IEEE Access
Multidisciplinary ⁞ Rapid Review ⁞ Open Access Journal

Received January 24, 2021, accepted February 22, 2021, date of publication March 1, 2021, date of current version March 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3062860

Benchmarking Deep Learning Models for Automatic Ultrasonic Imaging Inspection

JIAXING YE, (Member, IEEE), AND NOBUYUKI TOYAMA
National Metrology Institute of Japan, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8568, Japan
Corresponding author: Jiaxing Ye (jiaxing.you@aist.go.jp)

ABSTRACT The success of deep neural networks in carrying out a wide variety of cognitive tasks also raised expectations regarding the advent of AI for the ultrasonic testing (UT) data interpretation in the Non-destructive evaluation (NDE) field. Though it is a growing area of research, we identify two main barriers that hinder research in the field: the lack of real-world, annotated datasets accessible to the public and the scarceness of benchmarked performance of the state-of-the-art deep learning models. To address these issues, we first introduce a new dataset called "USimgAIST" which contains more than 7000 real ultrasonic inspection images with both normal cases and defective ones from 17 types of flaws. Using the dataset, we performed a comprehensive evaluation of representative deep learning models. Through the study, we expect to validate whether existing AI models can achieve *human-level* ultrasonic image understanding for defect characterization. Besides, all detailed benchmarking comparisons, including defect detection accuracy, model complexity, memory usage, and inference time, are shown. We hope this study exhibits an overview of performances of advanced learning models working for ultrasonic image analysis and lays the groundwork for prospective practitioners to compare their methods and results fairly.

INDEX TERMS Anomaly detection, computer vision, deep learning, non-destructive evaluation, ultrasonic inspection.

I. INTRODUCTION

All civil infrastructures, i.e., bridges, dams, and airports, have finite life spans and start to degrade since they are put into service. As time goes, deteriorations, such as corrosion, fatigue, erosion, wear, and overloads, will continue until the structures are no longer fit for their intended use. Among all issues related to infrastructure safety management, condition inspection is the foremost one since it is the decision-making stage for any further process. To characterize different types of damages, a couple of approaches for non-destructive evaluation (NDE) techniques had been developed, such as Radiographic testing (RT) and ultrasonic testing (UT) [2]. Those techniques had been applied to assess the health condition of structures without interrupting their continued usefulness or serviceability [3].

In recent years, the availability of advancement in information technologies, i.e., cheaper massive storage and high-speed wireless networks, greatly facilitate the vast collection of NDE data. Moreover, the rapid progress of deep neural networks (a.k.a. Deep learning) have made computerized interpretation of NDE data more achievable than ever before. In order to effectively exploit massive sensor data to distill crucial information for NDE with fewer/no expensive human participation, there has been increasing interest in the application of artificial intelligence techniques in NDE.

In this study, we focus on the topic of computer-aided interpretation of ultrasonic inspection images, which had been regarded as a mainstream NDE approach with high sensitivity to most material damage and superior proficiency in determining the defect location and size [4]. The general principle of ultrasonic inspection is that an electrical pulser is used to trigger an ultrasonic signal that propagates through the target object in the form of waves; once defect being encountered, part of the wave energy will reflect back to the surface. By collecting the echo waves, the defects can be discerned. As soon as it was possible to save and load ultrasonic inspection into a computer, researchers have built systems for automated analysis of non-destructive evaluation data [5]–[7]. The objective is twofold: From the efficiency aspect, the current NDE data interpretation relies on the inspector's expertise, the process can be time-consuming as

The associate editor coordinating the review of this manuscript and approving it for publication was Jinjia Zhou.

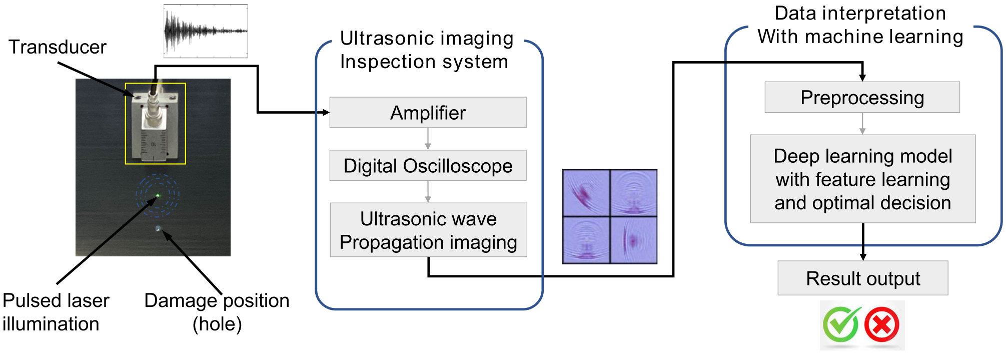

Transducer
Pulsed laser illumination
Damage position (hole)
Ultrasonic imaging Inspection system
Amplifier
Digital Oscilloscope
Ultrasonic wave Propagation imaging
Data interpretation With machine learning
Preprocessing
Deep learning model with feature learning and optimal decision
Result output

FIGURE 1. Chartflow of an ultrasonic NDE system with automatic data interpretation.

the workload increases. At the same time, the automatic inspection data analysis system is designated to alleviate the overwhelming workload of inspectors. Secondly, condition assessment is performed in a subjective manner based on the inspector's experience, making the results being vulnerable to human errors [8]. On the contrary, an AI-enabled interpreter is based on the collective intelligence that is distilled from the massive NDE database, which is anticipated to be an objective standard.

In broad terms, there are two approaches to NDE data analysis: model-driven and data-driven approaches [9]. Model-driven methods establish a high-fidelity physical model of the structure, usually by finite element analysis, and then establish a comparison metric between the model and the measured data from the real structure [10]. If the model is for a system or structure in normal (i.e., undamaged) condition, any departures indicate that the structure has deviated from the normal condition, and the damage is inferred. On the other hand, data-driven approaches also establish a model, but this is usually a statistical representation of the relationship between sensory data (vibration signal, acoustics, etc.) and system state indicators (OK or NG) without characterization of explicit knowledge of the physical behavior of the system [11]. More precisely, data-driven approach is the study of computational methods and algorithms for improving the performance of NDE data understanding by mechanizing the acquisition of knowledge from previous data collections. The design of a data-driven pattern analysis system requires careful attention to the following issues [12]: definition of pattern classes, pattern representation, feature extraction and selection, classifier design and learning, preparation of training and test samples, and performance evaluation. In recent years, a number of 'human-versus-machine' evaluations demonstrated that state-of-the-art data-driven machine learning systems are able to substitute the human role in various challenging tasks, such as visual content understanding and speech recognition, and thus the data-driven models for NDE applications had drawing plenty of research efforts lately. The objective of this paper is to perform a comprehensive comparative study upon the application of the latest deep learning approaches for automatic ultrasonic inspection image interpretation. A general processing flow of chart is presented in Fig 1, consisting of both sensing hardware and ultrasonic image data interpretation system. The significant contribution of this study is three-fold:

† One of the main challenges and hindrances to machine learning research for ultrasonic NDE data is the lack of real dataset that is accessible to the public. In this work, we introduce an open-access database – USimgAIST [1], consisting of over 7000 real ultrasonic inspection images with full annotations. The open dataset is assumed to lay the fundamentals for further studies and validations. The dataset is offered free of charge for non-commercial use only under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International [13]. We licensed the dataset for the purpose of academic research in a non-commercial context, also we allow third parties to copy and distribute the original dataset so long as the authors of the original are named.

† Based on the data collection, a series of comparative experiments had been conducted to evaluate state-of-the-art deep learning models for the task of automatic ultrasonic inspection image analysis. We present the detailed benchmarking results, which can be regarded as the baseline for future studies in the field of automatic ultrasonic inspection data interpretation. Moreover, the pre-trained deep learning model on our dataset can be used to establish an initialization point of the fine-tuning procedure to the weights of deep learning model. By continually updating the deep learning model on the task-specific dataset, the system is anticipated to achieve superior performance comparing to the case of training the network from scratch.

† We drew an extensive comparison of the critical specifications of each model, including architecture design, model size and processing speed. Those factors can be decisive when the model is deployed for real applications. We hope the review of the models can facilitate practitioners from industries.

The remainder of this paper is arranged as follows. In Section 2, we present a concise historical review of machine learning techniques employed for ultrasonic inspection data understanding from conventional models to nowadays deep learning. Section 3 describes the ultrasonic inspection image database, including the inspection device, test specimen, and data annotations. In section 4, we introduce the deep learning models under evaluation with a comprehensive comparison of model design and specifications. Section 5 presents the full benchmarking results in terms of data interpretation accuracies and efficiency. We finalize this study with a summary, an prospective discussion in Section 6.

II. RELATED WORK

The research aiming at building a human-like system to interpret ultrasonic inspection data autonomously has been a long-standing theme that can be traced back to half a century ago [14], [15]. The task can be reformulated as: building a mathematical model based on a massive collection of ultrasonic inspection images, known as "training data", in order to judge whether the particular flaw-induced patterns are presented in the observed data or not without explicit writing program to do so. Initial attempts to employ machine learning for NDE defect detection and classification focused on using simple neural networks (i.e., perception) to classify various types of NDE data [16]. As for the input ultrasonic signal, there are two groups: the raw echo signal and ultrasonic B-scan images. Accordingly, two different feature extraction approaches had been employed: Fourier and wavelet transform [17] had been introduced to process A-scan waves and co-occurrence features [18] had been investigated for ultrasound B-scan images as well.

Starting from the early 2000s, along with the increase in computational power, the used machine learning models have become more powerful. Many authors have reported favorable automatic ultrasonic results with advanced statistical pattern analysis approaches, such as sparse coding [19] and support vector machines [20]. Those research results confirmed that novel machine learning/pattern recognition techniques could substantially contribute to ultrasonic data interpretation.

Over the last decade, deep learning models have revolutionized many fields, including computer vision, natural language processing, speech recognition [21]. The significant advancements in deep learning models and computational hardware have facilitated more complex and powerful approaches that even reach human-level performance. In the meantime, the development of deep learning-enabled system for ultrasonic NDE data interpretation has emerged as an active research direction, and a series of articles had been published [23]. It is noteworthy that though the name of deep neural networks had been repeatedly mentioned [24], the most applied neural networks are limited to 4 layers [25], which is, indeed, NOT a deep learning solution. The primary factor accounted for this is that the ultrasonic inspection datasets used in the experimental validations are confined to small-scale so that there is no significant benefit to employ deep neural networks for generating efficient feature representations. From the viewpoint of machine learning, we may have a deep understanding of the limitations of current studies. Too little training ultrasonic inspection dataset will result in a poor approximation due to biased sampling to the whole pattern space of ultrasonic inspection images. On the contrary, choosing a powerful model such as a deep learning model with hundreds of layers, in turn, will likely overfit the training data and perform worse when encountering unseen data in the real application. To alleviate the situation, more latest works tried to increase the dataset's size by using data simulation data or augmentation [26]. In this study, we choose the most efficient but straightforward way to collect more real ultrasonic inspection image data. We hope to exploit whether a larger dataset facilitates the deep learning models to tune their many parameters to come up with somewhat generalizable models.

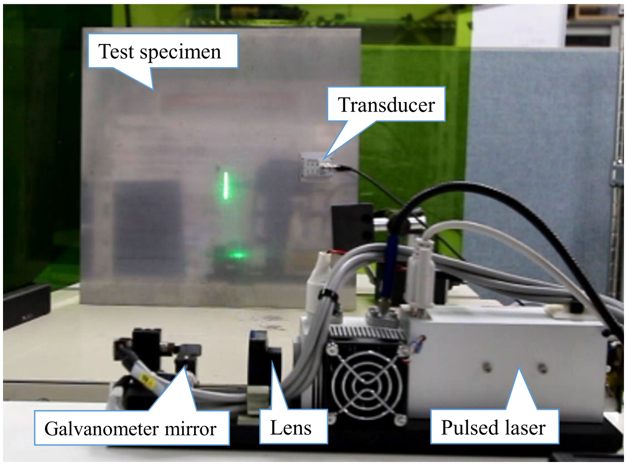


FIGURE 2. The ultrasonic inspection device used for database generation.

III. ULTRASONIC IMAGING INSPECTION DATASET

In this section, we introduce our dataset of real ultrasonic inspection images with annotations. The objective of this research is to investigate the performance of state-of-the-art computer vision techniques for replacing human roles in ultrasonic inspection image interpretation. Particularly, In the era of deep learning, openly accessible, fully annotated dataset addressing specific task is regarded as the foundation. Without domain-specific datasets, machine-learning algorithms would have no way of learning how to characterize the concealed knowledge available in the data.

Our research laboratory has a long research history regarding ultrasonic inspection device development and generated solid research records. We have previously developed a system for the visualization of ultrasonic wave propagation in general solid media [27], which employs a pulsed laser that scans an object for ultrasonic wave generation. Then a contact transducer is attached to the object specimen to capture a series of snapshots of the propagating waves. We have successfully applied it to the non-destructive inspection of various structural components. Fig.2 displays a photograph of the ultrasonic inspection imaging system, and key parts had been annotated.

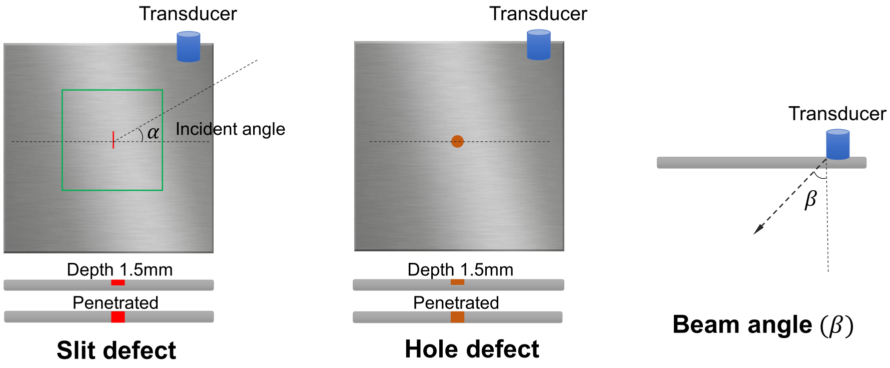


FIGURE 3. Demonstration of flaw types and sensor installation.

TABLE 1. Summary of settings of ultrasonic imaging inspection device.

Item	Settings
Probe frequency	1MHz
Beam angle β°	90
Pulse repetition frequency	500Hz
Incident angle α°	0, 22.5, 45, 67.5, 90

TABLE 2. Summary of the specimen specifications.

Specimen	Flaw type	Depth	Transducer side	Defect size(mm)
1 − 3	Hole	Penetrated	Front	$\phi1$ / $\phi3$ / $\phi5$
4 − 6	Hole	1.5mm	Front	$\phi1$ / $\phi3$ / $\phi5$
7 − 9	Hole	1.5mm	Back	$\phi1$ / $\phi3$ / $\phi5$
10 − 11	Slit	Penetrated	Front	5 / 10
12 − 14	Slit	1.5mm	Front	3 / 5 / 10
15 − 17	Slit	1.5mm	Back	3 / 5 / 10

A. ULTRASONIC INSPECTION IMAGE COLLECTION

During an ultrasonic inspection, several key parameter settings, such as probe frequency and pulse repetition frequency, can play an essential role in this analysis. To ensure the inspection data quality, we have adjusted the parameters, and the optimal settings are as presented in Tab. 1.

Moreover, The specifications of defects, such as shape, size, and depth of flaws, are variational factors affecting visual inspection. To evaluate the versatility of automatic imaging data interpretation methods, we prepared a batch of stainless steel plates with various types of flaws and details are shown in Fig. 3. In the evaluation, two types of defects were investigated: drill holes with diameters from 1mm to 5 mm and slits with lengths ranging from 3mm to 10mm. Tab. 2 summarized the specimen details. Notably, in addition to the 17 specimens with flaws, we further prepared one plate without damage to it. By including the ultrasonic images from the healthy specimen, we can efficiently exploit the key metric of false alarm rates (FAR) for anomaly inspection tasks. A laser scan is performed on the central region of 3mm thick specimens with 100mm by 100mm size (green zone) on both front/back sides of steel plates. Moreover, the incident angle of the transducer installation has been regarded as a critical parameter that significantly affects ultrasonic imaging patterns, especially when examining slit flaws. To exploit the robustness of the data interpretation system to those variations, we altered the incident angles from 0 to 90 with 22.5 intervals, and the resultant images had been stored in the dataset. Overall, we collected 7004 ultrasound images with a 3615/3389 split regarding without-defect and with-defects cases.

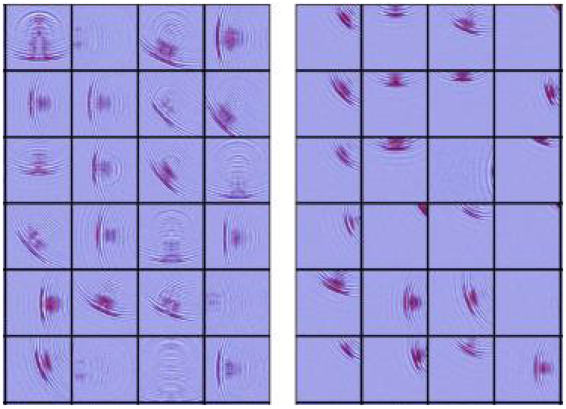


FIGURE 4. Examples of USimgAIST ultrasonic database (left: NG | right:OK).

In Fig. 4, we show two sets of ultrasonic inspection samples captured from healthy and defective specimens, respectively. By visual comparison, the flaws can be distinguished at a glance due to the evident flaw-induced reflection waves. However, from the viewpoint of an algorithm, it is quite difficult to define such reflection waves with explicit programming quantitatively. In this regard, we introduce cutting edge deep learning algorithms to verify how good they can replicate human performance in terms of making those judgments.

IV. REFERENCE DEEP LEARNING MODELS

Convolutional neural networks (CNN) denote a family of neural network architectures that is dedicated to processing matrix-shaped data, i.e., images [21]. This network structure was first proposed by Fukushima in 1980 [22]. However, the research and application toward neural networks are

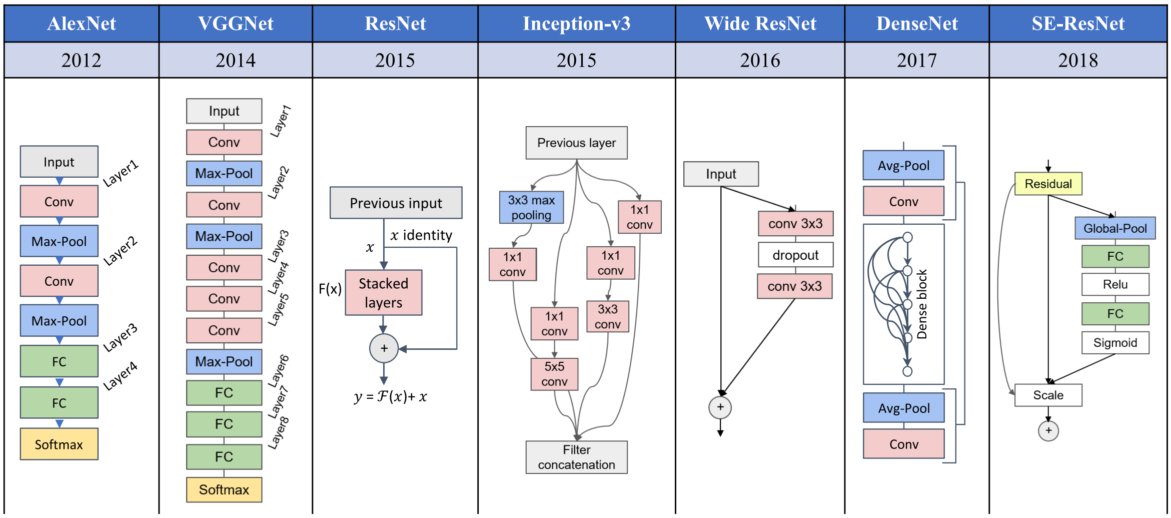


FIGURE 5. Representative design of deep learning models.

limited due to low computation hardware and insufficient datasets as then. Deep CNNs generally have complex architecture and time-intensive training phase that highly demand the parallel computation resources and larger memory. Since 2006, significant efforts have been made to tackle the CNN optimization problem, which revived the research in deep learning [28]. In this section, we elaborate on the most widely-applied architectures which would be employed to the evaluation for ultrasonic inspection image analysis through this study. In detail, our review consists of three parts: 1. A basic introduction to each model, 2. Summarization of the representative design of each model, 3. A comparison table of model specifications including detailed settings in the framework, parameter size and multiply accumulate operation (MACs) counts.

1) AlexNet [29]

In 2012, AlexNet significantly outperformed all the competitors using conventional hand-crafted features and won the difficult ImageNet challenge for visual object recognition called the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by reducing the top-5 error from 25% to 15.3%. It is the point in history where interest in deep learning increased rapidly. The AlexNet is composed of 5 layers of convolution operators followed by three fully connected layers. It is noteworthy that it also highlighted a batch of new techniques for computer vision research, such as the use of GPU to train a model, and many integral parts of *standard* deep learning models, including convolution kernel, max pooling, dropout, data augmentation, rectified linear unit (ReLU) activation and stochastic gradient descent (SGD) optimizer. After AlexNet, the exploration of novel architectures took off, and in the past five years, a trend emerged to build far deeper and wider models.

2) VGGNet [30]

VGGnet shows up as the runner-up at the 2014 version of the ILSVRC competition. The model came up with a deeper 16-layer structure and the appealing point is that it has a very uniform architecture with only 3 × 3 convolutions. It is still the most preferred choice in the community for extracting features from images. Moreover, because of the clarity in model design and superior reproducibility of coding, VGGNet has been used in many computer vision applications and challenges as a baseline feature extractor in nowadays.

3) ResNet [31]

It has been commonly acknowledged that with sufficient training data, increasing the network depth should increase the accuracy, as long as over-fitting can be carefully suppressed. Nevertheless, the problem does occur with increased depth; that is, the stimuli signal required to update the weights, which arises from the difference between the network output and ground-truth, becomes very small at the earlier layers and thus making the weights update intractable. The case is described as vanishing gradient. Residual networks are designated to tackle this issue by constructing the network through modules called residual models. The main idea of residual networks is to learn an additive residual function with respect to an identity mapping that is based on the preceding layer's inputs, accomplished by attaching an identity shortcut connection. Concretely, residual modules perform the following computation:

$$\begin{aligned} y_l &= h(x_l) + \mathcal{F}(x_l, W_{l,k|1\leq k\leq k}) \\ x_{l+1} &= f(y_l) \end{aligned} \tag{1}$$

where the input to the l-th residual module is denoted by x_l, represents its weights and biases, k is the number of layers in a module, $\mathcal{F}$ represents the residual function such as a stack of convolutional filters, f is the operation that follows element-wise addition, and h is an identity mapping of the form $h(xl) = xl$. An illustration of the residual module is shown in Fig 5. Residual networks have brought about some empirical success. Through the changes of the additive residual link mentioned above, ResNets were learned with

TABLE 3. Specifications of deep learning models under evaluation.

	AlexNet	VGGNet	ResNet-18	Inception-v3	Wide ResNet	DenseNet	SE-ResNet-50
	2012	2014	2015	2015	2016	2017	2018
Input Size	1 × 224 × 224	1 × 224 × 224	1 × 224 × 224	1 × 299 × 299	1 × 224 × 224	1 × 224 × 224	1 × 224 × 224
CONV Layer							
# CONV Layer	5	8	17	95	49	120	49
Depth	5	8	17	46	49	120	49
Kernel Size	3,5,11	3	1,3,7	1,3,5,7	1,3,7	1,3,7	1,3,7
Strides	1,2,4	2	1,2	1,2	1,2	1,2	1,2
FC Layer							
# of FC Layer	2	2	1	1	1	1	1
# of Channels	4096	4096	512	2048	1024	2048	2048
Complexity							
Parameters (M)	36.53	11.80	6.39	21.79	26.58	7.98	26.11
MACs(G)	0.72	7.63	1.82	5.75	11.46	2.88	3.90
Model Size(MB)	157.02 MB	190.29 MB	139.41 MB	199.56 MB	678.64 MB	126 MB	453.50 MB

network depth of as large as 152. It achieves better accuracy than VGGNet and GoogLeNet while being computationally more efficient than VGGNet. ResNet-152 achieves 95.51 top-5 accuracies on the ImageNet-1k challenge in 2015.

4) INCEPTION-v3 [33]

On top of the deeper networks, more complex designs have been exploited, aiming at improving model training efficiency and again reducing the number of parameters. Inception-v3 [33] is a representative example that had been proposed in 2015. The motivation of Inception-v3 is to extract multi-level visual features while keep relatively low computation complexity by employing parallel convolution kernels and efficient factorization. The detailed plot of inception model is presented in Fig. 5. With a batch of appealing components and tweaks, including RMSProp optimizer, Factorized and asymmetric convolutions for parameter reduction, and Auxiliary classifier, Inception-v3 significantly boosted the performance in ImageNet classification compared to the previous Inception models.

5) WIDE ResNet [34]

Stemmed from ResNet, Wide ResNet simply increased number of channels of ResNet. Though the change is subtle, it indeed produces superior performance compared to its predecessor. Another benefit brought by design is that increasing the width instead of depth makes training more computationally efficient. Due to its high performance and simplicity in computation, we add the method to the comparison list.

6) DenseNet [35]

So far as we reviewed, in standard conventional neural networks, the input image goes straight through multiple operation blocks to obtain higher-level features. In ResNet, identity mapping is added to facilitate the gradient propagation backward. In comparison, DenseNet established additional connections between each layer, and thus the 'collective knowledge' can be efficiently transmitted for better feature representation learning with respect to the given dataset. In addition to the superior feature learning ability, it is also noteworthy that the feature maps sharing between layers are beneficial from the viewpoint of computational efficiency and memory usage. Currently, DenseNet is regarded as one of the most applied models for various machine learning applications.

7) SQUEEZE-AND-EXCITATION (SE)-ResNet [36]

SE-ResNet is the latest winner of the ILSVRC 2017 classification challenge and achieved an impressive 25% relative improvement over the winning model of 2016 [36]. Upon the review to the representative deep learning model, the primary approaches are establishing various connections between stacking layers to boost accuracy. However, SE-Net considers how to evaluate the importance of the feature maps that learned through model training. To this end, SE-Net attempted to exploit global information to emphasize informative features and suppress less useful ones selectively. In other words, SE-Net block, shown in Fig.5, intrinsically introduces dynamics conditioned on the input, helping to enhance feature discriminability. The authors proposed two variations of SE-Net: SE-ResNet and SE-Inception, which were built on top of the successful models of ResNet and InceptionV3, respectively [36]. We choose the SE-ResNet for the evaluation due to the superior performance. We expect to see whether SE-ResNet can dominant the ultrasonic image understanding task again.

V. BENCHMARKING EXPERIMENTS

A. IMPLEMENTATION DETAILS

In order to provide a uniform and rigorous benchmarking of deep learning models for ultrasonic images recognition, we exactly reproduce the same policies for experiments. The PyTorch package [38] is used for network processing with

TABLE 4. Computing resource details.

CPU	Intel Xeon Gold 6148 × 2
GPU	NVIDIA Tesla V100 SXM2
Memory	384GiB
Storage	1.6TB NVMe SSD

TABLE 5. Model training hyper parameters.

Batch size	64
Epoch #	20
Learning rate	0.0001
Momentum	0.9

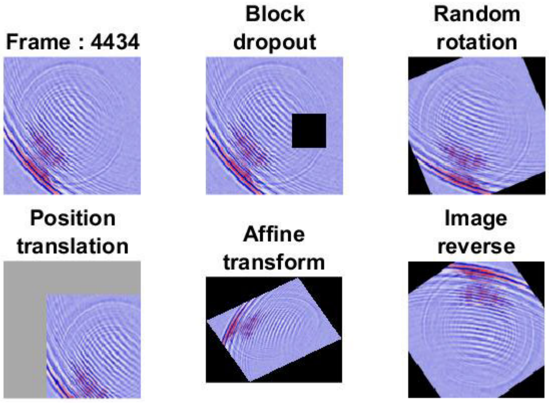


FIGURE 6. Demonstration of data augmentation methods.

supporting parallel computation libraries of cuDNN-v7.0 and CUDA-v10.1. Regarding the hardware specifications, we performed the experiments on AI Bridging Cloud Infrastructure (ABCI) [39], and the details are shown in Tab. 4.

In addition, we performed random search to determine the hyperparameters in model training. The resultant settings are presented in Tab. 5. As for optimization scheme selection, we applied root mean square propagation (RMSprop), which had been proven to be efficient and stably for parameter inference by automatically adjusting the learning rates.

All the other settings, such as weights initialization methods and dropout ratios, we assigned them by default values. In other words, we hope this benchmarking is for the off the shelf models without extensive tweaking.

To enhance the generalization power of models, we applied data augmentation to inject variabilities to original samples. For each ultrasonic inspection image, another five augmented images had been generated through block dropout, random rotation, position translations, affine transform, and image reverse. A demonstration of data augmentation is shown in Fig. 6.

At the training stage, we randomly selected 30% of all augmented images and fed them to deep learning models together with original image data. At the evaluation stage, we adopted a leave-one-specimen-out protocol; that is, at each iteration, we select one stainless steel plate specimen, and all the ultrasonic propagation imaging data generated from it would be assigned as testing data. In contrast, all the other images collected from all the other specimens will be split into training and validation sets on an 80%-20% ratio for each iteration. In the meantime, the undamaged specimen provided another 155 frames of ultrasonic images, which can further generate 775 augmented images to facilitate model training. The resultant 930 snapshots from the damage-free specimen are randomly inserted into training/testing/validation sets with a split of 65%-25%-10%. The evaluation is completed as the process repeats until all specimens had been iteratively tested. Finally, we obtain the predicted condition labels for the whole dataset. We normalized the images to 299 × 299 pixels for Inceptions-v3 model and 224 × 224 for all the other models considered.

B. RESULTS COMPARISON

In this part, we present results obtained throughout experimental validation. First, we introduce the evaluating metrics, which can be an essential part of assessing the prediction models. By comparing the prediction results with ground truth labels, we can derive four major statistics: True Positive (TP): defect images correctly detected. True Negative (TN): normal images classified to be non-defect. False Positive (FP): normal images incorrectly detected to have a defect. False Negative (FN): defect images incorrectly classified to be no-defect. Upon them, we introduce four metrics, i.e., precision (Pr), Recall (Re), Accuracy (ψ), and F-score (γ):

$$Pr = \frac{TP}{TP + FP}, \quad Re = \frac{TP}{TP + FN} \tag{2}$$

$$\psi = \frac{TP + TN}{P + N}, \quad \gamma = 2 \cdot \frac{Pr \cdot Re}{Pr + Re} \tag{3}$$

Furthermore, Tab.2 manifests that the size and shape of flaws are varying significantly. In order to evaluate the robustness of the deep learning model to those variations, we further extracted the specimen-wise averaging upon the above key metrics. The computation details are as follows:

$$\bar{Pr} = \frac{1}{17} \sum_{p=1}^{17} Pr_p, \quad \bar{Re} = \frac{1}{17} \sum_{p=1}^{17} Re_p \tag{4}$$

$$\bar{\psi} = \frac{1}{17} \sum_{p=1}^{17} \psi_p, \quad \bar{\gamma} = \frac{1}{17} \sum_{p=1}^{17} \gamma_p, \tag{5}$$

where $p \in [1, \ldots, 17]$ is the induce of specimen. The models are ultimately benchmarked in terms of the $\bar{Pr}$, $\bar{Re}$, $\bar{\psi}$, $\bar{\gamma}$.

The full evaluation results are presented with Tab.6. By seeing the table, we can find several interesting points. First, we added some representative computer vision techniques to the comparison, including histogram of gradients (HoG) and gradients of local auto-correlations (GLAC). We re-implemented those systems as presented in [37] and put them into the benchmarking on the USimgAIST dataset. Since we followed the same settings at model training/testing stage, and thus the comparison is fair and meaningful. According to the results, we find the latest models, even without fine-tuning at all, can easily outperform the conventional

TABLE 6. Results of automatic ultrasonic image pattern classification by using machine learning methods.

Model	HoG [37]	LBP [37]	GLAC [37]	AlexNet [29]	VGGNet [30]	ResNet-18 [31]	Inception-v3 [33]	Wide ResNet [34]	DenseNet [35]	SE-ResNet-50 [36]
Precision $\bar{Pr}$ (%)	82.81 ±0.159	83.97 ±0.123	95.20 ±0.038	67.70 ± *0.152*	66.79 ± *0.154*	92.11 ± *0.120*	92.51 ± *0.090*	94.49 ± *0.067*	95.21 ± *0.062*	**95.35** ± *0.046*
Recall $\bar{Re}$ (%)	90.58 ±0.083	86.62 ±0.143	90.32 ±0.100	94.88 ± *0.054*	95.37 ± *0.037*	91.43 ± *0.106*	94.01 ± *0.059*	94.45 ± *0.057*	**95.82** ± *0.034*	95.36 ± *0.043*
Accuracy $\bar{\psi}$ (%)	83.37 ±0.097	83.77 ±0.119	91.06 ±0.052	86.96 ± *0.035*	86.67 ± *0.039*	95.36 ± *0.044*	96.08 ± *0.030*	96.08 ± *0.031*	**97.35** ± *0.20*	97.20 ± *0.018*
F1-score $\bar{\gamma}$ (%)	85.45 ±0.098	84.81 ±0.121	90.47 ±0.053	77.87 ± *0.097*	77.76 ± *0.101*	90.95 ± *0.091*	92.93 ± *0.057*	94.17 ± *0.034*	**95.33** ± *0.030*	95.19 ± *0.021*

TABLE 7. Efficiency comparison of deep learning models.

Model	AlexNet [29]	VGGNet [30]	ResNet-18 [31]	Inception-v3 [33]	Wide ResNet [34]	DenseNet [35]	SE-ResNet-50 [36]
Training speed (fps)	2588.53	573.90	570.84	409.42	126.46	304.75	569.12
Test speed (fps)	5798.95	1753.56	1806.47	1335.63	400.33	1028.63	1766.57

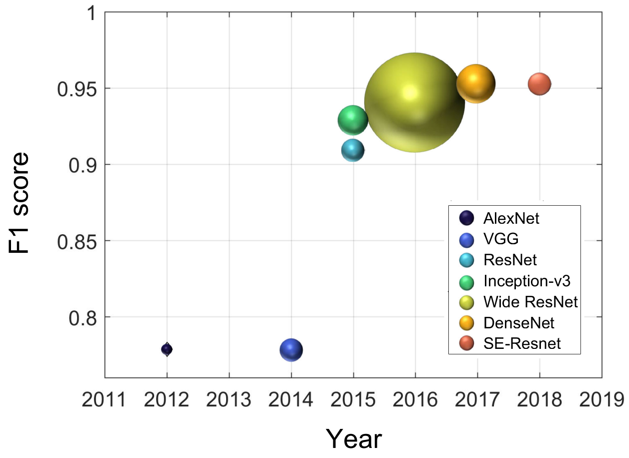


FIGURE 7. Summary chart reporting model performance (F1-score). The size of each ball corresponds to the model complexity for inference.

computer vision-based approaches. Although the dataset is not that big, the power of deep learning architectures can be clearly manifested. The second finding is that the evolution of deep learning models in the last five years is so significant that the latest ones like SE-Net and DenseNet are quite efficient compared to the predecessors in just two years. Moreover, in [37], we had proposed hand-crafted neural networks with residual module design, and after an extensive fine-tuning process, the averaged accuracy on 17 specimens reached to 95.68%. In contrast, the benchmarking tells us that both DenseNet and SE-Net models can achieved accuracies over 97% which outperformed the previous result with a clear margin. It should be noted that the deep learning models tested are just off the shelf ones without detailed tuning in a batch of hyperparameters. The results show that deeper stacking networks with efficient cross-layer connections are preferable to characterize rich discriminant from image data from the relatively small dataset. Besides, we also benchmarked the processing efficiencies of each model, and the results are shown in Tab. 7. Interestingly, the SE-ResNet, ranked the 2nd among all other models in classification accuracy, also maintained a 3rd place in processing efficiency in terms of the number of images being processed per second (fps) at testing stage. The high computation speed can be appealing merit when the model is considered to be deployed to practical applications. Finally, we conclude the benchmarking with Fig. 7, which manifests both model accuracy comparisons and computation efficiencies.

VI. CONCLUSION

Automatic data interpretation for ultrasonic imaging inspection garnered a lot of research interests recently, and plenty of research articles had been presented. In this paper, we contribute two aspects to the research field. First, we introduce the USimgAIST dataset, which consists of 7004 real ultrasonic inspection images collected from 18 stainless steel plates. Meanwhile, the data annotations are released together with the ultrasonic images. Moreover, we benchmarked the dataset by using state-of-the-art deep learning models. The results can be regarded as the baseline to facilitate the evaluation and comparison of methods for prospective studies. In addition, there are numerous possible ways to take advantage of the dataset, some of which include: Compiling a replicable survey comparing various approaches proposed in recent research papers and combining the presented dataset with private datasets to validate novel learning schemes of transfer learning and semi-supervised learning. Moreover, for practical development, the pre-training deep learning model obtained from our dataset can be used to initialize the weights of network. By continually updating the weights on the individual dataset, the resultant model is anticipated to achieve superior performance comparing to the one that trained from scratch. We hope the dataset together with benchmarks can open the path to new research in the field of automatic ultrasonic inspection image interpretation for non-destructive evaluation.

ACKNOWLEDGMENT

This article is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

[1] USimgAIST. *A Public Dataset for Research on Computer-Vision-Based Pattern Investigation for Ultrasonic Wave Propagation Images in Non-Destructive Testing (NDT)*. Accessed: Jun. 2020. [Online]. Available: https://sites.google.com/site/yejiaxingweb/projects

[2] S. K. Dwivedi, M. Vishwakarma, and P. A. Soni, "Advances and researches on non destructive testing: A review," *Mater. Today: Proc.*, vol. 5, no. 2, pp. 3690–3698, 2018.

[3] A. F. Grandt, Jr., *Fundamentals of Structural Integrity: Damage Tolerant Design and Nondestructive Evaluation*. Hoboken, NJ, USA: Wiley, 2003.

[4] J. Blitz and G. Simpson, *Ultrasonic Methods of Non-Destructive Testing*, vol. 2. Springer, 1995.

[5] R. T. Chin and C. A. Harlow, "Automated visual inspection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-4, no. 6, pp. 557–573, Nov. 1982.

[6] S. W. Doebling, C. R. Farrar, M. B. Prime, and D. W. Shevitz, "Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: A literature review," Los Alamos Nat. Lab., Los Alamos, NM, USA, Tech. Rep. LA-13070-MS, 1996.

[7] X. Gros, *NDT Data Fusion*. Amsterdam, The Netherlands: Elsevier, 1996.

[8] M. Bertovic, "Human factors in non-destructive testing (NDT): Risks and challenges of mechanised NDT," Ph.D. dissertation, Tech. Univ. Berlin, Berlin, Germany, 2015.

[9] C. R. Farrar and K. Worden, *Structural Health Monitoring: A Machine Learning Perspective*. Hoboken, NJ, USA: Wiley, 2012.

[10] J. D. Achenbach, "Quantitative nondestructive evaluation," *Int. J. Solids Struct.*, vol. 37, nos. 1–2, pp. 13–27, 2000.

[11] K. L. Rens, T. J. Wipf, and F. W. Klaiber, "Review of nondestructive evaluation techniques of civil infrastructure," *J. Perform. Constructed Facilities*, vol. 11, no. 4, pp. 152–160, Nov. 1997.

[12] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.

[13] *Creative Commons License (CC BY-NC-SA 4.0)*. Accessed: Nov. 2013. [Online]. Available: https://creativecommons.org/licenses/by-nc-sa/4.0/

[14] M. B. Fenton, S. L. Jacobson, and R. N. Stone, "An automatic ultrasonic sensing system for monitoring the activity of some bats," *Can. J. Zool.*, vol. 51, no. 2, pp. 291–299, Feb. 1973.

[15] J. L. Rose and G. P. Singh, "A pattern recognition reflector classification feasibility study in the ultrasonic inspection of stainless steel pipe welds," *Brit. J. Non-Destructive Test.*, vol. 21, no. 6, pp. 308–311, 1979.

[16] S. W. Lawson and A. Graham Parker, "Automatic detection of defects in industrial ultrasound images using a neural network," *Proc. SPIE*, vol. 2786, pp. 37–47, Aug. 1996.

[17] M. C. Robini, I. E. Magnin, H. Benoit-Cattin, and A. Baskurt, "Two-dimensional ultrasonic flaw detection based on the wavelet packet transform," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 44, no. 6, pp. 1382–1394, Nov. 1997.

[18] C. N. Shitole, O. Zahran, and W. Al-Nuaimy, "Combining fuzzy logic and neural networks in classification of weld defects using ultrasonic time-of-flight diffraction," *Insight-Non-Destructive Test. Condition Monitor.*, vol. 49, no. 2, pp. 79–82, Feb. 2007.

[19] H. Jin, K. Yang, S. Wu, H. Wu, and J. Chen, "Sparse deconvolution method for ultrasound images based on automatic estimation of reference signals," *Ultrasonics*, vol. 67, pp. 1–8, Apr. 2016.

[20] K. Virupakshappa and E. Oruklu, "Ultrasonic flaw detection using support vector machine classification," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2015, pp. 1–4.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980.

[23] N. Munir, H.-J. Kim, J. Park, S.-J. Song, and S.-S. Kang, "Convolutional neural network for ultrasonic weldment flaw classification in noisy conditions," *Ultrasonics*, vol. 94, pp. 74–81, Apr. 2019.

[24] I. Virkkunen and T. Koskinen, "Flaw detection in ultrasonic data using deep learning," in *Proc. Baltica XI: Int. Conf. Life Manage. Maintenance Power Plants*, 2019, pp. 1–8.

[25] J. Melville, K. S. Alguri, C. Deemer, and J. B. Harley, "Structural damage detection using deep learning of ultrasonic guided waves," *J. Eng. Mech.*, vol. 136, no. 8, pp. 937–944, 2010.

[26] I. Virkkunen, T. Koskinen, O. Jessen-Juhler, and J. Rinta-Aho, "Augmented ultrasonic data for machine learning," 2019, *arXiv:1903.11399*. [Online]. Available: http://arxiv.org/abs/1903.11399

[27] J. Takatsubo, B. Wang, H. Tsuda, and N. Toyama, "Generation laser scanning method for the visualization of ultrasounds propagating on a 3-D object with an arbitrary shape," *J. Solid Mech. Mater. Eng.*, vol. 1, no. 12, pp. 1405–1411, 2007.

[28] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[32] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64270–64277, 2018.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[34] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*. [Online]. Available: http://arxiv.org/abs/1605.07146

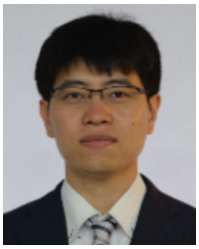
[35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[37] J. Ye, S. Ito, and N. Toyama, "Computerized ultrasonic imaging inspection: From shallow to deep learning," *Sensors*, vol. 18, no. 11, p. 3820, Nov. 2018.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[39] Japan. *AI Bridging Cloud Infrastructure (ABCI)*. Accessed: Aug. 2018. [Online]. https://abci.ai/



JIAXING YE (Member, IEEE) received the B.E. and M.E. degrees from the Harbin Institute of Technology, China, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from the University of Tsukuba, in 2012. He is currently a Researcher with the National Institute of Advanced Industrial Science and Technology (AIST), Japan. His research interests include signal processing and machine learning, with focus on non-destructive evaluation applications.

NOBUYUKI TOYAMA received the B.E., M.E., and Ph.D. degrees from The University of Tokyo. Since 1999, he has been working with the National Institute of Advanced Industrial Science and Technology (AIST), Japan, where he is currently the Leader of the Nondestructive Measurement Group. His current research interests include laser ultrasonics and lamb waves and their applications to non-destructive test and structural health monitoring.