# REPEAT: Improving Uncertainty Estimation in Representation Learning Explainability

**Kristoffer K. Wickstrøm**[1*], **Thea Brüsch**[2], **Michael C. Kampffmeyer**[1, 3], **Robert Jenssen**[1, 3, 4]

[1]Department of Physics and Technology, UiT The Arctic University of Norway
[2]Department of Applied Mathematics and Computer Science, Technical University of Denmark
[3]Norwegian Computing Center, Oslo, Norway
[4]Pioneer Centre for AI, University of Copenhagen, Denmark

## Abstract

Incorporating uncertainty is crucial to provide trustworthy explanations of deep learning models. Recent works have demonstrated how uncertainty modeling can be particularly important in the unsupervised field of representation learning explainable artificial intelligence (R-XAI). Current R-XAI methods provide uncertainty by measuring variability in the importance score. However, they fail to provide meaningful estimates of whether a pixel is certainly important or not. In this work, we propose a new R-XAI method called REPEAT that addresses the key question of whether or not a pixel is certainly important. REPEAT leverages the stochasticity of current R-XAI methods to produce multiple estimates of importance, thus considering each pixel in an image as a Bernoulli random variable that is either important or unimportant. From these Bernoulli random variables we can directly estimate the importance of a pixel and its associated certainty, thus enabling users to determine certainty in pixel importance. Our extensive evaluation shows that REPEAT gives certainty estimates that are more intuitive, better at detecting out-of-distribution data, and more concise.

**Code** — https://github.com/Wickstrom/REPEAT/

## Introduction

Representation learning through self-supervision is the cornerstone of recent improvements in the computer vision domain (Caron et al. 2018; He et al. 2022; Assran et al. 2023; Bardes, Ponce, and LeCun 2022). Transforming images into a new representation has been shown to improve performance in a wide range of unsupervised tasks (Trosten et al. 2023; Wickstrøm et al. 2023; Sehwag, Chiang, and Mittal 2021). Despite the benefits, unsupervised representation learning also suffers from some significant drawbacks, particularly a lack of explainability. Current methods in explainable artificial intelligence (XAI) are designed with supervised learning in mind, where a scalar model output is explained in relation to its input (Petsiuk, Das, and Saenko 2018; Bach et al. 2015; Sundararajan, Taly, and Yan 2017). Using these methods to explain *representations* is either not possible or requires major modifications of the underlying algorithms (Crabbé and van der Schaar 2022).

*Corresponding author: kristoffer.k.wickstrom@uit.no


Tackling this drawback has lead to a new direction within XAI, namely representation learning XAI (R-XAI). Methods within R-XAI solve the problem of explaining representations by either making adaptations of existing XAI methods (Crabbé and van der Schaar 2022) or by designing new methods that are particularly designed to tackle the representation learning setting (Wickstrøm et al. 2023; Lin et al. 2023; Bertolini, Clevert, and Montanari 2023; Møller et al. 2024).

A key ingredient in recent R-XAI research is uncertainty estimation (Wickstrøm et al. 2023), where importance is accompanied by a corresponding uncertainty estimate. Providing an indication of certainty is highly desirable, for instance in safety critical areas such as healthcare (Tonekaboni et al. 2019; Kompa, Snoek, and Beam 2021). However, existing frameworks are limited to measuring the variation in the importance scores. This only gives an indication of how the numerical importance scores spread out, not if we are certain of importance. A more critical aspect is *how certain are we that a pixel is important*. Consider an estimated importance map, where all pixels with importance scores higher than 2 are considered important. Now, take one pixel with importance value $5.6 \pm 0.1$ and another with importance value $5.6 \pm 1.2$. Due to the higher variance of the second pixel, current R-XAI methods would assign high uncertainty to this pixel. However, since all values within the 95% confidence interval of the pixel would still be above the importance threshold. As such, we would still be certain that this pixel is important, despite the higher uncertainty of the exact value.

Fig. 1 shows an example that illustrates the distinction between these two questions, where a prior metho is compared to our proposed solution. The example shows which pixels are important for the representation of this image and corresponding uncertainty estimates that indicate how certain the importance is. Both methods mostly agree on the important input pixels, but have vastly different estimates of uncertainty. This shows the effect of modeling certainty in importance, as opposed to variability in importance scores.

With the aim of answering the question of whether we are certain that a pixel is important or not, we present a new R-XAI method called REPEAT. The key idea of REPEAT is to consider each input pixel as a Bernoulli random variable (RV) that indicates if the pixel is important for the representation of the input image. To generate samples of

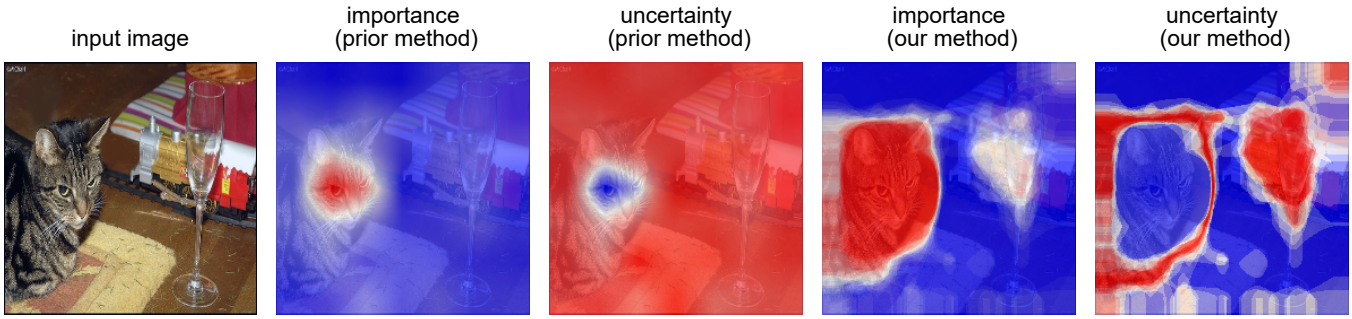


Figure 1: Motivating example to show the difference between current uncertainty estimation techniques in R-XAI and our proposed approach. An image from Pascal-VOC (Everingham et al. 2009) is encoded into a new representation using a ResNet18 feature extractor (He et al. 2016) and important pixels are determined. The importance (red indicates high and blue indicates low importance/uncertainty) is accompanied by uncertainties that specify how confident the importance of a pixel is. The results show that while the importance estimates somewhat agree, the uncertainty estimates are very different, which is due to the different type of uncertainty being captured (variation in pixel importance vs. certainty of importance).

these Bernoulli RVs, we leverage the stochasticity of prior R-XAI methods combined with classic image thresholding techniques to threshold an image into important and non-important pixels. By repeating the thresholding process on numerous importance estimates, a set of Bernoulli samples can be generated to estimate the probability of a pixel being important to the representation of an image and a corresponding uncertainty estimate which indicates how certain we are of the pixel being important or not. Fig. 2 shows an overview of the proposed REPEAT framework. Our contributions are:

1. A new R-XAI framework called REPEAT that models the importance of a pixel for the representation of an image as a Bernoulli RV. This RV indicates the probability of the pixel being important or not and indicates the certainty of the pixel being important.
2. Extensive evaluation across numerous feature extractors and datasets and comparison with state-of-the-art baselines. Results show that REPEAT produces more intuitive uncertainty estimates that are better at detecting out-of-distribution data and has lower complexity, compared to other state-of-the-art methods.
3. Evaluation on a downstream task where uncertainty is used to detect poisoned data in the unsupervised representation learning setting (He, Zha, and Katabi 2023).

## Related Work

**Existing R-XAI literature:** There are two main approaches to extending the field of XAI to handle representations of data; adapt existing methods to handle the representation learning setting or design new methods designed for this particular use case. For adaptation approaches, Crabbé and van der Schaar proposed Label-Free Feature Importance, where an auxiliary scalar function allows standard XAI-methods to be used on each component of the representation (Crabbé and van der Schaar 2022). For R-XAI specific methods, Wickstrøm et al. introduced RELAX, where a representation is explained through similarity measurements between masked and unmasked representations of a particular image (Wickstrøm et al. 2023). Lin et al. extended existing methods in R-XAI to allow for explanation of a reference corpus in relation to a contrastive foil set (Lin et al. 2023). Møller et al. proposed a trainable explanations network aimed at increating the latency of the explanation process (Møller et al. 2024). DeTomaso and Yosef proposed Hotspot, which is focused on explaining representations in single-cell genomics (DeTomaso and Yosef 2021). Lastly, Bertolini, Clevert, and Montanari introduced an aggregation method that generalizes attribution maps between any two convolutional layers of a neural network (Bertolini, Clevert, and Montanari 2023). R-XAI has also been used in many applications, for instance in healthcare (Chen et al. 2023; Wickstrøm et al. 2023; Weinberger, Lin, and Lee 2023) and business (Feng, Li, and Zhang 2023).

**Uncertainty in XAI:** Several approaches have been proposed for modeling uncertainty in XAI. A number of works have investigated how to model uncertainty in surrogate-based XAI (Zhang et al. 2019; Slack et al. 2021; Wang, Zhang, and Lim 2021; Schulz, Santos-Rodriguez, and Poyiadzi 2022), but these approaches are not transferable to the unsupervised setting. Other works have used Monte Carlo Dropout (Gal and Ghahramani 2016) to estimate uncertainty in importance (Wickstrøm, Kampffmeyer, and Jenssen 2020) but this is restrictive because it requires Dropout (Srivastava et al. 2014) in the feature extractor. Another work has used ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) for uncertainty estimation in XAI (Wickstrøm et al. 2021), but this is not directly applicable in this context since we are interested in explaining a single feature extractor. In the general field of uncertainty modeling, test-time-augmentation has been shown to be a generally-applicable and effective tool for uncertainty estimation (Wang et al. 2019; Kahl et al. 2024), but has not been explored in the context of R-XAI. The R-XAI framework RELAX (Wickstrøm et al. 2023) provides uncertainty estimates with its importance scores, but the uncertainty estimates measure the variability in the importance scores and

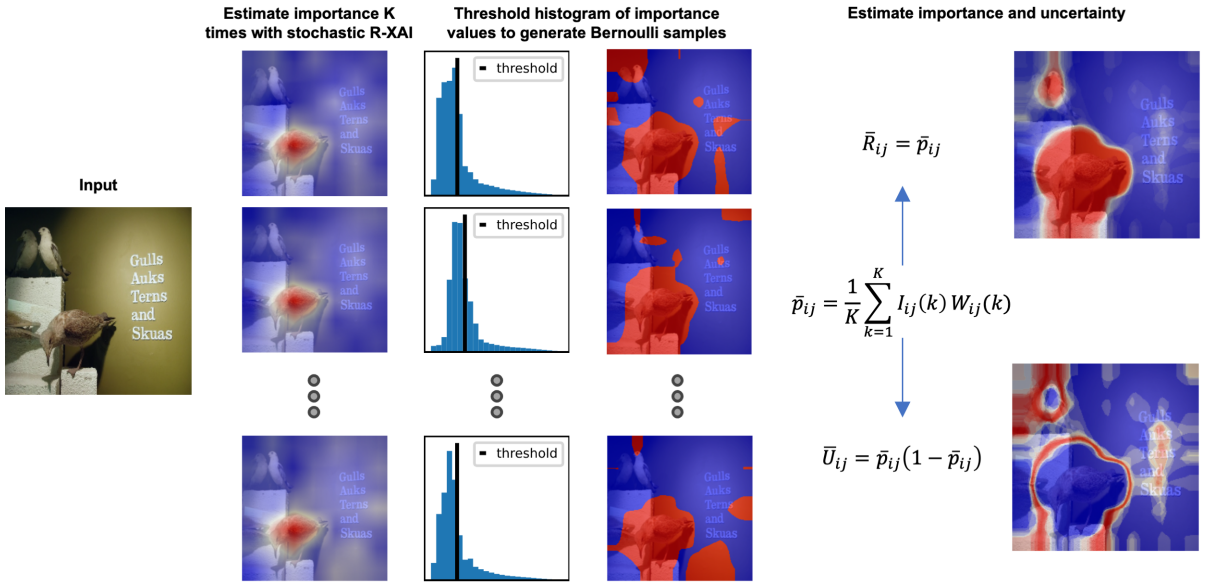


Figure 2: Overview of REPEAT. An image is transformed into a new representation and pixel importance is estimated using stochastic R-XAI. A histogram is constructed from pixel importance and Bernoulli samples are generated by thresholding the importance scores into foreground and background. Then, importance and uncertainty is estimated using the Bernoulli samples.

not certainty in pixel importance.

## REPEAT: A New Method for R-XAI With Improved Uncertainty Estimation

We present REPEAT, a new method for R-XAI that indicates which pixels in an image are most important for the representation of the image and provides uncertainty estimates that specify if a pixel is certainly important or not.

### Interpreting Input Pixels As Bernoulli RVs

Let $X_{ij}$ be a RV following a Bernoulli distribution such that $\Pr(X_{ij} = 1) = p_{ij}$ indicates the probability of pixel $\{i, j\}$ being important for the representation $\mathbf{h}$ of $\mathbf{X}$ by the feature extractor $f$, and $\Pr(X_{ij} = 0) = q_{ij} = (1 - p_{ij})$ indicates the opposite case. We consider the importance of a pixel as:

$$R_{ij} = \mathrm{E}\left[X_{ij}\right] = p_{ij}. \tag{1}$$

Furthermore, we consider the uncertainty associated with the importance as:

$$U_{ij} = \mathrm{Var}\left[X_{ij}\right] = p_{ij}(1 - p_{ij}). \tag{2}$$

The value of $p_{ij}$ is unknown, but can be estimated from data. To perform this estimation, we require realizations of $X_{ij}$. The following subsection presents how to generate these realizations.

### Generating Samples using Stochastic R-XAI and Thresholding

To estimate $p_{ij}$ we require samples that indicate whether or not a pixel is important to the representation of an image. We propose to leverage a base stochastic R-XAI method to generate samples as follows:

$$I_{ij}\left(k\right) = \begin{cases} 1 & \text{if } \bar{R}_{ij}^{\text{base}}(k) \geq \tau \\ 0 & \text{else} \end{cases}. \tag{3}$$

Here, for the $k^{\text{th}}$ realization, $I_{ij}\left(k\right)$ is an indicator function that activates if the importance score is above a certain threshold $\tau$ (see next subsection for threshold selection), and $\bar{R}_{ij}^{\text{base}}(k)$ is the estimated importance score from the base stochastic R-XAI method. If the scores are above the threshold, the pixel is considered important for the representation of the image in question. The stochasticity requirement for the base R-XAI method is critical, since this will ensure diversity and allow new realizations of the RV $X_{ij}$ to be generated. By repeating this process we obtain a set of samples that can be used to estimate $p_{ij}$. The intuition is that pixels which are assigned high importance across numerous realizations will have a high probability of being important. Similarly, pixels that are regularly assigned low scores will have a low probability of being important. And importantly, pixels that fluctuate above and below the threshold will be highlighted as having high uncertainty.

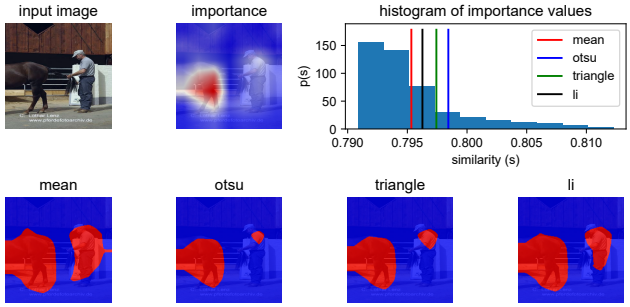


Figure 3: Top row: input image from Pascal VOC (Everingham et al. 2009), importance and histogram of similarities. Bottom row: thresholding of importance.

## Setting the Threshold

Finding a proper value for $\tau$ in Eq. 3 is crucial to generate meaningful samples. We propose to approach this from a foreground-background thresholding perspective, where we consider important pixels as foreground pixels and unimportant pixels as background pixels. Thresholding is a classic problem in image processing with a vast amount of established literature (Gonzalez and Woods 2006). Thresholding algorithms can roughly be separated into two categories; histogram-based or local methods. Histogram-based methods use the histogram of pixel intensities, while local methods process each pixel by considering its neighborhood. Local methods are more computationally demanding, and since we will repeat the thresholding procedure multiple times, we will only focus on histogram-based methods. We consider four widely used approaches to histogram-based image thresholding: mean (Glasbey 1993), Otsu's method (Otsu 1979), triangle method (Zack, Rogers, and Latt 1977), and Li's method (Li and Lee 1993; Li and Tam 1998). In App. A we provide a more detailed explanation of these methods. Fig. 3 displays an example of the thresholding procedure. Here, we feed an image into a ResNet50 encoder (He et al. 2016), and determine importance using an existing stochastic R-XAI method (Wickstrøm et al. 2023). We compute a histogram based on the importance values of all pixels and apply the four thresholding methods to set a threshold for separating foreground (important pixels) from background (unimportant pixels). In this example, the mean thresholding is more conservative and keeps more pixels as important, while Otsu's method produces a higher threshold value that assigns more pixels as unimportant. We found that this behavior was quite consistent across several R-XAI methods and datasets. In App. B, we compared all four methods and their potential in REPEAT, and found that mean thresholding yielded the best performance. Therefore, we use mean thresholding as the standard thresholding method for the remainder of this work.

## Importance and Uncertainty With REPEAT

With repeated use of Eq. 3, new data points can be generated to estimate $p_{ij}$. Assuming $K$ importance maps are generated, we propose to use a weighted sample mean that takes into consideration the base importance scores as follows:

$$\bar{p}_{ij} = \frac{1}{K}\sum_{k=1}^{K} I_{ij}(k)W_{ij}(k) = \bar{R}_{ij}, \tag{4}$$

where

$$W_{ij}(k) = \frac{\bar{R}_{ij}^{\text{base}}(k)}{A(k)} \tag{5}$$

and $A(k)$ is the maximum value of $\bar{\mathbf{R}}^{\text{base}}(k)$. Weighting the indicator function with the importance scores facilitates ranking among the most important pixels in an image, and scaling by the maximum value ensures that the importance scores are comparable across the repeated explanations. In App. A, we also provide an interpretation of REPEAT from the perspective of multiple kernel learning (Gönen and Alpaydin 2011), which shows that REPEAT can be understood as a weighted linear scoring function in a reproducing kernel Hilbert space. Finally, the uncertainty of pixel importance in REPEAT can be calculated in the standard way for Bernoulli RVs:

$$\bar{U}_{ij}^{\text{rtr}} = \bar{p}_{ij}(1 - \bar{p}_{ij}). \tag{6}$$

# Evaluation Protocol and Experimental Setup

Here, we describe how we evaluate and compare REPEAT to state-of-the-art alternatives and detail our experimental setup.

**Evaluation Protocol** We describe how we evaluate and compare uncertainty estimates. Quantitative evaluation is under active development (Hedström et al. 2023), and our evaluation follows the very recent procedures on well known tasks in both the uncertainty and the XAI literature.

**Sanity check:** The Model Parameter Randomisation Test (MPRT) (Adebayo et al. 2018) is widely used in XAI to investigate if XAI method behaves as expected (Barkan et al. 2023a; Lei et al. 2023; Barkan et al. 2023b). The general idea is to see if explanations deteriorate when the parameters of a model are randomized before the decision of the model is explained. However, the MPRT can be highly computationally demanding (Hedström et al. 2024), which makes it difficult to provide a comprehensive analysis. Therefore, we instead use the recently proposed efficient MPRT (eMPRT) (Hedström et al. 2024). The eMPRT compares the relative rise in explanation complexity for an explanation of a trained model compared to a completely randomized model. The intuition is that explanations of random models should be mostly random and therefore have high entropy, while explanations of a trained model should be more focused and therefore have lower entropy. A high positive value is desirable, as it indicates that the explanations of the random model are more complex, while a negative value indicates that the explanations of the trained model are more complex and is not desirable. Originally, both the MPRT and the eMPRT were designed for evaluation of explanations, but here we use eMPRT on the uncertainty estimates.

**Out-of-distribution detection:** A standard task in uncertainty estimation is out-of-distribution (OOD) detec-

| explainability method | resnet50 | vit |
|---|---|---|
| saliency* | -0.16 | -0.10 |
| guided backpropagation* | -0.23 | -0.10 |
| integrated gradients* | -0.17 | -0.19 |
| RELAX | <u>-0.08</u> | -0.07 |
| REPEAT | **0.48** | **0.13** |

| explainability method | resnet50 | vit |
|---|---|---|
| saliency* | -0.16 | -0.11 |
| guided backpropagation* | -0.21 | -0.11 |
| integrated gradients* | -0.18 | -0.19 |
| RELAX | -0.08 | -0.08 |
| REPEAT | **0.44** | **0.13** |

Table 1: Results for R-XAI method for the sanity check (higher is better) of uncertainty estimates on PASCAL-VOC (left) and MS-COCO (right) with a ResNet50 and a ViT encoder. The best and second best performance for each column are indicated by **bold** and <u>underlined</u>, respectively. The * highlights that these methods are adapted to the R-XAI setting using Label-Free Feature Importance (Crabbé and van der Schaar 2022).

tion (Lakshminarayanan, Pritzel, and Blundell 2017; Maddox et al. 2019). When presented with unfamiliar data this should be reflected in the uncertainty. We follow the approach of prior works (Lakshminarayanan, Pritzel, and Blundell 2017; Hein, Andriushchenko, and Bitterwolf 2019) and measure to what degree the uncertainty estimates can be used to differentiate between importance scores for in-distribution and OOD data, based on the uncertainty estimates. Since we are in the unsupervised setting, we propose to detect the OOD data using a Gaussian mixture model (Reynolds 2009) with two components and treat the component with the highest mean as the OOD detector. We choose the component with the highest mean, since we expect the OOD data to have the highest uncertainty.

**Complexity:** In the greater XAI literature, a desirable property of explanations is conciseness, often referred to as low complexity (Chalasani et al. 2020; Bhatt, Weller, and Moura 2021). The same property is also desirable for uncertainty estimates, since a model that is uncertain about all importance scores is not informative. Instead, the model should be confident about clearly important and unimportant pixels, and only uncertain about some critical pixels where there is ambiguity. In this work, we measure complexity following the standard approach proposed by Bhatt *et al.* (Bhatt, Weller, and Moura 2021), where complexity is calculated by taking the entropy of the uncertainties for an image.

**Experimental Setup** We investigate the performance of REPEAT and competing methods across numerous feature extractors, datasets, and baselines that are described below.

**REPEAT design choices:** In all presented results, we generate K=10 realizations of the Bernoulli RVs and use the mean to perform the thresholding. Both of these choices are determined by quantitative evaluation that is reported in App. B. As the base stochastic R-XAI method we use RELAX (Wickstrøm et al. 2023), due to its high performance in recent works. However, to demonstrate REPEAT's ability to leverage any stochastic R-XAI method, we also evaluated the performance of REPEAT with Kernel SHAP (Lundberg and Lee 2017), which we show in the results. We also reduce the computational demand of RELAX by developing a new bound on the estimation error in RELAX (see App. A). In App. B, we also evaluate the time complexity of REPEAT and competing methods.

**Datasets:** We use four widely used computer vision datasets; MS-COCO (Lin et al. 2014), Pascal-VOC (Everingham et al. 2009), EuroSAT (Helber et al. 2018), and FashionMNIST (Xiao, Rasul, and Vollgraf 2017).

**Baseline XAI methods:** We compare the performance of REPEAT with several strong baselines. First, RELAX (Wickstrøm et al. 2023), which is designed for the R-XAI setting. Apart from RELAX, all other methods are adopted to the R-XAI setting using Label-Free Feature Importance (Crabbé and van der Schaar 2022). These baselines are: Saliency (Morch et al. 1995), Guided-Backpropagation (Springenberg et al. 2015), and Integrated Gradients (Sundararajan, Taly, and Yan 2017).

**Feature extractors:** We leverage two state-of-the-art feature extractors to create the representations we want to explain; the ResNet50 (He et al. 2016) and the Vision Transformer (ViT) (Dosovitskiy et al. 2021). For the ResNet50, we take the representation to be the output of the adaptive pooling layer at the end of convolutional neural network backbone. For the ViT, we use the base model and treat the classification token as the representation. For simplicity and reproducibility, we use the pretrained weights from Pytorch (Paszke et al. 2019) for supervised classification of ImageNet (Deng et al. 2009).

**Baseline uncertainty estimation:** REPEAT and RELAX provide uncertainty estimates as part of their framework. The remaining baseline methods do not have this capability, and external methods must be used to estimate uncertainty in their importance scores. Due to its flexibility and performance (Kahl et al. 2024), we propose to use test-time augmentation (Abdar et al. 2021) to estimate uncertainty for the remaining baseline methods. Specifically, we follow Wang *et al.* (Wang et al. 2019), where Dropout is applied to the input (Dropout probability of 0.5). Here, we create 10 Dropout-versions of each image and calculate importance using the baseline methods. Uncertainty is computed by taking the standard deviation across all 10 importance maps.

## Results

This section present the main results of our work. First, we present the results of the outlined evaluation protocol. Then, we show that REPEAT is also applicable beyond RELAX. In all experiments, we randomly sample 1000 images from the dataset used for evaluation. We found that this was enough samples to provide reliable estimates of performance while still being computationally tractable. Due to their inherent

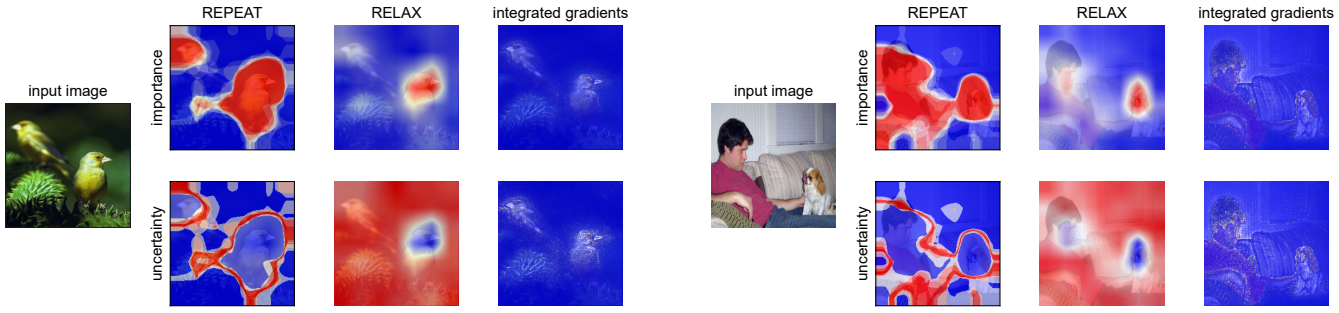


Figure 4: Qualitative examples on images from PASCAL VOC.

| explainability method | resnet50 | vit |
|---|---|---|
| saliency* | 0.797 | <u>0.002</u> |
| guided backpropagation* | 0.782 | 0.001 |
| integrated gradients* | <u>0.998</u> | <u>0.002</u> |
| RELAX | 0.000 | 0.000 |
| REPEAT | **1.000** | **0.973** |

| explainability method | resnet50 | vit |
|---|---|---|
| saliency* | 0.969 | 0.001 |
| guided backpropagation* | 0.823 | 0.000 |
| integrated gradients* | <u>0.999</u> | <u>0.002</u> |
| RELAX | 0.000 | 0.000 |
| REPEAT | **1.000** | **0.939** |

Table 2: Results for R-XAI methods for OOD detection using EuroSAT as the OOD dataset. The table shows AUROC when classifying in-domain (PASCAL-VOC (left) or MS-COCO (right)) vs out-of-domain clusters using a Gaussian mixture model (higher is better). The best and second best performance for each column are indicated by **bold** and <u>underlined</u>, respectively. The * highlights that these methods are adapted to the R-XAI setting using Label-Free Feature Importance (Crabbé and van der Schaar 2022).

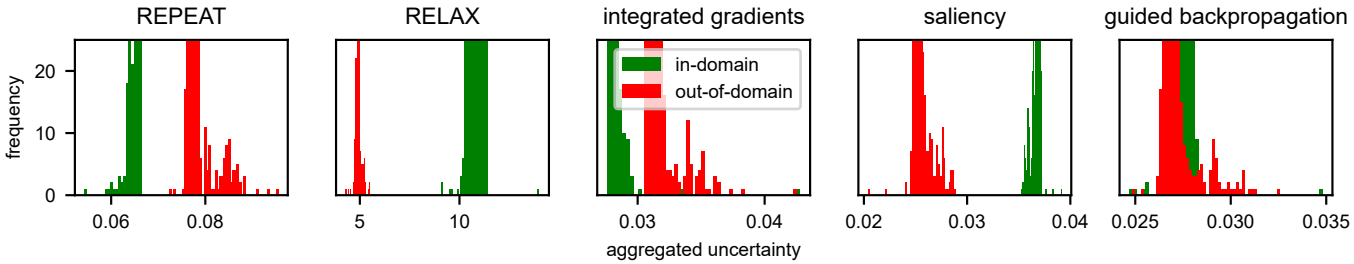


Figure 5: Histogram of aggregated uncertainty scores for in-distribution (PASCAL-VOC in green) and OOD (EuroSAT in red) data. This example illustrates how REPEAT gives the clearest separation between in-distribution and OOD data under the assumption that OOD data should have highest uncertainty.

stochasticity, RELAX and REPEAT experiments were repeated 3 times. This revealed (shown in App. B) that performance was stable with a standard deviation less than the decimal precision reported here. Therefore, for clarity we do not report the standard deviation here. Also, in App. B we evaluate the performance of the explanations produced by REPEAT, which shows good performance, and provide several qualitative examples.

## Quantitative Evaluation

**Sanity check:** Tab. 1 shows the results for the sanity check. The results show that REPEAT outperforms all baselines methods across all datasets and encoders. Interestingly, REPEAT is the only method that provides uncertainty estimates that pass the sanity check. This highlights that the proposed method brings a significant advantage compared to competing methods. Lastly, Fig. 4 shows some qualitative examples, and more can be found in App. B.

**OOD detection:** Tab. 2 shows the results of the OOD experiment with Eurosat as the OOD data. A similar experiment is shown in App. B with FashionMNIST as the OOD data. The results show how REPEAT clearly outperforms all other methods. Particularly RELAX has low performance, because it gives lower uncertainty scores to the OOD data, and thus falsely classifies in-domain data as OOD. Fig. 5 displays an example of the distribution of aggregated uncertainties for all methods. Note how the in-distribution and OOD data are clearly separable, with the uncertainty for the OOD data being much larger than the in-distribution data. In contrast, both RELAX and Saliency has lower uncertainty

| explainability method | resnet50 | vit |
|---|---|---|
| saliency* | 10.52 | <u>10.46</u> |
| guided backpropagation* | <u>10.41</u> | <u>10.46</u> |
| integrated gradients* | 10.44 | 10.49 |
| RELAX | 10.82 | 10.82 |
| REPEAT | **9.97** | **9.99** |

| explainability method | resnet50 | vit |
|---|---|---|
| saliency* | 10.45 | <u>10.40</u> |
| guided backpropagation* | <u>10.35</u> | <u>10.40</u> |
| integrated gradients* | 10.37 | 10.42 |
| RELAX | 10.75 | 10.74 |
| REPEAT | **9.91** | **9.94** |

Table 3: Results for R-XAI method for complexity (lower is better) of uncertainty estimates on PASCAL-VOC (left) and MS-COCO (right) with a ResNet50 and a ViT encoder. The best and second best performance for each column are indicated by **bold** and <u>underlined</u>, respectively. The * highlights that these methods are adapted to the R-XAI setting using Label-Free Feature Importance (Crabbé and van der Schaar 2022).

| explainability method | resnet50 | vit |
|---|---|---|
| REPEAT (Kernel SHAP) | <u>0.999</u> | **0.999** |
| REPEAT (RELAX) | **1.000** | <u>0.973</u> |

| explainability method | resnet50 | vit |
|---|---|---|
| REPEAT (Kernel SHAP) | **1.000** | **1.000** |
| REPEAT (RELAX) | **1.000** | <u>0.939</u> |

Table 4: Results for R-XAI methods for OOD detection using the EuroSAT dataset as the OOD dataset, with different base R-XAI methods in REPEAT. The table shows AUROC when classifying in-domain (VOC or COCO) vs out-of-domain clusters using a Gaussian mixture model. The best and second best performance for each column are indicated by **bold** and <u>underlined</u>, respectively.

| explainability method | resnet50 | vit |
|---|---|---|
| REPEAT (Kernel SHAP) | <u>10.40</u> | <u>10.36</u> |
| REPEAT (RELAX) | **9.97** | **9.99** |

| explainability method | resnet50 | vit |
|---|---|---|
| REPEAT (Kernel SHAP) | <u>10.28</u> | <u>10.28</u> |
| REPEAT (RELAX) | **9.91** | **9.94** |

Table 5: Results for R-XAI method for complexity (lower is better) of uncertainty estimates for different base R-XAI methods in REPEAT on PASCAL-VOC (left) and MS-COCO (right) with a ResNet50 and a ViT encoder. The best and second best performance for each column are indicated by **bold** and <u>underlined</u>, respectively.

for the OOD data, which is the complete opposite of the desired behavior. For Guided Backpropagation, the two distributions are indistinguishable. Integrated Gradients is adequate at separating the two distributions, but has much less separation compared to REPEAT. Motivated by REPEAT's successful OOD detection abilities, we also conducted experiments on poisoned data (Goldblum et al. 2023) with encouraging performance. These results can be seen in App. B, and show that uncertainty is essential to obtain good performance on this downstream task.

**Complexity evaluation:** Tab. 3 displays the result of complexity evaluation of the different uncertainty estimates, which shows that REPEAT outperforms all method across all settings, thus providing more concise uncertainty estimates compared to existing methods.

**REPEAT beyond RELAX:** In this work, we have focused on RELAX as the base R-XAI method in REPEAT. However, REPEAT is more general and can be used with any stochastic R-XAI method. To illustrate this, we have conducted the same OOD detection and complexity experiments as earlier but with Kernel-SHAP (Lundberg and Lee 2017) (adapted to the R-XAI setting using Label-Free Feature Importance (Crabbé and van der Schaar 2022)) as the base stochastic R-XAI method. These results are shown in Tab. 4 and Tab. 5, and demonstrate the flexibility of REPEAT. The performance for OOD detection is very similar for RELAX compared to Kernel-SHAP, but RELAX gives lower complexity compared to using Kernel-SHAP as the base R-XAI method.

## Conclusion

Current methods for determining certainty in pixel importance are limited, as they only estimate the variability in the importance values. This reduces the reliability of R-XAI, as users cannot decide if a pixel is certainly important or not. In this work, we proposed a new method called REPEAT that addresses this limitation. REPEAT treats each pixel in an image as a Bernoulli RV that is either important or unimportant to the representation of the image. From these Bernoulli RV we can directly estimate the importance of a pixel and its associated certainty, thus enabling users to ascertain certainty in pixel importance. We conducted an extensive evaluation which showed that REPEAT provides more intuitive uncertainty estimates that are better at identifying OOD data and with lower complexity. Further, we also show that REPEAT works effectively with different types of stochastic R-XAI methods. We believe REPEAT can play an important role in moving the field of R-XAI forward.

## Acknowledgements

This work was financially supported by the Research Council of Norway (RCN), through its Centre for Research-based Innovation funding scheme (Visual Intelligence, grant no. 309439), and Consortium Partners. It was further funded by RCN FRIPRO grant no. 315029 and RCN IKTPLUSS grant no. 303514.

## References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; Makarenkov, V.; and Nahavandi, S. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76: 243–297.

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 9525–9536. Red Hook, NY, USA: Curran Associates Inc.

Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; and Ballas, N. 2023. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7): e0130140.

Bardes, A.; Ponce, J.; and LeCun, Y. 2022. VICRegL: Self-Supervised Learning of Local Visual Features. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, 8799–8810. Curran Associates, Inc.

Barkan, O.; Elisha, Y.; Asher, Y.; Eshel, A.; and Koenigstein, N. 2023a. Visual Explanations via Iterated Integrated Attributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2073–2084.

Barkan, O.; Elisha, Y.; Weill, J.; Asher, Y.; Eshel, A.; and Koenigstein, N. 2023b. Deep Integrated Explanations. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*.

Bertolini, M.; Clevert, D.-A.; and Montanari, F. 2023. Explaining, Evaluating and Enhancing Neural Networks' Learned Representations. In *Artificial Neural Networks and Machine Learning – ICANN 2023: 32nd International Conference on Artificial Neural Networks, Heraklion, Crete, Greece, September 26–29, 2023, Proceedings, Part V*, 269–287. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-44191-2.

Bhatt, U.; Weller, A.; and Moura, J. M. F. 2021. Evaluating and aggregating feature-based model explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20. ISBN 9780999241165.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep Clustering for Unsupervised Learning of Visual Features. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Chalasani, P.; Chen, J.; Chowdhury, A. R.; Wu, X.; and Jha, S. 2020. Concise Explanations of Neural Networks using Adversarial Training. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1383–1391. PMLR.

Chen, Y.; Bijlani, N.; Kouchaki, S.; and Barnaghi, P. 2023. Interpreting Differentiable Latent States for Healthcare Time-series Data. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Crabbé, J.; and van der Schaar, M. 2022. Label-Free Explainability for Unsupervised Models. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 4391–4420. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Computer Vision and Pattern Recognition*, 248–255. Ieee.

DeTomaso, D.; and Yosef, N. 2021. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell Systems*, 12: 446–456.e9.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Everingham, M.; Gool, L. V.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2009. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 303–338.

Feng, X. F.; Li, C.; and Zhang, S. 2023. Visual Uniqueness in Peer-to-Peer Marketplaces: Machine Learning Model Development, Validation, and Application. *SSRN Electronic Journal*.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Balcan, M. F.; and Weinberger, K. Q., eds., *Proceedings of The 33rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 1050–1059. New York, New York, USA: PMLR.

Glasbey, C. 1993. An Analysis of Histogram-Based Thresholding Algorithms. *CVGIP: Graphical Models and Image Processing*, 55(6): 532–537.

Goldblum, M.; Tsipras, D.; Xie, C.; Chen, X.; Schwarzschild, A.; Song, D.; Madry, A.; Li, B.; and Goldstein, T. 2023. Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(02): 1563–1580.

Gönen, M.; and Alpaydin, E. 2011. Multiple Kernel Learning Algorithms. *Journal of Machine Learning Research*, 12(64): 2211–2268.

Gonzalez, R. C.; and Woods, R. E. 2006. *Digital Image Processing (3rd Edition)*. USA: Prentice-Hall, Inc. ISBN 013168728X.

He, H.; Zha, K.; and Katabi, D. 2023. Indiscriminate Poisoning Attacks on Unsupervised Contrastive Learning. In *The Eleventh International Conference on Learning Representations*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 CVPR*, 770–778.

Hedström, A.; Bommer, P. L.; Wickstrøm, K. K.; Samek, W.; Lapuschkin, S.; and Höhne, M. M. 2023. The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus. *Transactions on Machine Learning Research*.

Hedström, A.; Weber, L.; Lapuschkin, S.; and Höhne, M. 2024. A Fresh Look at Sanity Checks for Saliency Maps. In *Explainable Artificial Intelligence*, 403–420. Cham: Springer Nature Switzerland.

Hein, M.; Andriushchenko, M.; and Bitterwolf, J. 2019. Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 41–50. Los Alamitos, CA, USA: IEEE Computer Society.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2018. Introducing EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, 204–207. IEEE.

Kahl, K.-C.; Lüth, C. T.; Zenk, M.; Maier-Hein, K.; and Jaeger, P. F. 2024. ValUES: A Framework for Systematic Validation of Uncertainty Estimation in Semantic Segmentation. In *The Twelfth International Conference on Learning Representations*.

Kompa, B.; Snoek, J.; and Beam, A. L. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6405–6416. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Lei, Y.; Li, Z.; Li, Y.; Zhang, J.; and Shan, H. 2023. LICO: Explainable Models with Language-Image COnsistency. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 61870–61887. Curran Associates, Inc.

Li, C.; and Lee, C. 1993. Minimum cross entropy thresholding. *Pattern Recognition*, 26(4): 617–625.

Li, C.; and Tam, P. 1998. An iterative algorithm for minimum cross entropy thresholding. *Pattern Recognition Letters*, 19(8): 771–776.

Lin, C.; Chen, H.; Kim, C.; and Lee, S.-I. 2023. Contrastive Corpus Attribution for Explaining Representations. In *The Eleventh International Conference on Learning Representations*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, 740–755. Springer International Publishing.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4768–4777. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Maddox, W. J.; Izmailov, P.; Garipov, T.; Vetrov, D. P.; and Wilson, A. G. 2019. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Møller, B. L.; Igel, C.; Wickstrøm, K. K.; Sporring, J.; Jenssen, R.; and Ibragimov, B. 2024. Finding NEM-U: Explaining unsupervised representation learning through neural network generated explanation masks. In *Forty-first International Conference on Machine Learning*.

Morch, N.; et al. 1995. Visualization of neural networks using saliency maps. In *International Conference on Neural Networks*, 2085–2090.

Otsu, N. 1979. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1): 62–66.

Paszke, A.; Gross, S.; Massa, F.; et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, 8024–8035.

Petsiuk, V.; Das, A.; and Saenko, K. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Reynolds, D. 2009. *Gaussian Mixture Models*, 659–663. Boston, MA: Springer US. ISBN 978-0-387-73003-5.

Schulz, J.; Santos-Rodriguez, R.; and Poyiadzi, R. 2022. Uncertainty Quantification of Surrogate Explanations: an Ordinal Consensus Approach. *Proceedings of the Northern Lights Deep Learning Workshop*, 3.

Sehwag, V.; Chiang, M.; and Mittal, P. 2021. SSD: A Unified Framework for Self-Supervised Outlier Detection. In *International Conference on Learning Representations*.

Slack, D.; Hilgard, A.; Singh, S.; and Lakkaraju, H. 2021. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 9391–9404. Curran Associates, Inc.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2015. Striving for Simplicity: The All Convolutional Net. In *ICLR Workshop*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 3319–3328. PMLR.

Tonekaboni, S.; Joshi, S.; McCradden, M. D.; and Goldenberg, A. 2019. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106, 359–380.

Trosten, D. J.; Løkse, S.; Jenssen, R.; and Kampffmeyer, M. C. 2023. On the Effects of Self-Supervision and Contrastive Alignment in Deep Multi-View Clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23976–23985.

Wang, D.; Zhang, W.; and Lim, B. Y. 2021. Show or suppress? Managing input uncertainty in machine learning model explanations. *Artificial Intelligence*, 294: 103456.

Wang, G.; Li, W.; Aertsen, M.; Deprest, J.; Ourselin, S.; and Vercauteren, T. 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338: 34–45.

Weinberger, E.; Lin, C.; and Lee, S.-I. 2023. Isolating salient variations of interest in single-cell data with contrastiveVI. *Nature Methods*, 20(9): 1336–1345.

Wickstrøm, K.; Kampffmeyer, M.; and Jenssen, R. 2020. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical Image Analysis*, 60: 101619.

Wickstrøm, K.; Mikalsen, K. Ø.; Kampffmeyer, M.; Revhaug, A.; and Jenssen, R. 2021. Uncertainty-Aware Deep Ensembles for Reliable and Explainable Predictions of Clinical Time Series. *IEEE Journal of Biomedical and Health Informatics*, 25(7): 2435–2444.

Wickstrøm, K. K.; Trosten, D. J.; Løkse, S.; Boubekki, A.; Mikalsen, K. Ø.; Kampffmeyer, M. C.; and Jenssen, R. 2023. RELAX: Representation Learning Explainability. *Int. J. Comput. Vis.*, 1584–1610.

Wickstrøm, K. K.; Østmo, E. A.; Radiya, K.; Øyvind Mikalsen, K.; Kampffmeyer, M. C.; and Jenssen, R. 2023. A clinically motivated self-supervised approach for content-based image retrieval of CT liver images. *Computerized Medical Imaging and Graphics*, 102239.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR*, abs/1708.07747.

Zack, G. W.; Rogers, W. E.; and Latt, S. A. 1977. Automatic measurement of sister chromatid exchange frequency. *Journal of Histochemistry and amp; Cytochemistry*, 25(7): 741–753.

Zhang, Y.; Song, K.; Sun, Y.; Tan, S.; and Udell, M. 2019. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. In *Workshop on AI for Social Good*.