Sensors & Transducers, Vol. 249, Issue 2, February 2021, pp. 25-35

www.sensorsportal.com

Sensors & Transducers

Published by IFSA Publishing, S. L., 2021
http://www.sensorsportal.com

Uncertainty Estimation for Deep Learning-based Thermographic Imaging

[1, *] Bernhard LEHNER, [2] Thomas GALLIEN, [1, 4] Péter KOVÁCS, [5] Gregor THUMMERER, [5] Günther MAYR, [6] Peter BURGHOLZER and [1, 3] Mario HUEMER

[1] Silicon Austria Labs (SAL), JKU LIT SAL eSPML Lab, Altenbergerstr. 69, 4040, Austria
[2] Silicon Austria Labs (SAL), Inffeldgasse. 33, 8010, Austria
[3] Johannes Kepler University Linz, Institute of Signal Processing, JKU LIT SAL eSPML Lab, Altenbergerstr. 69, 4040, Austria
[4] Department of Numerical Analysis, Eötvös Loránd University, 1117, Hungary
[5] Josef Ressel Centre of Thermal NDE of Composites, University of Applied Sciences Upper Austria, Roseggerstraße 15, 4600, Austria
[6] Research Center for Non Destructive Testing, Altenberger Straße 69, 4040, Austria
[1] Tel.: +43 5 0317
E-mail: bernhard.lehner@silicon-austria.com

Received: 1 October 2020 /Accepted: 15 January 2021 /Published: 28 February 2021

Abstract: Thermographic imaging is a contactless and nondestructive way to detect defects inside the specimen. The current state-of-the-art approach combines model- and deep learning-based reconstructions in a hybrid fashion. The recently developed virtual wave concept (VWC) provides a framework to develop such hybrid solutions, and allows to utilize physical priors, such as non-negativity and/or sparsity. In combination with the superiority of deep learning approaches over hand-crafted features and heuristics, improved reconstruction accuracy compared to previous methods was achieved. However, the reconstruction results still deteriorate under low SNR conditions caused by defects located deeper underneath the surface. Therefore, it would be useful to have an uncertainty estimate that reflects the reliability of the reconstructions. This would enable the automatic identification of results that are likely to be inaccurate and require a closer inspection. In this paper, we propose two computationally very cheap methods to estimate the uncertainty of thermographic imaging results. In order to show the generalization capability of our approach, we thoroughly evaluate it under different conditions and even with different deep model architectures.

Keywords: Thermal tomography, Virtual wave, Nondestructive testing, Regularization, Sparsity, Deep learning.

1. Introduction

Deep learning-based approaches for thermographic imaging have been shown to significantly outperform pure numerical, model-based methods even under real-world conditions [1]. However, the thermographic reconstruction results still show signs of deterioration under low SNR conditions which are caused by deeper lying subsurface defects.

This phenomenon is to a large degree a manifestation of information loss, and can only be

partially compensated by priors in the form of non-negativity and sparsity. Therefore, it would be helpful to at least being able to automatically identify highly deteriorated thermographic reconstructions.

In our specific setting, uncertainty estimation should provide a single-valued indication of the reliability of thermographic reconstructions, and thus be highly correlated with the actual loss. Furthermore, it should be applicable in real-time scenarios and on embedded systems with restricted computational capabilities. Consequently, it is important that the method is computationally light-weight and fast. So, existing methods that require several forward passes, like e.g., dropout-based uncertainty estimation [2] or Bayesian ensembling [3] cannot be used.

Our previous work on uncertainty estimation [4] is the starting point of this paper, and we extend it in several aspects. First, we present novel results in a different setting that demonstrates the generalization capability of our approach. Second, we provide more detailed insights to the results so far, and also address possible extensions. Third, we discuss the potential limitations of our method and point to possible future work in order to further improve our current approach.

In the following sections, we first introduce the basic principles of thermographic imaging along with some practical applications for it. Afterwards, we briefly discuss the challenges that are posed by thermodynamics and entropy for the task at hand. We then present two techniques designed to alleviate the deterioration of the results related to these challenges. First, we discuss model-based reconstruction techniques that require modelling the physical aspects of the problem along with regularization. Second, we review deep learning-based reconstruction techniques, where deep learning is either applied in an end-to-end fashion, or along with domain knowledge.

1.1. Thermographic Image Reconstruction

Non-destructive evaluation (NDE) has become an integral part in multiple industry processes in which a product failure might result in an accident or body injury. NDE allows to identify defects before they become critical and thus helps to avoid serious damage of components [5]. Due to rapid developments in the field of infrared (IR) detectors and advanced approaches in signal processing, the emerging NDE technique of active IR thermography (IRT) has become a powerful tool [6]. Contactless and non-intrusive, this technique provides a rapid means for characterizing thermal properties and for locating and sizing of subsurface defects. Since 2017, active IRT has been qualified for testing safety-critical components in the aviation industry. Due to the non-ionizing thermal radiation, active IRT also has emerging applications in medicine [7].

In most thermal NDE applications, one-dimensional (1D) physic-based forward models are used for image enhancement and depth estimation [8]. The 1D models become inaccurate for the reconstruction of finite sized defects, taking into account that the anisotropic heat conduction, e.g., in composite materials, amplifies these effects [9]. To increase the depth resolution, radar inspired signal processing techniques, as matched filtering, were implemented to detect a pre-known signal waveform in a highly noisy channel [10].

Image reconstruction can be seen as the solution of a mathematical inverse problem in which the cause (measured surface temperature) is inferred from the effect (internal heat sources or reflectors). In general, inversion of thermal diffusion is severely ill-posed due to the diffusive nature of the heat propagation. A physical model describes the relationship between the surface temperature and the heat distribution inside the material. Assuming a particular source shape, a parameter estimation problem has to be solved by the minimization of an objective function for a set of parameters (e.g., diameter, depth, thickness). If the shape is not known a priori, regularization procedures can be used for the reconstruction of any energy distribution in a predefined mesh grid [11]. In particular, pulse-echo photo-thermal imaging is the most widely implemented method because the sample is often only accessible from one side. A uniformly spatially distributed pulse, e.g., flashlights, launches a package of plane diffusion waves into the material which propagates normal to the surface. When the scattered waves return to the surface, it affects the surface temperature decay. Holland and Schiefelbein proposed a full 3D image reconstruction technique, which represents the surface temperature as a linear combination of heat contributions [12]. All these methods are based on a large-scale linear inversion that considers (admittedly imperfectly) lateral heat flows. Furthermore, the surfaces have to be segmented into overlapping tiles due to the high computational costs.

A completely new imaging approach is to employ photoacoustic techniques for photothermal signals, the so-called virtual wave concept (VWC) [13]. The reconstruction process is a two-stage process. In the first step, a virtual wave signal is calculated for every location of the measured surface temperature signal. As the virtual wave obeys the wave equation, it can be reversed in time in a second step to reconstruct the internal heat sources and reflectors. This computationally cheap two-stage process for imaging can be used for 1D, 2D and 3D heat conduction problems.

However, the resolution limit of the thermographic reconstruction decays linearly with depth [14]. This diffusion-based blurring can be compensated by incorporating a priori information (e.g., sparsity, non-negativity, etc.) for the computation of the virtual wave field. Additionally, the signal-to-noise ratio (SNR) can be significantly increased by using well-known acoustic reconstruction methods, such as synthetic aperture focusing techniques (SAFTs). Here, the heat flow in all directions is taken into account for reconstruction.

However, handcrafted features and heuristics can be outperformed by deep learning approaches, as was demonstrated in [1]. Here, two training strategies were applied. The first was carried out in an end-to-end fashion, where the surface temperature data was fed directly to a neural network. The second approach incorporated domain knowledge, where the VWC [13] served as a mid-level representation that was fed to the neural network.

1.2. Practical Applications

The model- and deep learning-based reconstruction techniques mentioned in the last sections can be utilized in a wide range of NDE applications [14]. As a case-study, we considered the problem of thickness estimation in 1D [15], and active thermographic computed tomography in 2D [1] and 3D [13]. To this end, we built up various test specimens, such as steel step wedges, and physical phantoms made of epoxy resin, which includes steel beads and graphite bars. With the help of these specimen, the positive impact of prior knowledge (e.g., non-negativity, sparsity) in the reconstruction process could be demonstrated.

Furthermore, we could test the generalization ability of the deep learning-based reconstruction methods. This is extremely important, as the deep neural networks for thermographic image reconstruction were trained on simulated data only. It was shown in [1] that the neural networks can successfully generalize the learned reconstruction algorithm to real-world data even under previously unseen conditions (e.g., noise, rotation). In order to foster the development of other reconstruction techniques, we released both the synthetic and the real-world datasets along with the implementation of the deep learning-based methods to the research community[1].

1.3. Principles of Thermography

Thermography uses heat diffusion to characterize thermal properties and to estimate the location and size of subsurface defects. In active IRT, the sample is heated up by absorption of light, by eddy-currents for electrically conductive internal structures, or by other energy sources, such as ultrasound or microwave absorption. The surface temperature development caused by this is observed with an IR camera. Depending on the direction of the heat diffusion, there are two scenarios. First, the subsurface structures to be imaged can be heated directly, e.g., by eddy-currents. Here, the heat diffuses to the surface, and we consider this a one-way diffusion. Second, the sample surface is heated up, e.g., by a short laser pulse. In that case, the heat diffuses into the sample and the subsurface structures influence the measured surface temperature evolution. We consider this *a two-way diffusion*.

In Fig. 1, we can see an example of a measurement setup, where the specimen is heated up by eddy-currents. All investigations in this paper refer to such a *one-way diffusion* scenario.

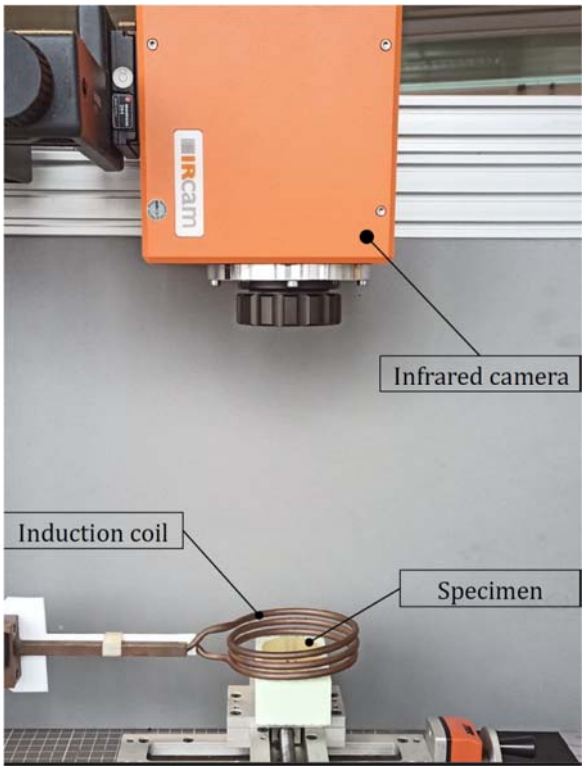

Fig. 1. Measurement setup where the specimen is heated up by eddy-currents.

1.4. Challenges of Thermographic Imaging

Only ideal processes in which dissipative effects such as friction are neglected can be reversed in time. An example of such an ideal process would be a pendulum where damping is ignored, hence oscillating forever. However, real processes like heat conduction do not exhibit a rewind button. From the 2nd law of thermodynamics, we know that heat always flows from the hot to the cold and never in the other direction due to entropy production. Since Szilard [16] and Landauer [17] with his aphorism "Information is physical" we know that the acquisition of information, such as the measuring process in thermography, must obey the laws of (non-equilibrium) thermodynamics. Heat diffusion in active thermography causes entropy production which turns out to be equal to information loss, and in imaging processes less information always corresponds to a limited spatial resolution.

In the VWC we link the measured thermal signal to an ideal solution of the wave equation, the so-called virtual wave, which is reversible in time. Due to the

[1] Available at: https://git.silicon-austria.com/pub/confine/ThermUNet

information loss for the thermal signal, the calculation of the virtual wave is an ill-posed inverse problem that requires regularization. In frequency domain, the loss of information due to entropy production during heat diffusion results in a limited bandwidth [14]. For all frequency components above a truncation frequency the distribution density for the signal does not differ significantly from the equilibrium density. Therefore, these high frequency components cannot contribute to the reconstructed virtual wave. The spatial resolution limit is then diffraction limited to half of the wavelength at the truncation frequency. It turns out that this information related criterion results in the same truncation frequency as when the signal amplitude in frequency domain gets less than the noise level, described by the SNR. For thermographic imaging with its high entropy production from heat diffusion, it is essential to use additional information such as positivity and sparsity by implementing iterative algorithms, e.g., the alternating direction method of multipliers (ADMM) to get a better resolution.

1.5. Model-based Reconstruction Techniques

Thanks to the VWC, the problem of thermographic imaging can be addressed by the two-stage reconstruction process that we already discussed in Section 1.1. Mathematically, each stage is an ill-posed inverse problem. Therefore, regularization is inevitable in order to obtain feasible solutions. Penalized least-squares methods provide a general framework to solve such problems:

$$\tilde{\boldsymbol{x}} = \mathrm{argmin}_{\boldsymbol{x}}\{||\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}||_2^2 + \lambda^2\Omega(\boldsymbol{x})\}, \qquad (1)$$

where $\boldsymbol{b}$ represents the measurement data, $\boldsymbol{A}$ is a matrix that models the physical phenomenon (e.g., heat diffusion, wave propagation), and $\tilde{\boldsymbol{x}}$ is an approximate solution to the virtual waves or to the initial temperature profile of the specimen. The penalty function $\Omega(\cdot)$ is of particular importance that allows to utilize domain knowledge, such as smoothness, non-negativity, sparsity, joint or group sparsity [18]. There are various ways to incorporate such constraints into the reconstruction process, e.g., by using spherical projections [19], Abel- or curvelet transformations [20].

Besides its numerical challenges, there are many computational issues in the reconstruction process. One of these is related to the proper choice of $\lambda > 0$, which is a crucial parameter of the regularization that controls the trade-off between overfitting and underfitting. Several state-of-the-art algorithms can be adopted to estimate the optimal value of λ, such as the L-curve method [21], and the Picard condition [22] that takes the special structure of $\boldsymbol{A}$ into account. This matrix has a block diagonal structure that allows the efficient implementation of the regularization by using the ADMM [23] in conjunction with the L-curve method.

However, this does not apply to the second stage, where the forward operator $\boldsymbol{A}$ cannot be explicitly formulated. An alternative to resolve this problem would be to use time- and frequency-domain SAFT [24, 25]. Another approach would be to utilize deep learning, as will be described in the next section.

1.6. Deep Learning-based Reconstruction Techniques

Besides their rigorous mathematical background, the performance of the model based reconstruction methods is limited. For instance, convex relaxation is one of the most common techniques for solving sparse approximation problems [26, 27]. These iterative regularization algorithms usually have high computational complexity. Hence, their convergence and thus the reconstruction process, which is often recalculated multiple times (e.g., for each cross-sectional data of a 3D measurement), may be slow. Additionally, modeling assumptions such as linearity, convexity, and normality restrict the eligible class of prior knowledge. These handcrafted features and heuristics are often incorporated into Eq. (1) via mathematical constraints and penalty functions Ω, which can lead to suboptimal solutions.

Deep learning is a promising alternative to resolve some of the remaining issues, as was shown recently [18, 1]. However, supervised deep learning methods typically require large amounts of labeled data, which is seldom easy to obtain. In thermography, that would involve the production of physical, so-called phantoms, with varying material properties, e.g., defects at various positions, sizes, and shapes. Fortunately, relatively simple analytic models of the heat diffusion process can be used to synthesize arbitrarily large amounts of data. In order to ensure proper generalization capabilities of a deep learning approach trained with such data, an evaluation under previously unseen, even real-world conditions is of utmost importance.

In [1], the superiority of deep learning approaches over handcrafted features and heuristics was demonstrated via two training strategies. The first was carried out in an end-to-end fashion. That is, the surface temperature data was fed directly to a neural network. The second approach incorporated domain knowledge. Here, the virtual wave concept [13] was involved as a feature extraction step before the neural network was applied.

Unfortunately, limitations of the physical world as discussed in Section 1.4 cannot simply be overcome by just applying deep learning. For that reason, the reconstruction results still deteriorate under low SNR conditions, and for defects located deeper underneath the surface, albeit to a much lesser extent compared to the numerical model-based methods. Therefore, it would be useful to have an uncertainty estimate that reflects the reliability of any given output. This would in turn enable the identification of results that are likely to be inaccurate, hence requiring a closer

investigation. In [4], we introduced two uncertainty estimation methods specifically targeting outputs from the deep learning approaches previously mentioned. However, due to the limitations of a short paper, some details and questions about these methods had to be left out. The remainder of this paper is therefore dedicated to fill in those gaps.

2. Uncertainty Estimation

In this section, we propose two different methods for uncertainty estimation. In order to keep the methods lightweight, we directly infer the uncertainty by analyzing the predicted image containing the defects in a single forward pass. This is a clear advantage over methods that require several forward propagations, like e.g., dropout-based uncertainty estimation [2] or Bayesian ensembling [3].

In the following section, we start by describing the neural network architectures and the data that we use throughout our experiments. With these networks and the data in mind, we developed our uncertainty estimation methods, which will be discussed afterwards.

2.1. Network Architectures

We propose to use a u-net variant that was introduced for various medical image segmentation tasks [28, 29]. This architecture won the International Symposium on Biomedical Imaging (ISBI) cell tracking challenge in 2015, and is known to work well even with relatively little training data. It is basically an autoencoder with skip-connections from the contracting path to the expansive path, enabling the network to localize. The networks are trained to minimize the mean squared error (MSE) between the predicted and the actual target masks. This loss will also serve as a reference to assess the quality of the uncertainty estimation later on. For a more detailed description we refer to the original paper [28].

In this paper, we focus on the two specific architectures that we introduced in [18]. One was designed to be extremely compact in terms of number of parameters, while still delivering satisfying results. It consists of just 16 filters in the first (single channel) layer, and has three layers in each the contracting and the expansive path, resulting in only 109 thousand weights. The second, larger architecture was designed to maximize the performance on our development data, and no restrictions in terms of network capacity were considered. Surprisingly, a very similar architecture as the compact one turned out to work best. It has also just 16 filters in the first (single channel) layer, five layers in each the contracting and the expansive path, and about 1.8 million weights.

Interestingly, the specific training setup did not make a difference to the outcome of the architecture search. That is, for both the end-to-end and the hybrid approach the exact same architectures were found to be most useful. The difference between the end-to-end and the hybrid approach lies in the input data for the network, which we describe in the following section.

2.2. Data

Since the vast amounts of data that is typically required for deep learning approaches is not available, we created a data set comprising simulated data for 2D. We started by randomly placing up to five square-shaped defects with side lengths between two and six pixels in order to end up with a binary target mask. The corresponding surface temperature measurements were then simulated by assuming adiabatic boundary conditions (i.e. there is no heat flow into or out from the specimen) [1]. These surface temperature measurements were later used as input to the end-to-end approach. For our hybrid method, the virtual waves from the temperature measurements were computed by ADMM in conjunction with the Abel transformation [9]. As a result, each sample is represented by three images: the target mask, the temperature measurements, and the virtual waves. All of them are single channel images with a dimensionality of 256 by 64 pixels. In Fig. 2, we can see an example of this data in the form of an initial temperature signal (representing the 2D target), a simulated surface temperature signal (corresponding to 2D thermographic data) and a simulated ideal virtual wave signal computed from it.

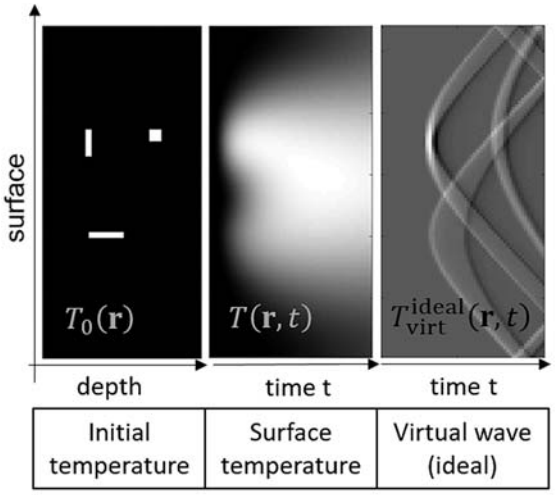

Fig. 2. Example of a target mask, the corresponding temperature measurement and virtual waves.

Additionally, we computed ten different versions of each sample, representing SNRs (i.e., initial temperature vs. noise) from -20 dB to 70 dB in 10 dB steps. The noise was added to the temperature before calculating the virtual waves. This is supposed to increase robustness against changes in the level of SNR during training. Furthermore, having multiple versions per sample is useful for a more detailed

evaluation. This data is available to the public, and for a more detailed description of it we like to refer to [1].

We divided this data into three non-overlapping subsets as follows: Our **training data** consisted of 8,000 samples in ten versions for different SNRs, thus comprising a total of 80,000 samples. We normalized the samples to have zero mean and unit standard deviation using only the training data. For development and validation purposes, we always used the same 1,000 samples. Considering the ten different versions, we had 10,000 samples in our **validation data**. A **test dataset** was unseen in every regard and also normalized based solely on the training data. It also consisted of 1,000 samples in ten versions for different SNRs, comprising 10,000 samples in total.

2.3. Uncertainty Estimation Methods

While developing our methods, we considered several requirements that are relevant for real-world use-cases. First and foremost, the estimate should highly correlate with the true MSE. Furthermore, it should be suitable for real-time scenarios and deployment on systems with restricted computational capabilities, e.g., edge devices.

Our first intuition was to compute the MSE on an estimation of the ground truth (i.e., the defects). The better this estimate, the higher the correlation to the true MSE, i.e. computed with respect to the actual ground truth. For this, we started by defining a mask operator $\mathbf{g}_\theta : \mathbf{X} \in \mathbb{R}^{M\times N} \rightarrow \{0,1\}^{M\times N}$ that will be applied on the output of the neural network, where

$$[\mathbf{g}_\theta(\mathbf{X})]_{ij} = \begin{cases} 1 & \mathbf{X}_{ij} > \theta,\ 0 < \theta \leq 1 \\ 0 & \text{else} \end{cases} \qquad (2)$$

With this mask operator, we then compute two binarized predictions $\widehat{\mathbf{X}}_{\text{sens}} = \mathbf{g}_\alpha(\widehat{\mathbf{X}})$ and $\widehat{\mathbf{X}}_{\text{conf}} = \mathbf{g}_\beta(\widehat{\mathbf{X}})$, where α and β are set to 0.01 and 0.95, respectively. This leads to $\widehat{\mathbf{X}}_{\text{sens}}$ being highly sensitive with respect to the magnitude of each output pixel. $\widehat{\mathbf{X}}_{\text{conf}}$ on the other hand involves only high confident pixels, i.e. with high magnitude. In Fig. 3, we can see some examples of this binarization step. These two binarized masks then serve as the basis for our uncertainty estimates.

For our 1st method, we start by evaluating the MSE of the predicted image $\widehat{\mathbf{X}}$ on only the non-zero pixels of a masked version of $\widehat{\mathbf{X}}$. This is the numerator in the uncertainty estimate that we define as

$$UCRT_1 = \frac{\text{mse}\left(\widehat{\mathbf{X}} \circ \widehat{\mathbf{X}}_{\text{sens}}, \widehat{\mathbf{X}}_{\text{sens}}\right)}{1 + \frac{||\widehat{\mathbf{X}}_{\text{conf}}||_0}{||\widehat{\mathbf{X}}_{\text{sens}}||_0}}, \qquad (3)$$

where ∘ defines the Hadamard product used to apply the mask pixel per pixel. The L^0-norm is used to denote the number of non-zero pixels of the corresponding mask. We use it to compute the ratio of the number of non-zero pixels of the confident to the sensitive binary masks. This ratio $\in \mathbb{R}[0,1]$ tends towards 1 as the predictions become more accurate, since every pixel in the sensitive mask is also an element of the confident mask. In contrast, as the SNR decreases and the predictions become ever more deteriorated, this ratio tends towards 0. As a result, the denominator $\in \mathbb{R}[1,2]$ scales the initial estimate of the MSE.

For our 2nd method, we take the relative estimated area of the defects into account as follows

$$UCRT_2 = \left(1 - \frac{||\widehat{\mathbf{X}}_{\text{conf}}||_0}{||\widehat{\mathbf{X}}_{sens}||_0}\right)\frac{||\widehat{\mathbf{X}}_{\text{sens}}||_0}{n}, \qquad (4)$$

where n denotes the number of pixels of the output image. Notice, that the ratio of the number of non-zero pixels of the confident to the sensitive masks is again utilized as a scaler.

In order to be able to assess the performance of our approaches relative to another common approach, we introduce our baseline in the following Section.

2.4. Uncertainty Estimation Baseline

Where appropriate, we compare our method with an ensemble based approach similar to the one presented by [30]. Here, they model the prediction by means of a Gaussian mixture model. The uncertainty measure is derived by calculating the mean of the Kullback-Leibler divergence of the individual mixture components with respect to the ensemble distribution. However, since thermographic imaging differs significantly from the proposed regression setting, we do not explicitly model the predictions to be Gaussian. Instead, we derive the uncertainty measure rather from an empirical estimate of the ensemble variance. That is, we take the mean of the pixelwise variances as our uncertainty baseline, and refer to it as UCRT_ENS for the remainder of this paper. The MSE of the ensemble – which consists of five models in our scenario – is computed on the averaged predictions of it.

3. Experiments

In this section, we present the results of three conducted experiments, each designed to shed light on a specific aspect of our uncertainty estimations. First, we will assess how our uncertainty estimates correlate with the actual loss for different architectures and both the end-to-end and the hybrid approach. Second, we will demonstrate that our uncertainty estimates generalize to unseen conditions in terms of SNR. Third, we consider a practical use-case scenario and utilize the uncertainty estimations for rejection.

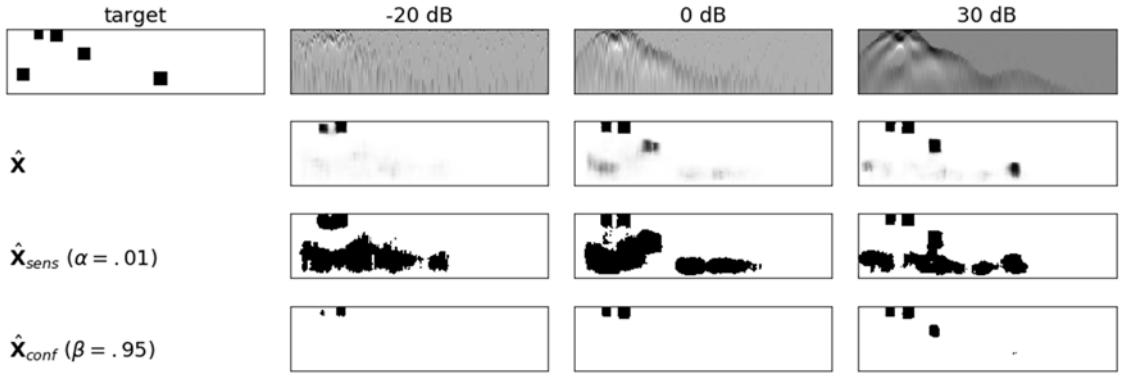

Fig. 3. 1st row: target (representing defects) and input images (virtual waves) for the neural network under several SNR conditions. 2nd row: raw output of the neural network. It can be seen how the quality of the output suffers as the SNR decreases. In the 30 dB scenario, only the most left defect deep inside the specimen is missed, whereas under the -20 dB condition only the two defects closest to the surface appear in the output of the neural network. 3rd row: the highly sensitive binarized mask used to compute the MSE. 4th row: another binarized mask based on high confident pixels.

3.1. Correlation with Loss

With this experiment, we want to show the high correlation between the actual loss (MSE) and the uncertainty estimates. For this, we take the two different approaches introduced in [1]. The first approach is end-to-end, where we feed the temperature measurements directly to the neural networks. This approach has the least computational load, but at the same time the highest task complexity. The second approach is hybrid, where we use virtual waves instead of temperature measurements as input. Compared to the end-to-end approach, some of the computational burden is taken away from the neural network, which yields better results. By combining them with each the compact and the large architecture (see Section 2.1), we get results from four different setups in total.

In Fig. 4, we can see a comparison between our uncertainty estimates and the actual loss for the four different setups.

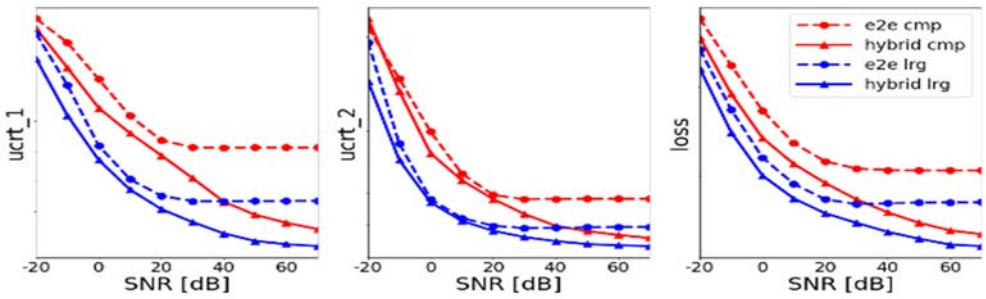

Fig. 4. Uncertainty estimations according to the two approaches compared to actual loss (MSE). The monotonic decrease of the loss corresponding to increased SNRs is well reflected in all scenarios.

This was done by computing the average loss and uncertainties for all SNR conditions on the unseen test data (see Section 2.2). The large and compact architectures are referred to as *lrg* and *cmp*, respectively. As can be seen, there is a high correlation between the uncertainty estimates and the MSE. Also the relative performance in terms of loss between each setup is represented by the uncertainty estimates, at least to a certain extent. Even the fact that the end-to-end models (see *e2e cmp*, *e2e lrg*) are not producing ever better results from approximately 30 dB SNR upwards is well reflected in both uncertainty estimations.

Intuitively, both uncertainty estimates seem to give satisfying results, and UCRT_1 seems to better reflect the relative performance between the setups.

3.2. Performance on Unseen SNR Conditions

With this experiment, we want to investigate the behavior of our uncertainty estimation methods under previously unseen SNR conditions compared to our baseline (see Section 2.4). In order to do so, we cannot re-use the models from the previous experiment, but have to create a new setup. Therefore, we train two neural networks, a full SNR range model and a high SNR range model using the large architecture (see Section 2.1). The high SNR range model is expected to perform worse on data corresponding to untrained SNR conditions, and will be helpful to determine the capabilities of the methods. We do the same thing for the baseline (see Section 2.4), except that we need to train five models for each ensemble.

The specific training regime for the two networks and the baseline is as follows. For training and validating the full SNR range models, we use all available ten different versions of each sample, representing SNRs from -20 dB to 70 dB in 10 dB steps. All in all, we use 27,000 and 10,000 samples for training and validation of the full range models, respectively. For training and validating the high SNR range models, we use just seven SNRs, ranging from 10 dB to 70 dB in 10 dB steps. All in all, we use 18,900 and 7,000 samples for training and validation of the high range model, respectively.

However, the unseen test data is identical to the original setting (see Section 2.2), and is normalized according to the training data respectively. As with the previous experiment, we compute the average loss and uncertainties for each SNR condition on the test data.

We start by showing a weakness of the baseline (see Section 2.4) in Fig. 5, where we plot the loss and uncertainty of the high SNR and the full SNR range models. For better comparison, we normalize the loss and uncertainty each to be between 0 and 1. As can be seen, the decreased loss corresponding to increased SNRs is well reflected for the high range model, but not for the full range model. This result seems counterintuitive at first, since the baseline performs well on unseen, and fails on previously seen conditions. The reason for that is, that even though the predictions are more corrupted under very low SNRs, they are getting more similar within the ensemble. This in turn leads to a relatively low variance of pixel values, which is reflected in the erroneous outcome of the baseline.

In Fig. 6, we can see that contrary to the baseline, both our uncertainty methods yield results that better reflect the actual loss curves.

3.3. Utilizing Uncertainty to Reject

In this experiment, we utilize the uncertainty estimates to implement a reject option. That is, we want to be able to reject samples with an expected loss above a specific threshold. To this end, we formulate a 2-class classification problem to distinguish *reject* and *no reject* categories. In order to end up with these categories, we fixed the threshold for the loss so that 30 % of the validation data results fall into the category *reject*.

The corresponding uncertainty threshold needs then to be fixed for each method and each model separately. From a practical point of view, different requirements may need to be fulfilled. Therefore, we decided to evaluate according to several uncertainty thresholds as follows. The uncertainty-based rejection should give maximum accuracy, equal error rate, a false positive rate of 5 %, and a false positive rate of 1 %. These four requirements lead to four different uncertainty thresholds, hence four evaluation results.

We present in Table 1 the results of the different scenarios on the validation data for all three methods. For each metric (precision, recall, f1-measure, and accuracy), the results on the full range model and the high range model are listed on the left and right column, respectively.

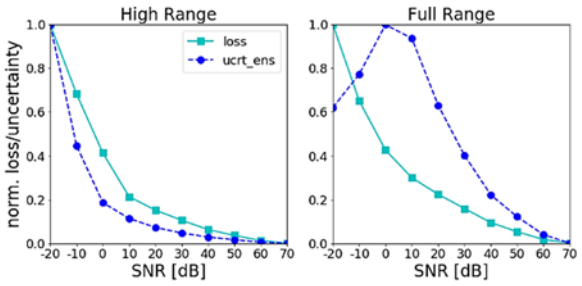

Fig. 5. Loss and uncertainty of the ensemble approach. Full Range: trained with all SNR conditions from [-20 to 70] dB; High Range: trained only with a subset ranging from [10 to 70] dB; This method reflects the decreased loss corresponding to increased SNRs only for the high range model.

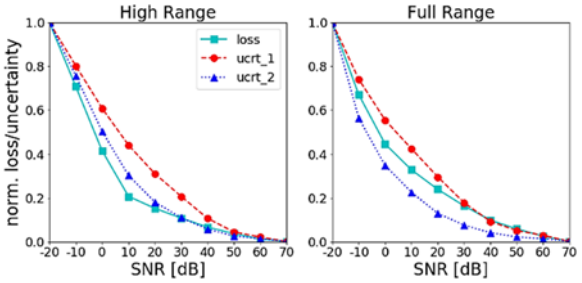

Fig. 6. Loss and uncertainty of our approaches. Compared to the ensemble approach in Fig. 5, both methods better reflect the decreased loss corresponding to increased SNRs.

Table 1. Results of the uncertainty based rejection on the validation set. UCRT_2 consistently outperforms the other methods.

	Precision		Recall		F1-Measure		Accuracy	
	full	high	full	high	full	high	full	high
Evaluation @Maximum Accuracy								
UCRT_ENS	.70	.84	.68	.73	.69	.78	.82	.88
UCRT_1	.73	.79	.60	.71	.66	.75	.81	.86
UCRT_2	.82	.85	.78	.80	.80	.82	.88	.90
Evaluation @Equal Error Rate								
UCRT_ENS	.64	.72	.80	.86	.71	.78	.80	.86
UCRT_1	.63	.69	.80	.84	.70	.76	.80	.84
UCRT_2	.75	.76	.87	.88	.80	.82	.87	.88
Evaluation @5% False Positives								
UCRT_ENS	.80	.86	.45	.70	.58	.77	.80	.88
UCRT_1	.79	.83	.44	.59	.57	.69	.80	.84
UCRT_2	.86	.87	.71	.76	.78	.81	.88	.89
Evaluation @1% False Positives								
UCRT_ENS	.89	.95	.19	.42	.31	.58	.75	.82
UCRT_1	.84	.88	.13	.17	.22	.29	.73	.75
UCRT_2	.95	.96	.46	.54	.62	.69	.83	.86

Interestingly, the full SNR range scenario seems to be more challenging for all uncertainty based rejection methods. This is due to the fact that the models trained only with high SNR ranges produce higher losses under the unseen, low SNR conditions (-20, -10 and 0 dB) compared to the models trained with the full

SNR range. Therefore, the average of the actual loss (MSE) of the 30 % that we categorize as reject tends to be higher compared to the full range scenario. This increased distance makes it easier to identify the samples to reject. The rejection based on method UCRT_2 seems to be the least affected by this phenomenon.

Additionally, the rejection based on UCRT_2 consistently outperforms the other methods in all regards, and on both the full and high range models. This suggests that for our specific task at hand, there is a correlation between the area of defects and the loss, since UCRT_2 factors this in.

The rejection based on method UCRT_1 gives results that are comparable to the baseline method UCRT_ENS when evaluated at maximum accuracy and equal error rate. When evaluated at a limited false positive rate, UCRT_1 gives inferior results compared to UCRT_ENS, especially for the high range model.

Finally, we keep the uncertainty thresholds and evaluate on the unseen test data set (see Section 2.1). In Table 2, the results of the different scenarios are listed for all three methods. In general, the test data results are very similar to the validation data results, suggesting good generalization capabilities. Consistently, the rejection based on UCRT_2 outperforms the other methods in all regards. Notice, that since we took the loss thresholds for the *reject* categorization from the validation data, the relative size of the class of interest *reject* is not exactly 30 %, but very close with 31.5 % and 30.6 % for the full and high SNR range scenarios, respectively.

Table 2. Results of the uncertainty based rejection on the unseen test set. In general, the results are very similar to the validation data results, and UCRT_2 consistently outperforms the other methods.

	Precision		Recall		F1-Measure		Accuracy	
	full	high	full	high	full	high	full	high
Evaluation @Maximum Accuracy								
UCRT_ENS	.69	.85	.69	.73	.69	.79	.80	.88
UCRT_1	.75	.80	.60	.71	.67	.75	.81	.86
UCRT_2	.84	.85	.79	.81	.81	.83	.89	.90
Evaluation @Equal Error Rate								
UCRT_ENS	.63	.73	.80	.86	.71	.79	.79	.86
UCRT_1	.65	.70	.80	.85	.72	.77	.80	.84
UCRT_2	.76	.77	.88	.90	.82	.83	.87	.89
Evaluation @5% False Positives								
UCRT_ENS	.79	.87	.47	.71	.59	.78	.79	.88
UCRT_1	.80	.84	.45	.59	.58	.70	.79	.84
UCRT_2	.87	.88	.71	.78	.78	.82	.88	.90
Evaluation @1% False Positives								
UCRT_ENS	.92	.96	.19	.43	.32	.59	.74	.82
UCRT_1	.86	.89	.12	.19	.22	.31	.72	.74
UCRT_2	.96	.97	.46	.55	.62	.70	.82	.86

4. Discussion

The consistency of the results on unseen data, unseen SNR conditions, different network architectures, and input data suggests good generalization capabilities. A possible direction of future work would be to assess our methods in different contexts. We think it could be useful for tasks in general, where the output can be converted into one or several binary masks, like semantic- or instance segmentation. Although, we consider it necessary to calibrate the α and β and hyperparameters to maximize the correlation with the loss first. The scope of application could even be extended towards a more general quality measurement that can also be applied to the results of the purely numerical methods.

Another interesting research direction would be dynamic model selection, where the best out of at least two models is dynamically selected based on their uncertainty estimates. Nevertheless, first experiments in that direction suggest that the difference in model performance needs to be substantial in order to reliably select the actual best model for each sample. Again, calibrating α and β might allow for a more fine-grained decision making. Therefore, implementing and evaluating an automatic calibration seems to be the logical next step to move our research forward.

5. Conclusion

In this paper, we presented two light-weight methods for uncertainty estimation of thermographic imaging results, and reported on three experiments. First, we demonstrated the high correlation of both methods with the actual loss (MSE) of different architectures (compact and large) and approaches (end-to-end and hybrid). Second, we intentionally limited the range of data in terms of SNRs during training in order to assess the performance on unseen, very low SNR conditions. While the ensemble-based baseline failed to give acceptable results on very low SNR conditions, both our methods exhibited the desirable behavior. Third, we showed how to use uncertainty to reject predictions that are expected to exceed a specific loss threshold.

In all conducted experiments, the ensemble based baseline was consistently outperformed by UCRT_2, and most often by UCRT_1. Another advantage of the proposed methods is, that just the result of a single forward pass is required, making them computationally less expensive than the ensemble-based baseline. Therefore, our uncertainty estimates are best suited for applications at the edge, even in real-time scenarios.

Acknowledgements

This work has been supported by Silicon Austria Labs (SAL), owned by the Republic of Austria, the Styrian Business Promotion Agency (SFG), the federal state of Carinthia, the Upper Austrian Research (UAR), and the Austrian Association for the Electric and Electronics Industry (FEEI); and by the University SAL Labs initiative of Silicon Austria Labs (SAL) and its Austrian partner universities for applied fundamental research for electronic based systems.

The financial support by the Austrian Federal Ministry of Science, Research and Economy and the National Foundation for Research, Technology and Development is gratefully acknowledged. Financial support was also provided by the Austrian research funding association (FFG) under the scope of the COMET program within the research project Photonic Sensing for Smarter Processes (PSSP) (contract number 871974). Parts of this work have been supported by the Austrian Science Fund (FWF), projects P 30747-N32 and P 33019-N.

References

[1]. P. Kovács, B. Lehner, G. Thummerer, G. Mayr, P. Burgholzer, M. Huemer, Deep learning approaches for thermographic imaging, *Journal of Applied Physics*, Vol. 128, Issue 15, 2020, 155103.

[2]. Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in *Proceedings of the 33rd Int. Conf. on Machine Learning*, New York, NY, USA: PMLR, Vol. 48, 2016, pp.1050–1059.

[3]. T. Pearce, F. Leibfried, A. Brintrup, M. Zaki, A. Neely, Uncertainty in neural networks: Approximately Bayesian ensembling, in *Proceedings of the 23rd Int. Conf. on Artificial Intelligence and Statistics*, Vol. 108, 2020, pp. 234-243.

[4]. B. Lehner, T. Gallien, Uncertainty estimation for non-destructive detection of material defects with u-nets, in *Proceedings of the 2nd Int. Conf. on Advances in Signal Processing and AI (ASPAI)*, 2020.

[5]. N. Ida, N. Meyendorf, Eds. Handbook of advanced nondestructive evaluation, *Cham, Switzerland: Springer*, 2019.

[6]. V. Vavilov, D. Burleigh, Infrared Thermography and Thermal Nondestructive Testing, *Cham: Springer Int. Publishing*, 2020.

[7]. S. Roointan, P. Tavakolian, K. S. Sivagurunathan, M. Floryan, A. Mandelis, S. H. Abrams, 3D dental subsurface imaging using enhanced truncated correlation-photothermal coherence tomography, *Scientific Reports*, Vol. 9, Issue 1, 2019, p. 16788.

[8]. V. P. Vavilov, D. D. Burleigh, Review of pulsed thermal NDT: Physical principles, theory and data processing, *NDT & E Int.*, Vol. 73, 2015, pp. 28–52.

[9]. G. Thummerer, G. Mayr, P. D. Hirsch, M. Ziegler, P. Burgholzer, Photothermal image reconstruction in opaque media with virtual wave backpropagation, *NDT & E International.*, Vol. 112, 2020, p. 102239.

[10]. P. Tavakolian, K. Sivagurunathan, A. Mandelis, Enhanced truncated-correlation photothermal coherence tomography with application to deep subsurface defect imaging and 3-dimensional reconstructions, *Journal of Applied Physics*, Vol. 122, Issue 2, 2017, p. 023103.

[11]. A. Mendioroz, K. Martínez, R. Celorrio, A. Salazar, Characterizing the shape and heat production of open vertical cracks in burst vibrothermography experiments, *NDT & E International.*, Vol. 102, 2019, pp. 234–243.

[12]. S. D. Holland, B. Schiefelbein, Model-based inversion for pulse thermography, *Experimental Mechanics*, Vol. 59, Issue 4, 2019, pp. 413–426.

[13]. P. Burgholzer, M. Thor, J. Gruber, G. Mayr, Three-dimensional thermographic imaging using a virtual wave concept, *Journal of Applied Physics*, Vol. 121, Issue 10, 2017, 105102.

[14]. P. Burgholzer, G. Mayr, G. Thummerer, M. Haltmeier, Linking information theory and thermodynamics to spatial resolution in photothermal and photoacoustic imaging, *Journal of Applied Physics*, Vol. 128, Issue 17, 2020, p. 171102.

[15]. G. Mayr, G. Stockner, H. Plasser, G. Hendorfer, P. Burgholzer, Parameter estimation from pulsed thermography data using the virtual wave concept, *NDT & E International*., Vol. 100, 2018, pp. 101–107.

[16]. L. Szilard, Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen, *Zeitschrift für Physik*, Vol. 53, Issue 11-12, 1929, pp. 840–856 (in German).

[17]. R. Landauer, Information is physical, *Physics Today*, Vol. 44, Issue 5, 1991, pp. 23–29.

[18]. P. Kovács, B. Lehner, G. Thummerer, M. Günther, P. Burgholzer, M. Huemer, A hybrid approach for thermographic imaging with deep learning, in *Proceedings of the 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 4277–4281.

[19]. G. Thummerer, G. Mayr, M. Haltmeier, P. Burgholzer, Photoacoustic reconstruction from photothermal measurements including prior information, *Photoacoustics*, Vol. 19, 2020, 100175.

[20]. B. Pan, S. R. Arridge, F. Lucka, B. T. Cox, N. Huynh, P. C. Beard, E. Z. Zhang, M. M. Betcke, Photoacoustic reconstruction using sparsity in curvelet frame, *arXiv arXiv:2011.13080*, 2020.

[21]. P. C. Hansen, Analysis of discrete ill-posed problems by means of the L-curve, *SIAM Review*, Vol. 34, Issue 4, 1992, pp. 561–580.

[22]. P. C. Hansen, Rank-deficient and discrete ill-posed inverse problems: Numerical aspects of linear inversion. Philadelphia, PA, USA: *SIAM Monographs on Mathematical Modeling and Computation*, 1998.

[23]. J. Eckstein, P. D. Bertsekas, On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators, *Mathematical Programming*, Vol. 55, Issue 1–3, 1992, pp. 293–318.

[24]. L. J. Busse, Three-dimensional imaging using a frequency-domain synthetic aperture focusing technique, *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, Vol. 39, Issue 2, 1992, pp. 174–179.

[25]. F. Lingvall, T. Olofsson, S. T., Synthetic aperture imaging using sources with finite aperture: Deconvolution of the spatial impulse response, *The Journal of the Acoustical Society of America*, Vol. 114, Issue 1, 2003, pp. 225–234.

[26]. J. A. Tropp, Just Relax: Convex Programming Methods for Identifying Sparse Signals in Noise, *IEEE Transactions on Information Theory*, Vol. 52, Issue 3, 2004, pp. 1030–1051.

[27]. J. A. Tropp, Greed is good: algorithmic results for sparse approximation, *IEEE Transactions on Information Theory*, Vol. 50, Issue 10, 2004, pp. 2231–2242.

[28]. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, *in Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI'2015)*, 2015, pp. 234–241.

[29]. M. D. Jenkins, T. A. Carr, M. I. Iglesias, T. Buggy, G. Morison, A deep convolutional neural network for semantic pixel-wise segmentation of road and pavement surface cracks, in *Proceedings of the 26th*

European Signal Processing Conference (EUSIPCO), Sep. 2018, pp. 2120–2124.

[30]. B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, *in Advances in Neural Information Processing Systems 30*, (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds.), *Curran Associates, Inc.*, 2017, pp. 6402–6413.