Domain adapted probabilistic inspection using deep probabilistic segmentation

Andersen, Rasmus Eckholdt; Boukas, Evangelos

Published in:
Ocean Engineering

Link to article, DOI:
10.1016/j.oceaneng.2022.113568

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Ocean Engineering 270 (2023) 113568

Contents lists available at ScienceDirect

Ocean Engineering

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

Domain adapted probabilistic inspection using deep probabilistic segmentation☆

Rasmus Eckholdt Andersen *, Evangelos Boukas

Department of Electrical and Photonics Engineering, Technical University of Denmark, Kgs. Lyngby, 2800, Denmark

ARTICLE INFO

Keywords:
Domain adaptation
Probabilistic segmentation
Marine vessels

ABSTRACT

This paper introduces the concept of domain-adapted probabilistic segmentation for marine vessel classification. The evolution of corrosion is continuous and it is, therefore, impossible to acquire marine vessel inspection datasets representative of the entire active fleet. Additionally, human surveyors introduce high levels of subjectiveness in the classification process, resulting in potentially multiple equally valid but ambiguous classification results. Consequently, deterministic inspection is flawed. The goal of this paper is to address these challenges by using a probabilistic approach to segmentation while performing domain adaptation to align the feature space across the different stages of age degradation. We test a Probabilistic U-Net on both simulated images and images from real vessels and compare it against two novel probabilistic models. We have evaluated the models using both quantitative — energy distance as distribution similarity — and qualitative — feature reduction visualization — approaches. Our results indicate that the combination of probabilistic segmentation and domain adaption could potentially have a high impact on marine vessel surveys in the future.

1. Introduction

Modern civil structures are continuously pushing the performance limits of construction materials due to increasing focus on environmental aspects, aesthetics, or due to external natural forces. These types of structures are usually long-term investments that have to be maintained to avoid natural degradation in the form of wear and tear, corrosion, and elements of nature, such as earthquakes, hurricanes, floods, etc. Thus, inspection is a critical aspect of the maintenance which provides an insight into the structural integrity and pinpoints where repairs should be done — ideally before irreversible damage has occurred.

While there already exist some works on using modern computer vision and deep learning to locate and classify faults and defects in many types of civil structures such as power lines (Nguyen et al., 2018), wind turbines (Liu et al., 2022), marine vessels (Bonnin-Pascual and Ortiz, 2019), and bridges (Abdallah et al., 2022), they typically rely on training data with a single ground truth. As a consequence, the output of the model is often deterministic and, therefore, fails to capture the inherent variability between surveyors performing a predominately visual assessment. Ideally, the output of the inspection process should be a distribution of assessments that captures the variability between surveyors.

1.1. Marine vessel inspection

A field where the variability between surveyors is prominent is the inspection process in the marine vessel industry. To ensure the structural integrity of the global marine vessel fleet, each vessel has to undergo periodic inspections. Ballast tanks are example areas that have to undergo inspections as corrosion, buckling, and coating breakdown are a direct risk to their structural integrity. The inspection process currently involves a human surveyor physically traversing the hard-to-access ballast tanks and subjectively assessing the condition of the structure and surfaces, based on loose definitions and experience. The faults and defects located by the surveyor are noted down for later repairs. Since the assessment of the severity of the defects is entirely up to the surveyor, there exists a high level of subjectivity which makes it difficult to reproduce the results of inspections between different surveyors. Even the same surveyor may produce different results depending on the time of day (Rizzo, 2008).

The current inspection process is divided into three main surveys; Overall Survey, Close-Up Survey, and Non-Destructive Testing. The overall survey gives an indication of the general condition and determines whether the structural integrity of the vessel is sufficient to

☆ This work has been funded by the Innovation Fund Denmark (IFD), through the Inspectrone (Autonomous and high-level commanded system for remote inspection of marine vessels to support classification and commercial operations) project, under contract number 8090-00080B.

* Corresponding author.
E-mail address: recan@dtu.dk (R.E. Andersen).

https://doi.org/10.1016/j.oceaneng.2022.113568
Received 3 September 2022; Received in revised form 30 November 2022; Accepted 26 December 2022
Available online 20 January 2023

Fig. 1. Two vessels of the same type, but with different age. On the top is a younger vessel and on the bottom is an older vessel. Note the difference in color as the coating degrades; from gray-toned to yellow-toned. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

allow the vessel back into service. One of the metrics used to evaluate vessels is the amount of area covered in corrosion which is discretized into three categories: Good (less than 3% coating breakdown), Fair (between 3% and 20% coating breakdown), and Poor (more than 20% coating breakdown). It is up to the surveyor to estimate which category an area under inspection belongs to. The purpose of the close-up survey is to identify and localize critical defects for repairs or further investigation using non-destructive testing.

Since the coating texture of a marine vessel ballast tank changes drastically over the course of a vessel's lifetime, it is extremely difficult to collect the amount of data required to represent all states a vessel can be in. Fig. 1 shows two vessels of the same type with different age. It is clearly visible that not only is there expectedly more corrosion visible, but the color of the surface has also changed from a gray color-tone to a yellow color-tone.

One of the biggest challenges when applying deep learning in any new field is the need for a large amount of data. This is partly due to the networks being highly sensitive to small perturbations in the input image (Su et al., 2019). Due to the difficulty of accessing real-life data in a high enough quantity, we have developed a simulation model that can procedurally generate an infinite amount of images with the three types of corrosion using the open source 3D computer graphics software toolset Blender (Hess, 2007). Using this simulation we have generated a set of images of an old vessel. For each image in this dataset, we have five independently annotated ground truths which we use as expert opinions. Examples of ground truth data are shown in Fig. 2. Since the goal of this work is to be deployed on a drone traversing ballast tanks, we have exported the simulation model to the Unreal Engine-based simulation framework AirSim (Shah et al., 2017) which can generate photo-realistic (albeit less than Blender) imagery. Additionally, we have a limited set of real-life images. We employ domain adaptation between different combinations of these datasets to verify that the model is able to generalize an expert distribution across domains.

1.2. Multi-expert GT's with probabilistic models

The optimization of a variational auto-encoder includes the joint optimization of an encoder and decoder to minimize a reconstruction loss, such as the Kullback–Leibler divergence, between a prior and posterior distribution. The prior is estimated through network parameters while the posterior is either a fixed known distribution or, if conditioned on the ground truth, estimated from another set of network parameters. The ability to map an input to a multivariate distribution means that it formulates a distribution for each latent attribute instead of a single-point estimate. Consequently, the model can be sampled to produce multiple unique outputs for a given input. For multiple ground truths in image segmentation, the posterior can be estimated using multiple samples and the prior can be estimated from the input image.

The purpose of this paper is to use a variety of different variational auto-encoders and probabilistic models to capture the variability of the human expert surveyor opinions when segmenting a given input image, while simultaneously applying domain adaptation. We introduce the notion of probabilistic inspection, which is defined as the ability to estimate an expert-provided likelihood of defect instances in images, by inferring a distribution of solutions for the otherwise ambiguous task of corrosion segmentation. The concept is illustrated in Fig. 2 where five different, but individually equally valid segmentations of corrosion show how the task is naturally ambiguous. By using domain adaptation, the network will narrow the feature space to only the features available in both the newer and older vessels.

The main contributions of this work are:

1. Introduce the notion of probabilistic inspection in maritime classification procedures.
2. Introduce two new variations of probabilistic models that do not rely on fixed distribution parameters.
3. Introduce domain adaptation to multiple probabilistic models.
4. Introduce a metric for deciding the optimal number of samples for the expected output of the model.

The remainder of this paper is structured as follows: Section 2 gives an overview of the current state of the art. Section 3 recaps the Probabilistic U-Net architecture. Section 4 introduces the datasets on which we train and test all our models. Section 5 describes the training and evaluation metrics we use to evaluate the performance of our models. Section 6 shows the results of all our experiments.

2. Related work

The use of computer vision techniques has already been explored to some extent in the context of marine vessel inspection. Some of the early works utilized more classic computer vision techniques such as saliency to compute a topographic map that is fed to a contrast-based detector and a combination of logical operators (Bonnin-Pascual and Ortiz, 2014, 2016a,b, 2018, 2017) to produce a full segmentation. The downside of this type of approaches is the high level of tuning required, which has to be performed by an expert. Our approach relies on a deep neural network architecture to generalize the features available in a large set of example data. Any features computed are not pre-determined by an expert but stochastically determined by a gradient descent algorithm.

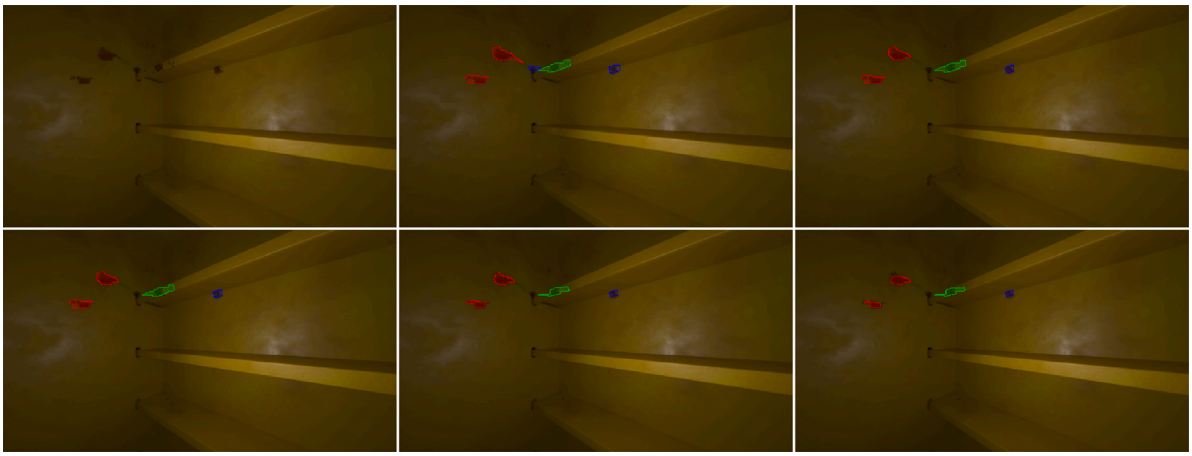


Fig. 2. Example of an input and the 5 associated ground truths. The input image is in the top left. Red=spot corrosion, blue=seam corrosion, green=edge corrosion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Some deep-learning approaches have also been applied. Often using object detectors like Faster R-CNN (Ren et al., 2015), the network is used to detect the corrosion using bounding boxes in the image frame. Examples of this approach include (Cha et al., 2018; Ortiz et al., 2018) where corrosion and other defect types are detected using a Faster R-CNN. The detection approach presented in Liu et al. (2018a,b,c, 2019) is part of a complete system with a drone traversing the ballast tanks of a vessel with the detection algorithm annotating the defects. While efficient, this approach limits the post-processing ability to automatically determine the severity or the area covered by the corrosion since bounding boxes naturally overestimate the size of the corrosion due to the rectangular shape constraint. The network used in this paper provides pixel-wise segmentation and, thus, increases the post-processing ability to determine the physical characteristics of the defect, given mapping from the image plane to the environment.

Similarly, in Andersen et al. (2021) a Faster R-CNN architecture is modified to output bounding boxes as well as the estimated percentage of the area covered in corrosion. While the goal was not to produce pixel-wise accuracy but, rather, to maintain a low inference speed, the method does not generalize to multiple experts. Since the definition of *area under consideration* is loosely defined, it leaves ample room for interpretation to the surveyor which (Andersen et al., 2021) does not capture. By using a probabilistic network, the expected output can be used instead of a deterministic result, providing implicit uncertainty estimations of the output segmentation.

Applying Bayesian methods to introduce variability in the output has previously been used to address the challenges of applying regular deterministic models. Specifically, predictive uncertainty has been used as a measure of confidence for self-labeling tasks (Kendall and Gal, 2017a). Though Bayesian networks capture uncertainty through the estimation of distributions over model parameters, they currently scale poorly with higher dimensions, leaving them infeasible for applications where high image resolution is a concern. Instead, the uncertainty can be approximated with variational models that estimate the posterior in a more controllable manner. This, along with an in depth description of *uncertainty* and *variational inference*, can be found in De Sousa Ribeiro et al. (2020). By maximizing the evidence lower bound, the Kullback–Leibler divergence is minimized resulting in an approximate probability distribution including uncertainty, which is the approach used in this paper.

In general, little work has been published addressing the subjective nature of the inspection process. While many detection models are able to detect corrosion in images, when provided with sufficient training data from a single expert, they lack the ability to incorporate multiple expert opinions. Furthermore, the system will also solely represent a single surveyor's opinion. In this paper, a distribution covering multiple possible solutions is created instead, eliminating the limitation to reflecting only a single opinion.

3. Variational auto-encoder for corrosion detection

The Probabilistic U-Net (Fig. 3) consists of a U-Net feature extractor that outputs a deterministic feature space for a given input image and a low-level latent space $\mathbb{R}^N$. In Kohl et al. (2018), the prior distribution parameters are estimated using a conditional variational auto-encoder (referred to as the prior net). The prior net outputs parameters for a distribution over a latent space. Each sample in the latent space, thus, encodes a segmentation solution conditioned on the input image. The parameters modeled using the prior net are used in an axis-aligned Gaussian distribution with parameters μ and σ, referred to as the prior probability distribution. Sampling the latent space for a given input image X is given by:

$$z \sim \mathcal{N}\left(\mu\left(X\right),\ \mathrm{diag}\left(\sigma\left(X\right)\right)\right) \tag{1}$$

The latent space samples are broadcasted into the same shape as the segmentation map of the U-Net with N number of channels. The segmentation and the broadcasted latent variables are then channel-wise concatenated before four 1×1 convolutions, $\left(f_{unet}\right)$, are performed to merge their information and reduce the number of channels to the number of desired classes. The four convolutions are defined as follows for an output S:

$$S = f_{comb}\left(f_{unet}\left(X\right), \mathbf{z}\right) \tag{2}$$

A downside of this approach is the requirement of multiple samples to produce an expected output, however, the concatenation of the sampled latent space and the features from the U-Net is done late in the network, multiple output masks can be sampled efficiently since there is no need to do inference on the prior net or the U-Net for a given image. Thus, more samples can be generated effectively with only a few 1×1 convolutions.

The Probabilistic U-Net uses a posterior net for training which is identical to the prior net but without sharing weights. The input to the posterior net is one of the labeled masks concatenated with the input image. The output is the same axis-aligned Gaussian parameters as for the prior distribution. Thus, the posterior is estimating the distribution parameters based on the information of a ground truth example. A Kullback–Leibler divergence loss is used between the prior probability distribution and posterior probability distribution. This means the two distributions will be drawn toward each other, and since the posterior net includes the information of the ground truth mask, the prior net will learn a similar distribution but conditioned only on the input image. A standard cross entropy loss between the produced segmentation sample S and the ground truth sample Y is used for a given input image X.

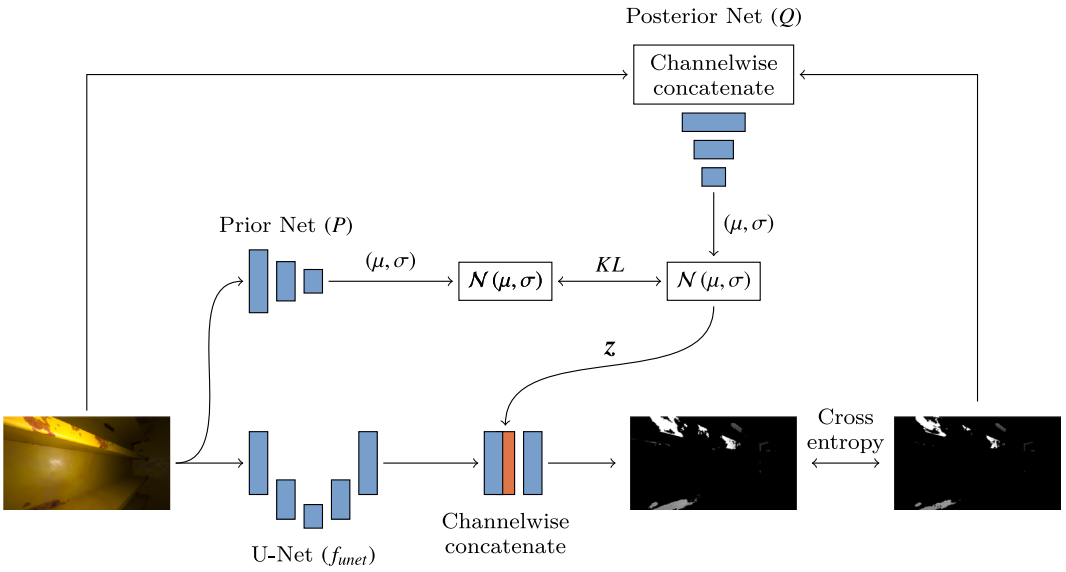


Fig. 3. The Probabilistic U-Net consists of a U-Net and a variational auto-encoder conditioned on the input image. The parameters produced by the auto-encoder get sampled and merged with the output of the U-Net.

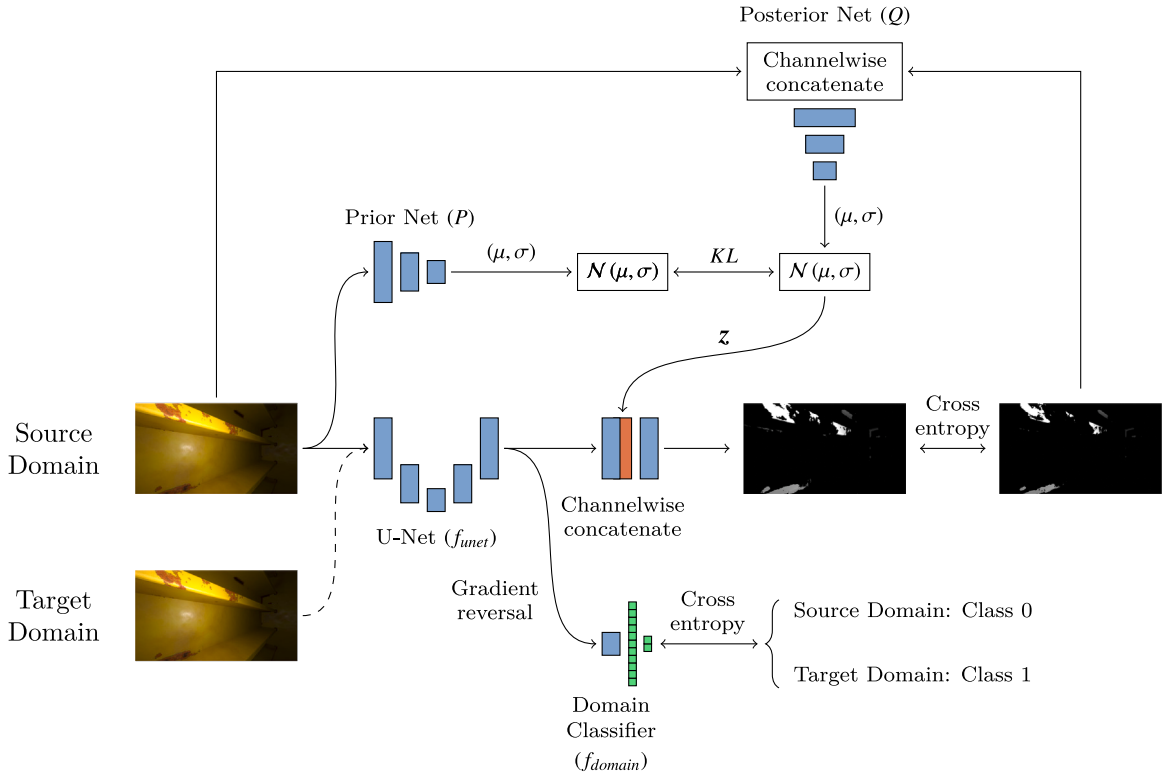


Fig. 4. Similar to the previous setup, the training consists of a prior and posterior network that act on a deterministic feature extractor.

Using the denotation P and Q for the prior and posterior probability distributions, the total loss function is a combination of the two:

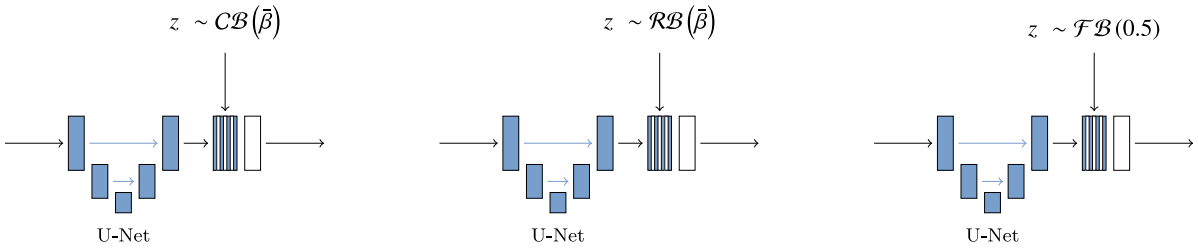
$$\mathcal{L}_s\left(Y_s, X\right) = \mathbb{E}_{z\sim Q(\cdot|Y_s,X)}\left[-\log P_c\left(Y_s|S\left(X,z\right)\right)\right] + \\ +\beta D_{KL}\left(Q\left(z|Y_s,X\right) \parallel P\left(z|X\right)\right) \tag{3}$$

Where P_c denotes the pixel-wise multinomial distribution.

The dimension of the latent space, N, is determined by the number of axis-aligned Gaussian distributions. Both the prior and the posterior are neural networks with the same shape as the first half of the U-Net. The output of both the prior and posterior networks is a vector of length $2 \times N$. To achieve this vectorized output, the spatial dimensions of the prior/posterior feature space are average-pooled.

3.1. Domain adapted probabilistic U-Net

Due to the large shift in colors that happens to the vessel's steel and coating over its lifetime (see Fig. 1), the Probabilistic U-Net has difficulties generalizing the feature space. Therefore, we have modified the network to include unsupervised domain adaptation by backpropagation (Ganin and Lempitsky, 2015) as shown in Fig. 4. The function of the domain classifier, D, is to classify which of the two domains



(a) Continuous Bernoulli (b) Relaxed Bernoulli (c) Fixed dropout

Fig. 5. Results from exploring using a binary latent space rather than a Gaussian distribution.

the input image belongs to. Using a standard entropy loss, the network parameters will produce features that separate the two domains. However, instead of optimizing toward domain separation, the gradients are reversed such that the resulting feature space contains inseparable features from both domains, i.e. a more overlapping projection. The gradient reversal is unsupervised, meaning that we do not need labeled segmentation data for the target domain — the only required information is which domain, Y_d, the image originated from. When the unlabeled domain is fed to the network, the gradients of the segmentation head of the Probabilistic U-Net are zeroed. Thus, the loss is modified to be:

$$\mathcal{L}\left(Y_s, X\right) = \alpha \mathcal{L}_s\left(Y_s, X\right) - \lambda \mathbb{E}\left[-\log D\left(Y_d | f_{unet}\left(X\right)\right)\right] \tag{4}$$

Where $D = f_{domain}\left(f_{unet}\left(X\right)\right)$ is the binary domain classification. $\alpha = 1$ when X belongs to the source domain and 0 otherwise, i.e. when we do not have labels for the target domain, we exclude the loss of the segmentation from the optimization. λ is a tuneable scaling factor for the gradients of the discriminator.

3.2. Variations of the probabilistic U-Net

We have explored different modifications of the Probabilistic U-Net (see Fig. 5), mainly to investigate different latent space encoding methods. The motivation behind evaluating these binary distributions stems from the fact that the ground truth samples are annotated using binary variables. Therefore, a binary distribution might result in an f_{comb} which requires fewer transformations to represent the binary nature of the ground truth. For this reason, we have additionally employed a latent encoding of two different binary distributions; Relaxed Bernoulli (Jang et al., 2016; Maddison et al., 2016) and Continuous Bernoulli (Loaiza-Ganem and Cunningham, 2019). Instead of concatenating latent variables to f_{unet}, we do a channel-wise multiplication, effectively scaling the channels of f_{unet} with some scaling factor based on a learned probability representing the importance of that feature layer. The concept is similar to dropout regularization which, if included at test time, also produces stochastic outputs (Kendall et al., 2015; Kendall and Gal, 2017b). In this case P and Q output vectors of independent probabilities with length equal to the number of channels in f_{unet}. Therefore, we replace S with S_{drop}

$$S_{drop} = f_{drop}\left(f_{unet}, z\right) \tag{5}$$

Where $z \sim \mathcal{RB}\left(\bar{\beta}\left(X\right)\right)$ or $z \sim \mathcal{CB}\left(\bar{\beta}\left(X\right)\right)$ for the relaxed- and continuous Bernoulli respectively and the $\bar{\beta}$ are the axis aligned independent parameters modeled with the prior net. The continuous Bernoulli and relaxed Bernoulli latent space can be viewed as extensions of the fixed dropout version introduced in Kendall et al. (2015). The main difference here is that we try to learn the dropout probability for each channel instead of fixing the dropout rate.

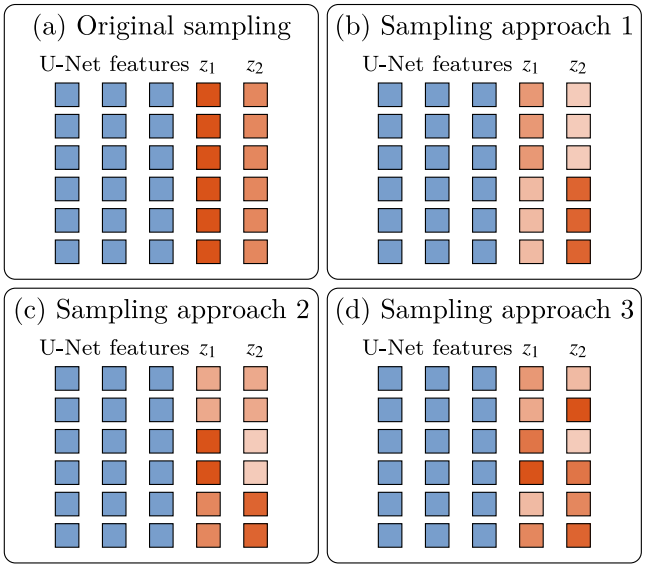


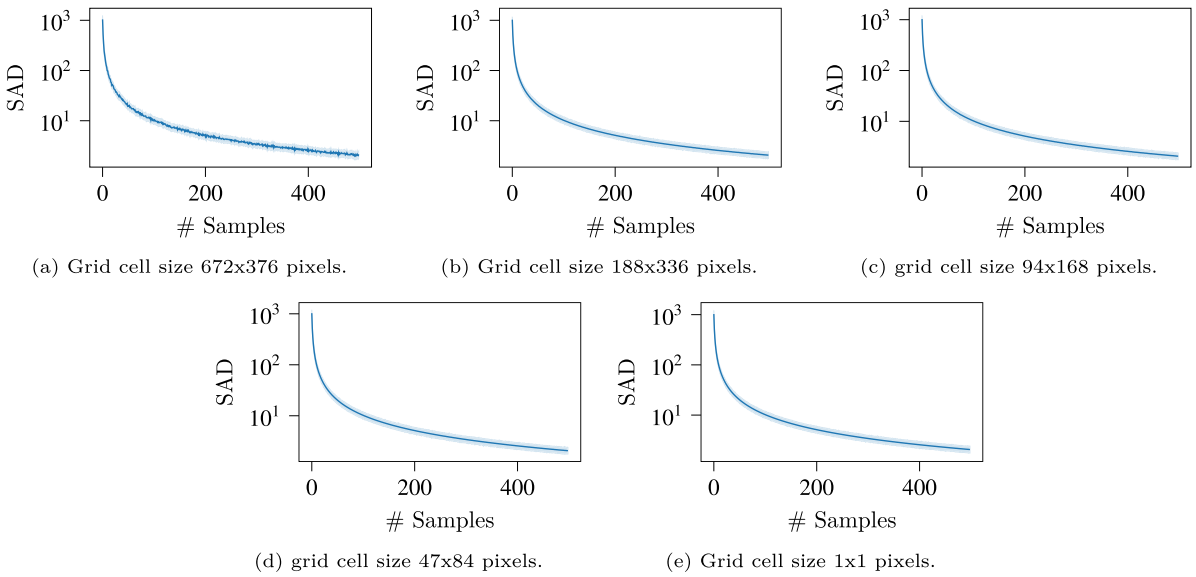
Fig. 6. Illustrates the concatenation of the latent space in the Probabilistic U-Net when the latent space is two dimensions. Each layer can be either sampled using only a single value from the latent space z as in figure a, or they can be sampled in a grid of varying size as shown in figure b–d. Figure d represents the case where each pixel is sampled individually.

4. Data generation and collection

The three dominant types of corrosion defects present on vessels are Edge corrosion, Seam Corrosion, and Spot Corrosion. A corroded area is defined as an edge or seam corrosion if it lies within 100 millimeters of an edge or a seam respectively. Any other present corrosion is defined as spot corrosion.

A high-resolution simulation was generated using a 3D scanned point cloud of a double-skinned oil carrier side ballast tank. The point cloud was converted to a mesh in Blender and a procedurally generated texture was applied. Seams and edges were masked to simulate the increased probability of corrosion appearing in these areas. The resulting mesh and textures are imported to the Unreal Engine-based simulator AirSim (Shah et al., 2017) which offers a higher level of photo-realism than many other simulators while still being able to run in real-time.

From Blender $\left(\mathcal{D}_B\right)$, we have generated 6183 images with five ground truths for each image (see Table 1). From AirSim $\left(\mathcal{D}_A\right)$, we have generated 5225 images with no label information. To efficiently compare the two domains, we have created a dataset, $\left(\mathcal{D}_{BA}\right)$, where the intrinsic and extrinsic parameters of the camera are the same — meaning that there are two images in each position, one from Blender and one from AirSim. Consequently, we can use the ground truth generated from Blender as the ground truth in AirSim. We use these two



(a) Grid cell size 672x376 pixels. (b) Grid cell size 188x336 pixels. (c) grid cell size 94x168 pixels.

(d) grid cell size 47x84 pixels. (e) Grid cell size 1x1 pixels.

Fig. 7. The effect of tiling each channel z_n. While we have shown that more latent space samples per image have little effect on the average output, the tiling procedure has a direct effect on the smoothness of the convergence. A grid cell size of 672 × 376 corresponds to the original Probabilistic U-Net approach.

Table 1
The number of annotations for each ground truth for $\mathcal{D}_B$. The number of training images is 6183.

	Annotations					Average
	GT 1	GT 2	GT 3	GT 4	GT 5	
Seam	25819	21013	21035	20929	21056	21970.4
Edge	24735	21424	21444	21412	21337	22070.4
Spot	13621	8514	8569	8612	8491	9561.4
Total	64175	50951	51048	50953	50884	–

Table 2
The number of annotations for each ground truth for $\mathcal{D}_{BA}$. The number of test images is 2782.

	Annotations					Average
	GT 1	GT 2	GT 3	GT 4	GT 5	
Seam	6801	3994	3975	3942	4054	4553.2
Edge	4824	2843	2896	2819	2934	3263.2
Spot	2507	1562	1593	1543	1600	1761
Total	14132	8399	8464	8304	8588	–

datasets for testing — i.e. none of the networks will see images from $\mathcal{D}_{BA}$ during training. The mixed dataset contains 2782 images with five ground truths for each image (see Table 2).

The field of view of the simulated cameras are 120 degrees with an image resolution of 672 × 376. The field of view was based on the Mynteye D1000-120 stereo camera and the resolution was chosen based on a trade-off between rendering time and image quality. Additionally, the images are rectified to reduce the sensitivity to camera distortions across domains.

To show that the pipeline presented in this paper also works on real data, we have collected a small dataset, $(\mathcal{D}_R)$, of real images. The dataset was annotated manually, which introduces a high level of subjectiveness in the labeling. Additionally, we only have a single annotation per image and the intrinsic parameters of the camera used varies across the images resulting in varying resolutions, signal-to-noise ratio, and lighting conditions — whereas the simulated data all have a camera with fixed intrinsic parameters and a light mounted with a static transform close to the camera. The real dataset contains 600 annotated images. In addition to the annotated images, we have 3847 non-annotated images. The non-annotated images were captured using an Intel RealSense L515 mounted on a custom drone for marine vessel inspection (Brogaard et al., 2020). The images are of size 640 × 480. The non-annotated images will be used to qualitatively determine the domain adaption on real images.

5. Training and evaluation metrics

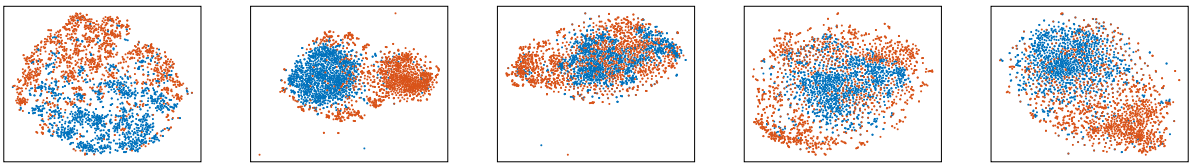
When training the Probabilistic U-Net, we use an Adam optimizer (Kingma and Ba, 2014) with a learning rate of $1e-5$. Additionally we use L2 regularization on the weights of f_{unet}, P, and Q. We set $\lambda = 1$ as we found these values worked sufficiently well. The posterior and prior are initialized with an orthogonal normal with parameters $(\mu = 0, \sigma = 0.001)$. The rest are initialized using a kaiming normal distribution (He et al., 2015).

5.1. Domain adaption on simulated data

In the original work, the Probabilistic U-Net is trained on monochrome images of size 128 × 128 whereas we are training on colored simulated images with size 672 × 376. This increase in dimensions makes the Probabilistic U-Net prone to divergence. To achieve stable and reproducible training sessions, we train the network initially with a fixed prior and posterior for 80000 steps. After which we reset the optimizer and train all model weights for another 32000 steps. We use a batch size of 16 throughout all experiments. To maintain a fair comparison, we use the same parameters where applicable for all baseline methods.

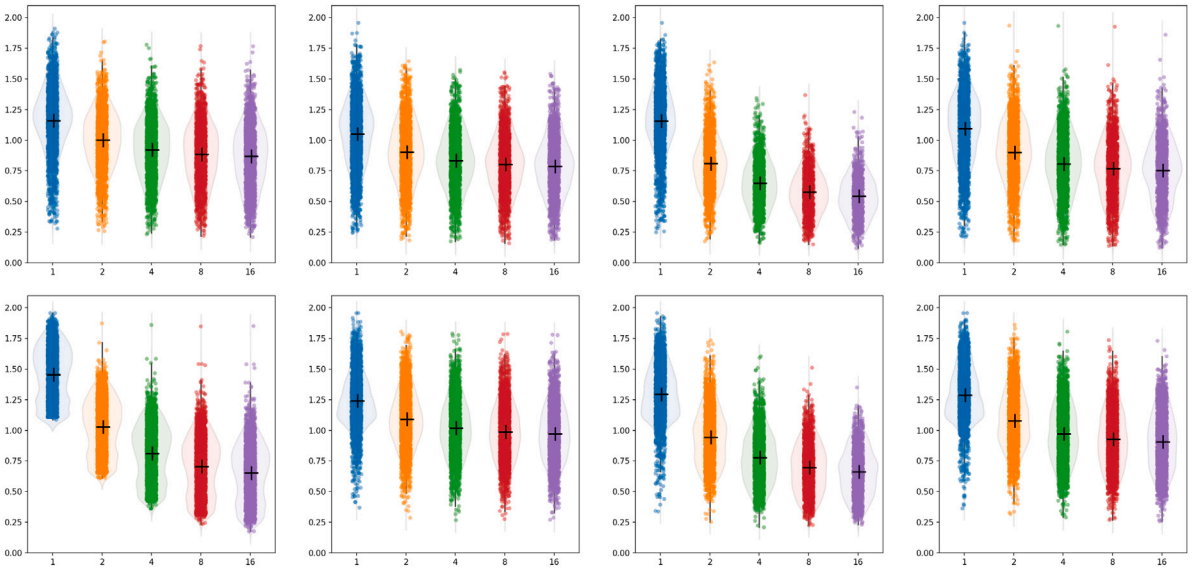
Since the Probabilistic U-Net produces samples from a distribution, we cannot evaluate the performance using standard intersection over union. We evaluate the domain adaption between the three domains as a difference in generalized energy distance where the distance metric we use is 1-IoU — the same metric as proposed in the original Probabilistic U-Net (Kohl et al., 2018). We baseline against three variations of the Probabilistic U-Net introduced in Section 3.2.

When evaluating real images, we use the same training process as was used on the simulated data with the exception of the prior and posterior. Since we only have a single ground truth for each image, we cannot learn an encoding of the experts. So we initialize the prior and posterior as before, but we do not update the weights during backpropagation. This way we do not have a fixed distribution (e.g. a standard Gaussian) but instead compute the distribution parameters based on the input image.



(a) No training (b) Gaussian (c) Continuous Bernoulli (d) Relaxed Bernoulli (e) Fixed dropout

Fig. 8. The t-SNE reduced feature space for each of the probabilistic models. Blue encodings are features from $D_{\mathcal{B}}$ and orange encodings are features from $D_{\mathcal{A}}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Gaussian (b) Continuous Bernoulli (c) Relaxed Bernoulli (d) Fixed dropout

Fig. 9. Results from exploring using a binary latent space rather than a Gaussian distribution. The top and bottom row shows energy distance from the source and target domain respectively.

Since we only have a single ground truth for each image, we cannot use the energy distance to measure the network's ability to capture the expert distribution as it reduces to a deterministic case. Instead, we qualitatively visualize the feature space using t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008) and by color-encoding the domain label.

5.2. Determination of latent space sampling size

Due to the non-linear transformation (from the non-linear activation layers) located directly after concatenating the latent space, the properties of the Gaussian distribution are no longer valid. Thus, to produce any statistics about the samples, we have to sample the network several times. The reasoning for not relying on a single sample drawn from the network is that all expert opinions are encoded in the latent space and can be reproduced with the correct latent variables. Thus, if we only needed one sample, we could simply produce one expert opinion in a traditional deterministic fashion.

As shown in Fig. 6a, in the original implementation of the Probabilistic U-Net, every latent space sample — a single number — is used to fill in each channel $z_n, n = 1, 2, \dots, N$ in both the height and the width dimension. While this means we can sample outlier expert opinions, it makes the average fluctuate, while sampling toward convergence. To reduce this fluctuation, we propose sampling using a grid as shown in Fig. 6b–c.

We have investigated the effect of sampling multiple latent samples per channel, in a grid manner with varying tile size up to 188×336. In this lowest granularity case, four samples per image channel are used which in 2D projection corresponds to Fig. 6b. The statistic we use to measure convergence is the sum of the absolute difference between two consecutive averages defined as

$$SAD = \sum_{w=0}^{W} \sum_{h=0}^{H} |\bar{s}_N(w,h) - \bar{s}_{N-1}(w,h)| \tag{6}$$

$$\bar{s}_N(w,h) = \frac{1}{N} \sum_{n=0}^{N} S_n(w,h) \tag{7}$$

where W and H are the width and height of the image and $\bar{s}_N$ is the pixel-wise average over N samples produced by the network S.

Since each pixel or each grid cell will be sampled from the same distribution, this is equivalent to performing inference on a single image. Computing the average over all pixels in the image then means we are computing the expected value using more samples from the distribution than just a single sample. In Fig. 7a, the convergence curve shows that the sampling is sensitive to outliers, thereby determining the minimum samples size is more difficult. As can be seen from the lack of fluctuations in Figs. 7b–d, sampling in a grid drastically smoothens the convergence at no loss in quality of distribution representation. Thus, in cases where the expected value of the hidden expert distribution is desired more than any single sample, sampling in a grid shape can make it easier to choose the number of samples required.

6. Experiments and assessment

In the following section, we present the results of the different variations of the domain-adapted probabilistic models. Firstly, using the simulated datasets $D_{\mathcal{B}}$ as the source and $D_{\mathcal{A}}$ as the target. Secondly, we train on the dataset with real images $D_{\mathcal{R}}$ using the annotated images

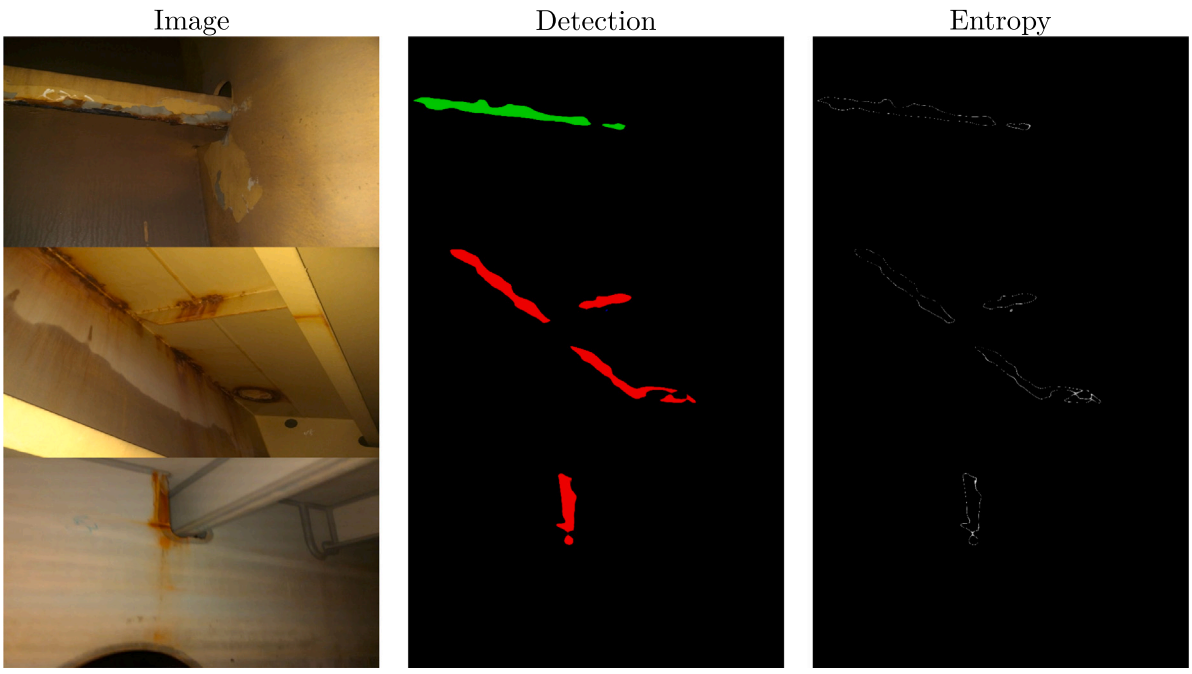


Fig. 10. Each row shows an image containing real corrosion from D_R. The first column is the input image. The second column is the average output of 100 samples from the Probabilistic U-Net. The last column is the pixel-wise entropy from the 100 samples. Brighter pixels indicate more randomness, i.e. the pixel is flipping between classes and/or the background.

as the source domain and the non-annotated images captured from the drone as the target domain.

6.1. Domain adaption on simulated data

The results reported in this subsection are from the aligned dataset $D_{B\mathcal{A}}$ which were never used during training. Additionally, we show the t-SNE feature reduction as a qualitative measure of training success. Although the images are similar and will thus most likely produce near-similar features, we can see from Fig. 8 that the models adapt to the domain with varying success. Specifically, the continuous and relaxed Bernoulli version of the probabilistic models has succeeded in unifying the feature space while the original Probabilistic U-Net and the fixed dropout rate have not managed to combine the two domains within the same cluster. However, as shown by the energy distance in Fig. 9, the Probabilistic U-Net and the fixed model both have similar performance as the continuous Bernoulli Probabilistic U-Net on the source domain $D_{\mathcal{B}}$. The downward trend as more samples are drawn indicates that more samples help approximate the expert distribution, with the relaxed Bernoulli model being the best-performing model. On the target domain, the Probabilistic U-Net is the worst performing model when sampled only once, but reaches a similar performance as the relaxed Bernoulli after multiple samples.

6.2. Domain adaption on real data

Since we do not have multiple ground truths for the expert labeled data, using the energy distance as a metric is of no use as it would reduce to a deterministic case. Instead, we qualitatively evaluate the performance by visualizing the entropy of the output mask as shown in Fig. 10. The output variation is mostly located around the edges of the blobs in the mask as expected since adding samples from a static distribution will simply add a random bias in the 1×1 convolutions in f_{comb}. In Fig. 11 we see that the features produced in the source and target domain are overlapping. The target images were taken from random parts of a video stream captured using a real drone. Since sequential images produce similar features, the visualized features exhibit some degree of inter-domain correlation.

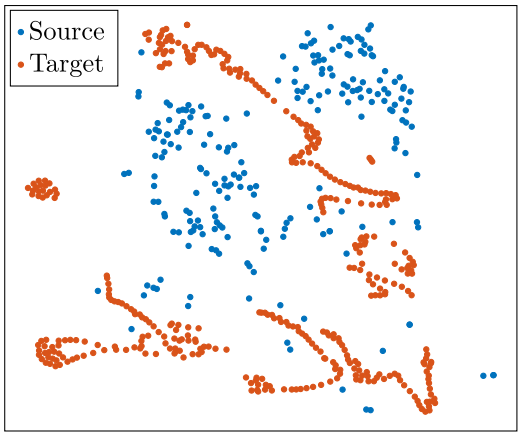


Fig. 11. T-SNE of the validation data from $D_{\mathcal{B}}$. The data was captured from a video stream from a drone. To preserve as much data for training as possible, we have only used approximately 400 images for clear visualization.

7. Conclusion

In this paper, we have demonstrated the use of probabilistic models with domain adaptation for inspections of marine vessels. While the Probabilistic U-Net architecture performs similarly to our proposed relaxed Bernoulli Probabilistic U-Net on the target domain, our version additionally showed better performance on the source domain and, hence, better domain adaptation as it seems to be less prone to catastrophic forgetting. In general, with the introduction of probabilistic inspection, the systems will be able to generalize the expert distribution to other domains. This means that the same expert distribution can be used across the entire lifetime of the vessel, as the differences between younger and older vessels can be seen as a shift in the feature domain. Additionally, we introduced a sampling strategy that simplifies the decision of the number of samples necessary for a representative expected output that does not exhibit unexpected fluctuations. We developed two baselines inspired by another probabilistic model based on

dropout regularization layers to introduce variability in the output. In the future, we will investigate the introduction of more complex latent space encoding methods, including attentional networks. Additionally, we will investigate the probabilistic implementation of more recent and well-performing segmentation models.

CRediT authorship contribution statement

Rasmus Eckholdt Andersen: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Visualization, Investigation, Reviewing and editing. **Evangelos Boukas:** Supervision, Writing – review & editing.

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

Abdallah, A.M., Atadero, R.A., Ozbek, M.E., 2022. A state-of-the-art review of bridge inspection planning: current situation and future needs. J. Bridge Eng. 27 (2), http://dx.doi.org/10.1061/(asce)be.1943-5592.0001812.

Andersen, R.E., Nalpantidis, L., Boukas, E., 2021. Vessel classification using a regression neural network approach. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 4480–4486.

Bonnin-Pascual, F., Ortiz, A., 2014. A probabilistic approach for defect detection based on saliency mechanisms. In: 19th IEEE International Conference on Emerging Technologies and Factory Automation. ETFA 2014, http://dx.doi.org/10.1109/etfa.2014.7005257.

Bonnin-Pascual, F., Ortiz, A., 2016a. A flying tool for sensing vessel structure defects using image contrast-based saliency. IEEE Sens. J. 16 (15), 6114–6121. http://dx.doi.org/10.1109/jsen.2016.2578360.

Bonnin-Pascual, F., Ortiz, A., 2016b. A generic framework for defect detection on vessel structures based on image saliency. In: 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation. ETFA, IEEE, pp. 1–4. http://dx.doi.org/10.1109/etfa.2016.7733668.

Bonnin-Pascual, F., Ortiz, A., 2017. A saliency-boosted corrosion detector for the visual inspection of vessels. Front. Artif. Intell. Appl. 300, 176–185. http://dx.doi.org/10.3233/978-1-61499-806-8-176.

Bonnin-Pascual, F., Ortiz, A., 2018. A novel approach for defect detection on vessel structures using saliency-related features. Ocean Eng. 149 (August 2017), 397–408. http://dx.doi.org/10.1016/j.oceaneng.2017.08.024.

Bonnin-Pascual, F., Ortiz, A., 2019. On the use of robots and vision technologies for the inspection of vessels: A survey on recent advances. Ocean Eng. 190, 106420. http://dx.doi.org/10.1016/j.oceaneng.2019.106420.

Brogaard, R.Y., Zajaczkowski, M., Kovac, L., Ravn, O., Boukas, E., 2020. Towards uav-based absolute hierarchical localization in confined spaces. In: 2020 IEEE International Symposium on Safety, Security, and Rescue Robotics. SSRR, IEEE, pp. 182–188.

Cha, Y.J., Choi, W., Suh, G., Mahmoudkhani, S., Büyüköztürk, O., 2018. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. Comput.-Aided Civ. Infrastruct. Eng. 33 (9), 731–747. http://dx.doi.org/10.1111/mice.12334.

De Sousa Ribeiro, F., Calivá, F., Swainson, M., Gudmundsson, K., Leontidis, G., Kollias, S., 2020. Deep bayesian self-training. Neural Comput. Appl. 32 (9), 4275–4291.

Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning. PMLR, pp. 1180–1189.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034.

Hess, R., 2007. The Essential Blender: Guide to 3D Creation with the Open Source Suite Blender. No Starch Press.

Jang, E., Gu, S., Poole, B., 2016. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144.

Kendall, A., Badrinarayanan, V., Cipolla, R., 2015. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680.

Kendall, A., Gal, Y., 2017a. What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30.

Kendall, A., Gal, Y., 2017b. What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S.M.A., Jimenez Rezende, D., Ronneberger, O., 2018. A probabilistic U-net for segmentation of ambiguous images. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 31, Curran Associates, Inc., pp. 1–11, URL https://proceedings.neurips.cc/paper/2018/file/473447ac58e1cd7e96172575f48dca3b-Paper.pdf.

Liu, Y., Hajj, M., Bao, Y., 2022. Review of robot-based damage assessment for offshore wind turbines. Renew. Sustain. Energy Rev. 158, 112187. http://dx.doi.org/10.1016/j.rser.2022.112187, URL https://www.sciencedirect.com/science/article/pii/S1364032122001113.

Liu, L., Tan, E., Cai, Z.Q., Yin, X.J., Zhen, Y., 2018a. CNN-based automatic coating inspection system. Adv. Sci. Technol. Eng. Syst. 3 (6), 469–478. http://dx.doi.org/10.25046/aj030655.

Liu, L., Tan, E., Cai, Z.Q., Zhen, Y., Yin, X.J., 2018c. An integrated coating inspection system for marine and offshore corrosion management. In: 2018 15th International Conference on Control, Automation, Robotics and Vision. ICARCV 2018, Institute of Electrical and Electronics Engineers Inc., pp. 1531–1536. http://dx.doi.org/10.1109/icarcv.2018.8581327.

Liu, L., Tan, E., Yin, X.J., Zhen, Y., Cai, Z.Q., 2019. Deep learning for Coating Condition Assessment with Active perception. In: Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference on - HPCCT 2019. ACM Press, New York, New York, USA, pp. 75–80. http://dx.doi.org/10.1145/3341069.3342966, URL http://dl.acm.org/citation.cfm?doid=3341069.3342966.

Liu, L., Tan, E., Zhen, Y., Yin, X.J., Cai, Z.Q., 2018b. AI-facilitated coating corrosion assessment system for productivity enhancement. In: Proceedings of the 13th IEEE Conference on Industrial Electronics and Applications. ICIEA 2018, Institute of Electrical and Electronics Engineers Inc., pp. 606–610. http://dx.doi.org/10.1109/iciea.2018.8397787.

Loaiza-Ganem, G., Cunningham, J.P., 2019. The continuous Bernoulli: fixing a pervasive error in variational autoencoders. Adv. Neural Inf. Process. Syst. 32.

Maddison, C.J., Mnih, A., Teh, Y.W., 2016. The concrete distribution: A continuous relaxation of discrete random variables. arXiv preprint arXiv:1611.00712.

Nguyen, V.N., Jenssen, R., Roverso, D., 2018. Automatic autonomous vision-based power line inspection: A review of current status and the potential role of deep learning. Int. J. Electr. Power Energy Syst. 99 (September 2017), 107–120. http://dx.doi.org/10.1016/j.ijepes.2017.12.016.

Ortiz, A., Yao, K., Bonnin-Pascual, F., Garcia-fidalgo, E., Company-corcoles, J.P., 2018. New steps towards the integration of robotic and autonomous systems in the inspection of vessel holds. In: Jornadas Nacionales de Robótica (Spanish Robotics Workshop).

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 28, Curran Associates, Inc., pp. 1–9, URL https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

Rizzo, C.M., 2008. 13 - Inspection of aged ships and offshore structures. In: Paik, J., Melchers, R. (Eds.), Condition Assessment of Aged Structures. In: Woodhead Publishing Series in Civil and Structural Engineering, Woodhead Publishing, pp. 367–406. http://dx.doi.org/10.1533/9781845695217.5.367, URL http://www.sciencedirect.com/science/article/pii/B9781845693343500139.

Shah, S., Dey, D., Lovett, C., Kapoor, A., 2017. AirSim: high-fidelity visual and physical simulation for autonomous vehicles. In: Field and Service Robotics. URL https://arxiv.org/abs/1705.05065, arXiv:arXiv:1705.05065.

Su, J., Vargas, D.V., Sakurai, K., 2019. One pixel attack for fooling deep neural networks. IEEE Trans. Evol. Comput. 23 (5), 828–841. http://dx.doi.org/10.1109/tevc.2019.2890858.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).