cases. MMD showed a highly variable behaviour, detecting 24% of cases for binary and 88% of cases for multi-class. H-Divergence missed over half the shifts achieving 46%–49% detection rates. Of note, all methods that are based in uncertainty quantification and OOD detection (except the SDC) achieved a detection rate of 0% for the binary segmentation task, indicating a complete failure to identify covariate shift using only 1 test image. It was found that in order these approaches to achieve TPR@5 = 1, at least 50 test images were required. The results were variable for the multi-class task, ranging from 17% for Predictive Entropy to 100% for the Energy Score. Finally, the SDC confidence score-based method achieved 89% detection rate for the binary task, however, it cannot be applied to the multi-class setting, as it is limited to binary segmentation.

The above performance gap between Segmetron and the baselines can be explained by the fact that the traditional methods (MMD, H-Divergence) try to measure distributional differences directly in the high-dimensional space. Therefore, these methods struggle as dimensionality increase. On the other hand, feature-distance methods (RMD) operate on latent spaces which may not capture the distribution shifts at the pixel space. Last, the failure of traditional uncertainty or OOD-based metrics can be justified by factors such as the MC Dropout unreliability in high-dimensional tasks and the common statistical power decay of uncertainty-based tests as the ambient dimension increases.

3.2. Core finding breakdown

Figs. 3 and 4 provide visual evidence that Segmetron successfully identifies when data has shifted away from the training distribution for the binary cardiovascular semantic segmentation task. These two figures act as a quality control system that can tell when the images it receives are fundamentally different from what it was trained on.

Fig. 3 shows the calibration phase, where Segmetron essentially learns what normal variation looks like. The histogram depicts how many disagreement pixels occurred in each round. The red line marks

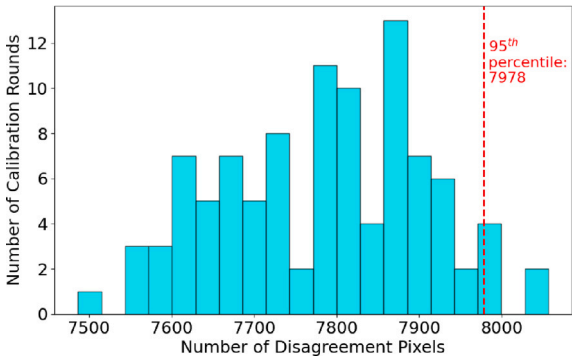


Fig. 3. The Segmetron hypothesis test: Shown is the distribution of the number of calibration rounds as a function of the number of disagreement pixels on $\mathbb{P}^*$. A different random seed for $\mathbb{P}^*$ was used for each calibration round. Also shown in red is the 95th percentile. The plot is taken from the binary cardiovascular semantic segmentation task. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

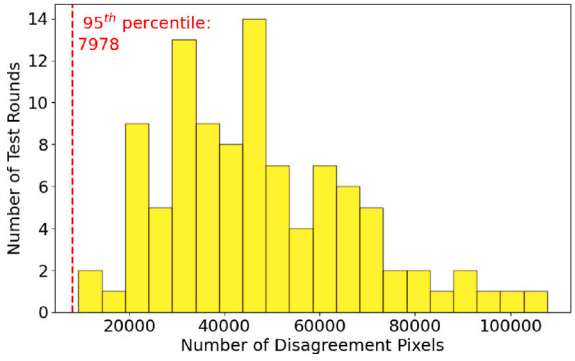


Fig. 4. The Segmetron hypothesis test: Shown is the distribution of the number of test rounds as a function of the number of disagreement pixels on $\mathbb{Q}$. A different random seed for $\mathbb{Q}$ was used for each test round. Also shown in red is the 95th percentile of the 100 calibration rounds on $\mathbb{P}^*$. The plot is taken from the binary cardiovascular semantic segmentation task. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the 95th percentile, meaning only 5% of the time one would expect to see disagreement levels above this threshold by random chance alone. It establishes the baseline of disagreement when there is no shift.

Fig. 4 illustrates the detection capability. 100 different shifted datasets ($\mathbb{Q}$) were tested, that is images with actual covariate shift. The key result is that in all 100 cases, the disagreement pixel count exceeded the 95th percentile from calibration. This means Segmetron correctly identified the shift 100% of the time. The separation between the distributions is clear and substantial.

3.3. Validation of decision choices

We first show the results of the analysis of the ensemble size. Fig. 5 illustrates the Dice (validation) accuracy as the ensemble size increases for the binary semantic segmentation task. Plotted are graphs for both shifted and unshifted datasets from Q. Both curves maintain high Dice accuracy ($\geq$0.95) for the various ensemble sizes. It can be seen that enforcing disagreement training does not accidentally compromise in-distribution performance.

Figs. 6 and 7 depict the relationship between the number of disagreed pixels and the ensemble size for shifted and unshifted data, respectively. These two images actually reveal the mechanism that

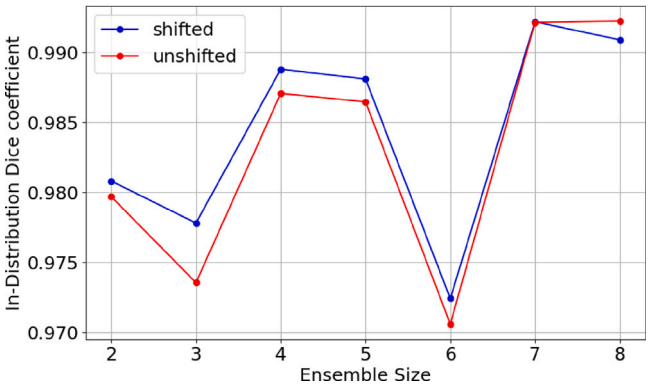


Fig. 5. Plot of the in-distribution Dice coefficient as a function of the ensemble size. Plotted are graphs for both shifted (blue) and unshifted (red) datasets from Q for the binary semantic segmentation task. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Lambda constant sensitivity results.

λ	TPR @5	Setting
1	1	shift
1/101	1	shift
1/1001	1	shift
1/10001	1	shift
1/100001	1	shift

makes Segmetron work. It can be observed that the number of disagreement pixels increases substantially faster in the f_Q curve (from 32,486 to 93,942) as the number of EDSs rises than in the f_P curve (from 6870 to 8292). The rapid increase in 6 indicates that the models are collectively identifying substantial differences in shifted data, as opposed to 7 where the nearly flat trajectory indicates a baseline disagreement level expected from random variation. To sum up, the disagreement gap between shifted and unshifted data widens dramatically with the ensemble size. This divergence creates a powerful signal for covariate shift detection. Apart from studying the ensemble size effect, the findings in these three figures also corroborate our loss function choices (DCE and Focal Tversky) as the learning objective used in the EDS training. In separate experiments conducted concerning the binary semantic segmentation task, it was found that Segmetron's covariate shift detection ability drops dramatically, when binary cross entropy (rather than Focal Tverksky) was used for an EDS to learn to agree. This finding is in line with the need that the EDS training should mirror the base segmenter model training scheme.

Table 2 lists the sensitivity analysis study of the λ as a function of covariate shift detection performance (TPR@5). It was found that Segmetron remarkably maintains its sensitivity (TPR@5=1.0) to covariate shift detection for a wide range of the tuning parameter λ. It was chosen for λ to equal the value $\frac{1}{|\mathbb{Q}|+1}$, to better balance the two distinct objectives and also to be consistent with theoretical literature on the PQ learning setting.

We further observed that, in the binary segmentation task, training the EDSs with ground-truth labels yielded the same covariate shift detection performance (TPR@5=1.0) as training with pseudo-labels produced by the base segmenter, indicating that the pseudo-labels are sufficient. Finally, Segmetron maintained equally strong covariate shift detection performance (TPR@5=1.0) across different test set sizes (1, 2, 5, and 10 images).

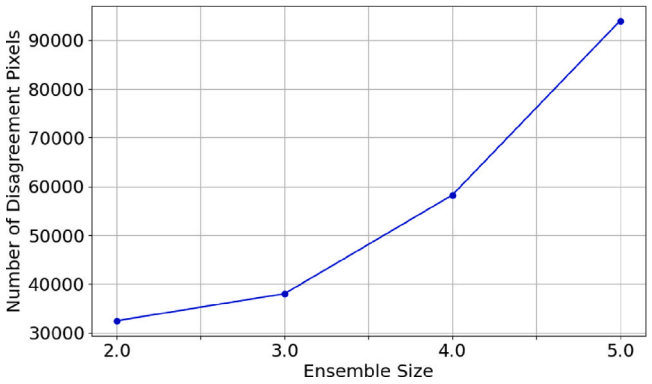


Fig. 6. Plot of the number of disagreement pixels as a function of the ensemble size. Plotted is the graph for shifted data from Q for the binary semantic segmentation task.

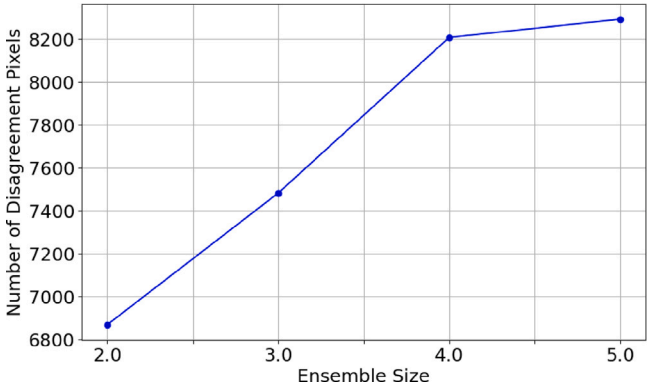


Fig. 7. Plot of the number of disagreement pixels as a function of the ensemble size. Plotted is the graph for unshifted data from Q for the binary semantic segmentation task.

3.4. Computational efficiency considerations

Table 3 reports inference runtime, carbon emissions, energy consumption, and hardware utilisation for all evaluated covariate shift detection methods using one test patient worth data. While Segmetron is not the fastest method in absolute terms, the results highlight a clear trade-off between computational cost and detection reliability, with Segmetron occupying a distinctive and justified position in this spectrum.

Segmetron incurs a higher inference time (27.57 s) compared to lightweight statistical baselines such as MMD or RMD; however, this cost is accompanied by very low GPU (1%) and CPU (2.2%) utilisation, indicating that the method is not hardware-intensive and does not rely on continuous accelerator usage. Importantly, its energy consumption (0.001 kWh) and carbon emissions (0.27 g) remain comparable to, or lower than, many alternative approaches, demonstrating that the increased runtime does not translate into disproportionate environmental cost. In contrast, uncertainty-based methods such as Predictive Entropy, Variance, and Mutual Information exhibit substantially higher inference times (up to 105.31 s) and sustained GPU utilisation ($\approx$ 82%), reflecting the need for multiple stochastic forward passes. These methods are therefore considerably more expensive in practice, both in terms of runtime and hardware load, despite offering weaker or less principled guarantees for covariate shift detection. Segmetron's inference cost is instead dominated by statistical aggregation and hypothesis testing at the image level, rather than repeated model evaluations. This design choice explains its moderate runtime while keeping energy and resource usage low.

4. Discussion

Semantic segmentation enables a higher level of understanding of a depicted scene. It also forms an essential capability towards delivering a plethora of life-changing technologies. However, the covariate shift itself, as well as the lack of trustworthy methods for detecting it, limit the applicability of semantic segmentation. In this study, the image-level covariate shift detection in semantic segmentation was investigated through the lens of statistical hypothesis testing. Segmetron was introduced as a segmenter with a trustworthy reject option that abstains from predicting on test images when semantic segmentation should not be made due to covariate shift. To detect covariate shift, two ensembles of segmenters were enforced to agree on training data and disagree on test data from their set. For training these ensemble models, ways of agreeing (i.e. loss function components) were innovatively put forward that are better suited to the semantic segmentation task and also mirror the base training. The average pixel disagreement rate between each EDS and the base segmenter was chosen as the discriminative test statistic upon which the Segmetron hypothesis test was built.

Importantly, Segmetron is able to deal with unforeseeable covariate shifts of any unknown arbitrary distribution that may occur during deployment. Therefore, this work is valuable because it aligns with "Responsible AI" principles and it happens at a time when the machine learning community is striving to increase society's willingness to accept AI. To obtain strong theoretical guarantees for arbitrary distributions, we extended recent theoretical work on the PQ learning setting of SC to the richer semantic segmentation task [28,29] Also markedly, Segmetron is sample-efficient being able to identify covariate shifts from a single image. Therefore, it could be useful for isolating shifted cases and removing them from automated semantic segmentation pipelines. Last, another favourable characteristic of Segmetron is that it is model-agnostic, since it recognises covariate shifts regardless of the baseline segmentation model that had been used in the initial training.

The ability of Segmetron to detect covariate shift was showcased in two real-world semantic segmentation tasks from the CMR field, involving two (aorta and background) or more (left ventricle, right ventricle, myocardium, background) semantic classes. The shifts in both tasks are malignant (rather than benign) [37], given that they resulted in gross baseline model inaccuracies. Our approach was found to have superior statistical power (TPR@5) to nine SOTA approaches (RMD, H-Divergence, Deep Kernel MMD, Predictive Entropy, Mutual Information, Variance, Maximum Softmax Probability, Energy Score, Soft Dice Confidence) on both tasks. An explanation of these results might be that the kernel-based and uncertainty-based multivariate two-sample tests have been shown to scale badly and their statistical power to decay with the dataset ambient dimension [38]. Moreover, estimating the discrepancy between two probability distributions using the H-divergence has been shown to be prone to underestimations, even if the two distributions are significantly different [39]. The fact that the data of our experiments was 3D is valuable also because established anomaly detection methods in 2D have been shown to not provide reliable performance on 3D data [40].

The validity of our design choices was corroborated by plotting graphs of the relationship between the ensemble size versus number of disagreement pixels and in-distribution accuracy. It was found that the disagreement level manifested a substantially more rapidly increasing trend for shifted than unshifted data from $\mathbb{Q}$. In addition, the EDSs of both ensembles were found to preserve high accuracy on data from the training distribution, also validating the loss function choices. A justification for the choice of the tuning parameter λ was provided. The pseudo-labels produced by the base segmenters were found sufficient for training the EDSs.

Segmetron has higher computational complexity than the SOTA approaches, and is similar in complexity to other ensemble approaches

Table 3
Inference runtime, carbon emissions, energy, and resource utilisation (All Methods).

Method	Inference (s)	Carbon (g)	Energy (kWh)	GPU Avg (%)	CPU Avg (%)	RAM Avg (%)
Segmetron	27.57	0.27	0.001	1	2.2	1.6
RMD	0.50	1.31	0.012	12	33.3	1.6
H-Divergence	2.28	2.71	0.001	7	37.9	1.6
MMD	0.33	1.33	0.006	0	1.9	1.6
Predictive Entropy	60.47	1.19	0.001	82	2.0	1.6
Variance	68.53	1.30	0.001	82	9.1	1.6
Mutual Information	105.31	1.78	0.001	82	2.3	1.6
Maximum Softmax Probability	4.21	0.10	0.001	79	2.3	1.6
Energy Score	7.53	0.13	0.001	46	2.4	1.6
Soft Dice Confidence	17.54	0.25	0.001	25	2.3	1.6

[41]. It achieves favourable energy efficiency and resource utilisation, making it suitable for deployment scenarios where computational sustainability and reliability are both critical [42,43], even if inference latency is moderate. Potential strategies to speed up Segmetron inference time might include pixel subsampling, simplifying the EDS architectures, and more efficient execution through CPU core parallelisation.

In this study, we dealt with the problem of "semantic segmentation with a trustworthy reject option". This problem is closely related to the ''segmentation quality assessment" problem [44,45], since they both make image-level decisions. In addition, given that it was demonstrated that the proposed method is able to identify covariate shift given just a single test dataset, the term could also be simplified to "anomaly detection" [46]. Image enhancement techniques could also have a significant impact on the semantic segmentation results through improving the quality of the images [47,48].

Segmetron offers a straight-forward way to assess the intensity of a covariate shift. The severity ordering could be based on the relative disagreement gap between the in-distribution and out-of-distribution samples. The more harsh the domain drift, the larger the gap between ξ_Q and the 95th percentile of ξ_P will be. Following this line of argument, it was found in our experiments that the covariate shift in the multi-class cardiac semantic segmentation problem was more severe than in the binary one. In a similar manner, the disagreement rate values could also be used to localise (and isolate) the failure regions in an image, a.k.a. "anomaly segmentation" [49].

The PQ learning framework assumes that the samples are independent identically distributed, but such an assumption does not hold for semantic segmentation. However, this has been handled by designing modern DL semantic segmentation architectures that capture spatial correlations and multi-scale representations, and by incorporating appropriate training strategies and loss functions.

The theoretical guarantees of the PQ learning setting, that both the selective segmenter's error on the unknown test distribution $\mathbb{Q}$ and the rejection rate on the training distribution $\mathbb{P}$ are jointly bounded with high probability, rely on the function class having finite VC dimension. In practice, deep neural networks (segmenters) have very high (often exponential in the number of parameters) or even unbounded VC dimension. This is a well-known limitation of classical learning-theoretic analyses. However, this does not invalidate PQ learning. The role of this assumption is to provide insight into the statistical properties of PQ learning under standard finite-capacity models, not to describe modern deep networks exactly. From empirical results, one can safely claim that this does not harm Segmetron's practical usefulness.

In the proposed framework, we aim to detect covariate shift from as few test samples as possible. Therefore, we propose to estimate disagreement on the same patient dataset that had been used to train the ensemble f_Q. While this could result in high variance and low statistical power, we dealt with this issue by estimating the relative increase in disagreement between the EDSs on $\mathbb{Q}$ and $\mathbb{P}$. In this study, and similar to previous research [50], we treated semantic segmentation as a pixel-wise classification problem. Even though such an approach seems to disregard for the structured semantic layout, it was found to lead to a powerful discriminative test statistic for detecting the image-level covariate shift in semantic segmentation. Last, Segmetron detects covariate shift but it does not correct it. Test-time unsupervised domain adaptation methods could come to our rescue in this aspect [51,52], and this will be a topic of future work.

5. Conclusions

In this study, Segmetron, a semantic segmenter with an image-level reject option, was introduced. The proposed hypothesis test is sample-efficient, distribution-free, model-agnostic and provides theoretical guarantees through PQ learning. It was found to outperform nine existing baselines on real-world CMR shifts that arose in both binary and multi-class semantic segmentation tasks. Segmetron treats semantic segmentation as pixel-wise classification which might lead to not capturing contextual cues. The PQ learning assumptions on finite VC dimension and i.i.d. samples may not hold for deep segmenters and real images, respectively. Training EDSs is computationally expensive, and the current evaluation is limited to two CMR datasets. The method detects covariate shift but does not adapt the model to new domains. Future work should focus on exploring test-time unsupervised domain adaptation to correct for detected shifts, validating the method to other imaging modalities and natural-scene datasets, and investigating ways to reduce computational cost. As the pattern recognition community continues to focus on building trust and acceptance of AI in society, Segmetron stands as a critical step forward in enhancing the robustness and reliability of semantic segmentation models. It can potentially enable the widespread adoption of semantic segmentation-based DL technologies across various fields.

CRediT authorship contribution statement

T.A. Bohoran: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **K.S. Parke:** Formal analysis, Data curation. **A. Cowley:** Formal analysis, Data curation. **G.S. Gulsin:** Data curation. **J. Yeo:** Data curation. **A. Dattani:** Data curation. **G.P. McCann:** Writing – review & editing, Supervision, Project administration, Formal analysis, Data curation. **A. Giannakidis:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

[1] J. Zhao, Y. Wu, R. Deng, S. Xu, J. Gao, A. Burke, A survey of autonomous driving from a deep learning perspective, ACM Comput. Surv. (ISSN: 0360-0300) 57 (10) (2025) http://dx.doi.org/10.1145/3729420.

[2] L. Li, K. He, X. Zhu, F. Gou, W. Jia, A pathology image segmentation framework based on deblurring and region proxy in medical decision-making system, Biomed. Signal Process. Control. (ISSN: 1746-8094) 95 (2024) 106439, http://dx.doi.org/10.1016/j.bspc.2024.106439.

[3] K. Hu, Z. Chen, H. Kang, Y. Tang, 3D vision technologies for a self-developed structural external crack damage recognition robot, Autom. Constr. (ISSN: 0926-5805) 159 (2024) 105262, http://dx.doi.org/10.1016/j.autcon.2023.105262.

[4] W. Zhang, H. Wang, P. Ren, W. Zhang, Underwater scene clarity reconstruction via multilayer information fusion and self-organized stitching, IEEE Trans. Circuits Syst. Video Technol. 36 (2) (2026) 1848–1861, http://dx.doi.org/10.1109/TCSVT.2025.3608828.

[5] H. Wang, W. Zhang, Y. Xu, H. Li, P. Ren, WaterCycleDiffusion: Visual–textual fusion empowered underwater image enhancement, Inf. Fusion (ISSN: 1566-2535) 127 (2026) 103693, http://dx.doi.org/10.1016/j.inffus.2025.103693.

[6] P.W. Koh, S. Sagawa, H. Marklund, S.M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R.L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S.M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, P. Liang, WILDS: A benchmark of in-the-wild distribution shifts, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th ICML, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 5637–5664.

[7] S. Kumari, P. Singh, Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives, Comput. Biol. Med. (ISSN: 0010-4825) 170 (2024) 107912, http://dx.doi.org/10.1016/j.compbiomed.2023.107912.

[8] P. de Jorge, R. Volpi, P.H.S. Torr, G. Rogez, Reliability in semantic segmentation: Are we on the right track? in: 2023 IEEE/CVF Conference on CVPR, 2023, pp. 7173–7182, http://dx.doi.org/10.1109/CVPR52729.2023.00693.

[9] B.L.C. Borges, B.M. Pacheco, D. Silva, Selective prediction for semantic segmentation under distribution shift, in: 5th Workshop on Practical ML for Limited/Low Resource Settings, 2024.

[10] Z. Gao, B. Li, M. Salzmann, X. He, Generalize or detect? Towards robust Semantic Segmentation under multiple distribution shifts, in: The Thirty-Eighth Annual Conference on NeurIPS, 2024.

[11] C. Viviers, A. Valiuddin, F. Caetano, L. Abdi, L. Filatova, P. de With, F. van der Sommen, Can your generative model detect out-of-distribution covariate shift? in: Computer Vision – ECCV 2024 Workshops: Milan, Italy, September 29–October 4, 2024, Proceedings, Part XVII, Springer-Verlag, Berlin, Heidelberg, 2025, pp. 184–201, http://dx.doi.org/10.1007/978-3-031-91585-7_12.

[12] C. Dechesne, P. Lassalle, S. Lefèvre, Bayesian U-net: Estimating uncertainty in semantic segmentation of earth observation images, Remote. Sens. (ISSN: 2072-4292) 13 (19) (2021) 3836, http://dx.doi.org/10.3390/rs13193836.

[13] T. Nair, D. Precup, D.L. Arnold, T. Arbel, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, Med. Image Anal. (ISSN: 1361-8415) 59 (2020) 101557, http://dx.doi.org/10.1016/j.media.2019.101557.

[14] K.-C. Kahl, C.T. Lüth, M. Zenk, K. Maier-Hein, P.F. Jaeger, ValUES: A framework for systematic validation of uncertainty estimation in Semantic Segmentation, in: The Twelfth ICLR, 2024.

[15] A. Jungo, F. Balsiger, M. Reyes, Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation, Front. Neurosci. (ISSN: 1662-453X) 14 (2020) 282, http://dx.doi.org/10.3389/fnins.2020.00282.

[16] P. McBee, F. Zulqarnain, S. Syed, D.E. Brown, Image-level uncertainty in pseudo-label selection for semi-supervised segmentation, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, 2022, pp. 4740–4744, http://dx.doi.org/10.1109/EMBC48229.2022.9871359.

[17] L. Gong, Y. Zhang, Y. Zhang, Y. Yang, W. Xu, Erroneous pixel prediction for semantic image segmentation, Comput. Vis. Media 8 (1) (2022) 165–175, http://dx.doi.org/10.1007/s41095-021-0235-7.

[18] J. Lee, G. Lee, T.-Y. Kwak, S.W. Kim, M.-S. Jin, C. Kim, H. Chang, MurSS: A multi-resolution selective segmentation model for breast cancer, Bioengineering (ISSN: 2306-5354) 11 (5) (2024) 463, http://dx.doi.org/10.3390/bioengineering11050463.

[19] Y. Ding, J. Liu, X. Xu, M. Huang, J. Zhuang, J. Xiong, Y. Shi, Uncertainty-aware training of neural networks for selective medical image segmentation, in: Proceedings of the Third Conference on Medical Imaging with Deep Learning, in: Proceedings of Machine Learning Research, vol. 121, PMLR, 2020, pp. 156–173.

[20] J. Küchler, D. Kröll, S. Schoenen, A. Witte, Uncertainty estimates for semantic segmentation: providing enhanced reliability for automated motor claims handling, Mach. Vis. Appl. (ISSN: 1432-1769) 35 (4) (2024) 66, http://dx.doi.org/10.1007/s00138-024-01541-3.

[21] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, D.J. Sutherland, Learning deep kernels for non-parametric two-sample tests, in: H. Daumé III, A. Singh (Eds.), Proceedings of the 37th ICML, in: Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 6316–6326.

[22] S. Zhao, A. Sinha, Y. He, A. Perreault, J. Song, S. Ermon, Comparing distributions by measuring differences that affect decision making, in: ICLR, 2022.

[23] J. Ren, S. Fort, J. Liu, A.G. Roy, S. Padhy, B. Lakshminarayanan, A simple fix to Mahalanobis distance for improving near-OOD detection, in: ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning, 2021.

[24] S. Jang, S. Park, I. Lee, O. Bastani, Sequential covariate shift detection using classifier two-sample tests, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th ICML, in: Proceedings of Machine Learning Research, vol. 162, PMLR, 2022, pp. 9845–9880.

[25] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (ISSN: 1939-3539) 39 (12) (2017) 2481–2495, http://dx.doi.org/10.1109/tpami.2016.2644615.

[26] D. Zhou, Z. Yu, E. Xie, C. Xiao, A. Anandkumar, J. Feng, J.M. Alvarez, Understanding the robustness in vision transformers, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th ICML, in: Proceedings of Machine Learning Research, vol. 162, PMLR, 2022, pp. 27378–27394.

[27] J. Liang, R. He, T. Tan, A comprehensive survey on test-time adaptation under distribution shifts, Int. J. Comput. Vis. (ISSN: 1573-1405) (2024) http://dx.doi.org/10.1007/s11263-024-02181-w.

[28] S. Goldwasser, A.T. Kalai, Y. Kalai, O. Montasser, Beyond perturbations: Learning guarantees with arbitrary adversarial test examples, in: NeurIPS, vol. 33, 2020, pp. 15859–15870.

[29] T. Ginsberg, Z. Liang, R.G. Krishnan, A learning based hypothesis test for harmful covariate shift, in: The Eleventh ICLR, 2023, pp. 1–34.

[30] M.D. Ernst, Permutation methods: A basis for exact inference, Statist. Sci. 19 (4) (2004) 676–685, http://dx.doi.org/10.1214/088342304000000396.

[31] T.A. Bohoran, K.S. Parke, M.P.M. Graham-Brown, M. Meisuria, A. Singh, J. Wormleighton, D. Adlam, D. Gopalan, M.J. Davies, B. Williams, M. Brown, G.P. McCann, A. Giannakidis, Resource efficient aortic distensibility calculation by end to end spatiotemporal learning of aortic lumen from multicentre multivendor multidisease CMR images, Sci. Rep. 13 (1) (2023).

[32] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M.A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V.A. Kollerathu, G. Krishnamurthi, M.-M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K.H. Maier-Hein, P.M. Full, I. Wolf, S. Engelhardt, C.F. Baumgartner, L.M. Koch, J.M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, P.-M. Jodoin, Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. Med. Imaging 37 (11) (2018) 2514–2525, http://dx.doi.org/10.1109/TMI.2018.2837502.

[33] X. Zhuang, Multivariate mixture model for myocardial segmentation combining multi-source images, IEEE Trans. Pattern Anal. Mach. Intell. 41 (12) (2019) 2933–2946, http://dx.doi.org/10.1109/TPAMI.2018.2869576.

[34] C.F. Baumgartner, L.M. Koch, M. Pollefeys, E. Konukoglu, An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation, in: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8, Springer, 2018, pp. 111–119.

[35] J. Massey, The Kolmogorov-Smirnov test for goodness of fit, J. Amer. Statist. Assoc. 46 (253) (1951) 68–78, http://dx.doi.org/10.2307/2280095.

[36] W. Liu, X. Wang, J.D. Owens, Y. Li, Energy-based out-of-distribution detection, in: NeurIPS, vol. 33, 2020, pp. 21464–21475.

[37] S. Rabanser, S. Günnemann, Z. Lipton, Failing loudly: An empirical study of methods for detecting dataset shift, in: NeurIPS, vol. 32, 2019.

[38] A. Ramdas, S.J. Reddi, B. Póczos, A. Singh, L. Wasserman, On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, 2015.

[39] Z. Zhao, L. Cao, R-divergence for estimating model-oriented distribution discrepancy, NeurIPS '23, in: NeurIPS, vol. 36, Curran Associates, Inc., 2023, pp. 56641–56659.

[40] A. Vasiliuk, D. Frolova, M. Belyaev, B. Shirokikh, Limitations of out-of-distribution detection in 3D medical image segmentation, J. Imaging 9 (9) (2023) 191, http://dx.doi.org/10.3390/jimaging9090191.

[41] Q. Zhou, X. Wu, S. Zhang, B. Kang, Z. Ge, L.J. Latecki, Contextual ensemble network for semantic segmentation, Pattern Recognit. (ISSN: 0031-3203) 122 (2022) 108290, http://dx.doi.org/10.1016/j.patcog.2021.108290.

[42] A. Siouras, S. Moustakidis, G. Chalatsis, T.A. Bohoran, M. Hantes, M. Vlychou, S. Tasoulis, A. Giannakidis, D. Tsaopoulos, Economical hybrid novelty detection leveraging global aleatoric semantic uncertainty for enhanced MRI-based ACL tear diagnosis, Comput. Med. Imaging Graph. (ISSN: 0895-6111) 117 (2024) 102424, http://dx.doi.org/10.1016/j.compmedimag.2024.102424.

[43] M.S. Lystbaek, M.J. Beliatis, A. Giannakidis, Low-resource GAN-stack for high-resolution floor plan generation with enhanced evaluation and contextual validation, J. Build. Eng. (ISSN: 2352-7102) 114 (2025) 114211, http://dx.doi.org/10.1016/j.jobe.2025.114211.

[44] S. Chabrier, B. Emile, C. Rosenberger, H. Laurent, Unsupervised performance evaluation of image segmentation, EURASIP J. Adv. Signal Process. 2006 (2006) 1–12, http://dx.doi.org/10.1155/ASP/2006/96306.

[45] B. Specktor-Fadida, L. Ben-Sira, D. Ben-Bashat, L. Joskowicz, SegQC: a segmentation network-based framework for multi-metric segmentation quality control and segmentation error detection in volumetric medical images, Med. Image Anal. (ISSN: 1361-8415) (2025) 103638, http://dx.doi.org/10.1016/j.media.2025.103638.

[46] D. Lee, S. Malacarne, E. Aune, Explainable time series anomaly detection using masked latent generative modeling, Pattern Recognit. (ISSN: 0031-3203) 156 (2024) 110826, http://dx.doi.org/10.1016/j.patcog.2024.110826.

[47] Y. Xu, H. Wang, X. Sun, Q. Xie, W. Zhang, P. Ren, F. Zhou, S. Rahardja, A comb concatenation diffusion model for hyperspectral image super-resolution, Eng. Appl. Artif. Intell. 160 (2025) 111985, http://dx.doi.org/10.1016/j.engappai.2025.111985.

[48] Y. Xu, H. Wang, F. Zhou, C. Luo, X. Sun, S. Rahardja, P. Ren, MambaHSISR: Mamba hyperspectral image super-resolution, IEEE Trans. Geosci. Remote Sens. 63 (2025) 1–16, http://dx.doi.org/10.1109/TGRS.2025.3560632.

[49] Y. Xia, Y. Zhang, F. Liu, W. Shen, A.L. Yuille, Synthesize then compare: Detecting failures and anomalies for semantic segmentation, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, in: Lecture Notes in Computer Science, vol. 12346, Springer, 2020, pp. 145–161, http://dx.doi.org/10.1007/978-3-030-58452-8_9.

[50] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, P. Pérez, Addressing failure prediction by learning model confidence, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), NeurIPS 32, Curran Associates, Inc., 2019, pp. 2902–2913.

[51] Z. Huang, K. Sheng, K. Li, J. Liang, T. Yao, W. Dong, D. Zhou, X. Sun, Reciprocal normalization for domain adaptation, Pattern Recognit. (ISSN: 0031-3203) 140 (2023) 109533, http://dx.doi.org/10.1016/j.patcog.2023.109533.

[52] X. Ye, K.I.-K. Wang, Deep generative domain adaptation with temporal relation attention mechanism for cross-user activity recognition, Pattern Recognit. (ISSN: 0031-3203) 156 (2024) 110811, http://dx.doi.org/10.1016/j.patcog.2024.110811.