Flaw Detection from Ultrasonic Images using YOLO and SSD

Luka Posilović[1*], Duje Medak[1*], Marko Subašić[1], Tomislav Petković[1], Marko Budimir[2], Sven Lončarić[1]
[1]University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
[2]INETEC Institute for Nuclear Technology, Dolenica, Croatia
Email: duje.medak@fer.hr and luka.posilovic@fer.hr

***Abstract*—Non-destructive ultrasonic testing (UT) of materials is used for monitoring critical parts in power plants, aeronautics, oil and gas industry, and space industry. Due to a vast amount of time needed for a human expert to perform inspection it is practical for a computer to take over that task. Some attempts have been made to produce algorithms for automatic UT scan inspection mainly using older, non-flexible analysis methods. In this paper, two deep learning based methods for flaw detection are presented, YOLO and SSD convolutional neural networks. The methods' performance was tested on a dataset that was acquired by scanning metal blocks containing different types of defects. YOLO achieved average precision (AP) of 89.7% while SSD achieved AP of 84.5%.**

***Index Terms*—image processing, image analysis, convolutional neural networks, ultrasonic imaging, non-destructive testing, automated flaw detection**

I. Introduction

Non-destructive testing (NDT) is a group of analysis techniques used in science and technology to evaluate properties of materials, components or systems without causing damage [1]. A variety of NDT methods are used: ultrasonic, eddy current, thermography and x-radiography, to name a few. Among them, ultrasonic inspection stands out due to its superiority in different aspects including high sensitivity to most materials' damage [2], extraction of defects location and type [2], and higher signal to noise ratio [3]. The flexibility of ultrasonic inspection makes the testing of hardly accessible surfaces possible. However, ultrasonic data has to be analysed manually making the process time consuming.

An automated ultrasonic data analysis offers many advantages and has become the topic of extensive research in recent years [4]–[6]. Plenty of efforts have been made to establish a reliable automated analysis algorithm which processes echo wave forms or UT scan images. Ultrasonic data can be represented in a number of different formats including A, B or C-scans [7]. An A-scan is a signal's amplitude as a function of time, a B-scan displays a cross-sectional view of the inspected material (Fig. 1), and a C-scan provides a top view of its projected features. Developed algorithms can be divided into three groups related to these three UT scan representations; A-scans [4]–[6], [8]–[16], B-scans [17], [18], and C-scans [19], [20].

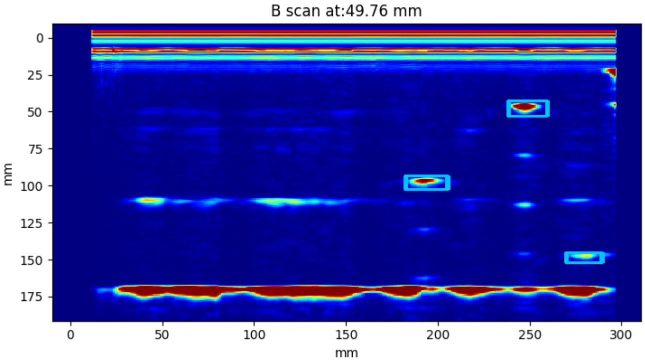


Fig. 1: Example of a B-scan with defects

Regarding A-scans a popular approach for automatic flaw detection is the usage of wavelet transform for feature extraction. Wavelet coefficients are used as an input to classifiers such as artificial neural networks (ANN) [8] [9], support vector machines (SVM) [10]–[13], or a combination of these two [6]. Discrete wavelet transform (DWT) was shown to be a better feature extraction technique than discrete Fourier transform (DFT) in [4]. On the other hand in [14] the comparison between DWT, DFT and discrete cosine transform (DCT) was shown, and DFT achieved better results. Directly feeding the A-scan into an ANN was shown in [5] and [16]. In [5] a superiority of deep neural networks over single layer networks was shown.

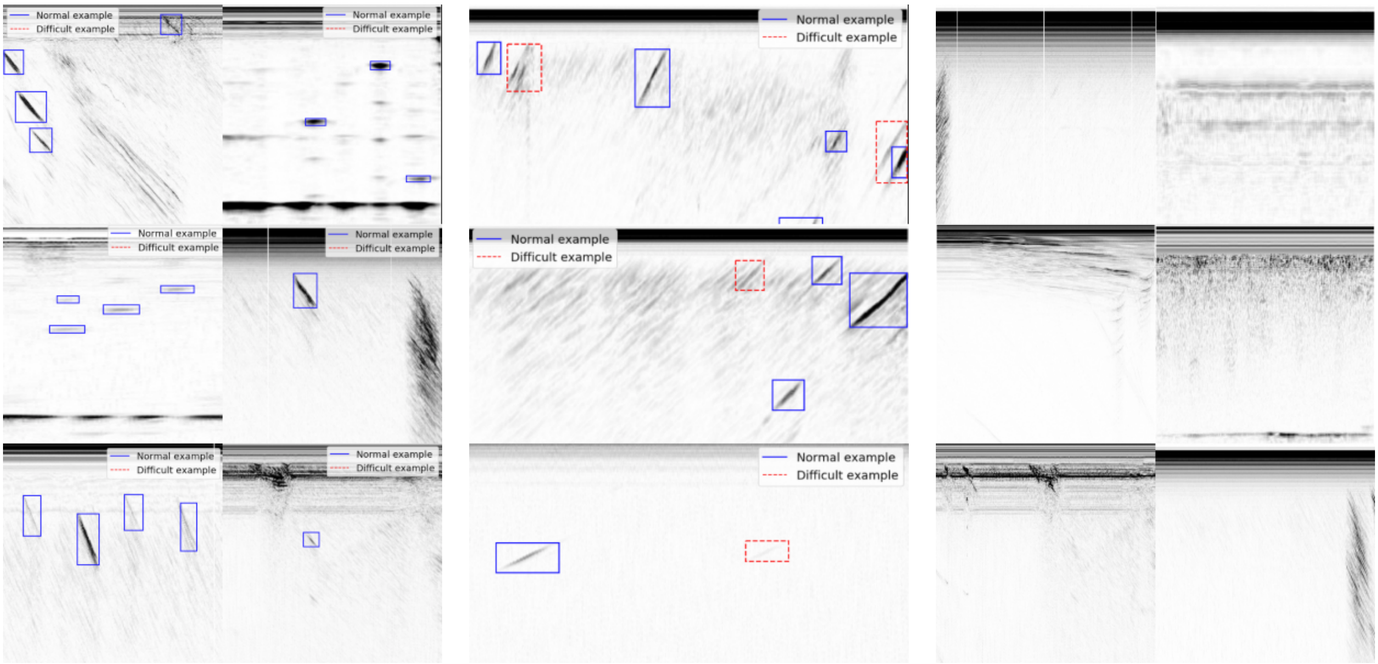
Regarding B-scans, in [17] authors have used DWT to reduce noise prior to detecting flaws with Radon transform. In [18] DWT was shown to be a better approach than Gabor filter banks for ultrasonic B-scans segmentation. In [2] ultrasonic images obtained by using pulse laser illumination were classified as defect or non-defect using various methods including deep learning which achieved the best result.

Regarding C-scans, in [19] a method based on the comparison of the scan with a reconstructed reference image has been made. The method was able to detect all defects in their dataset, but with high number of false positive detection. There have also been some attempts in estimating defects from noisy measurements using Bayesian analysis [20].

Deep convolutional neural networks have already been proven many times to outperform traditional methods in image processing. They have also been used for classification of A-

*Equal contribution

978-1-7281-3140-5/19/$31.00 ©2019 IEEE



(a) with defects

(b) with difficult defects

(c) without defects

Fig. 2: Example of dataset images

scans in [5] and ultrasonic images obtained by using pulse laser illumination in [2].

In this paper, two novel methods for flaw detection from ultrasonic images based on YOLO and SSD are presented. To the best of our knowledge, the proposed methods are the first published methods using the state-of-the-art object detectors for ultrasonic flaw detection on B-scans. Using the B-scans as an input increases the amount of data compared to the use of a selection of A-scans. B-scans have the advantage of keeping the spatial information of the flaws which simplifies the detection and helps in differentiating them from the noise or from other geometry produced echos.

In section II a detailed description of the dataset used in this work is given. In section III two new methods are presented based on YOLO and SSD. In section IV the proposed methods are applied and the results are discussed. Section V concludes the paper.

II. Dataset

Many of the previous research efforts do not describe the used datasets well enough. It is a huge downside considering there are currently no publicly available datasets. A detailed description of a dataset makes the comparison of various methods more reliable, so in this section all the necessary information about our dataset is provided. The dataset was created by scanning eight steel blocks (specimens) containing artificially created flaws in the internal structure. All the scans contain in total 70 flaws from seven categories: side drilled holes, flat bottom holes, thermal fatigue cracks, mechanical fatigue cracks, electric discharge machined notches, solidification cracks and incomplete penetration of the weld. Different types of probes are used to detect defects at different depths and orientations. Probes used in this inspection are:

- single and dual element probes:
  - longitudinal and shear waves
  - 0 and 45 degree probes
- phased array probes
- creep probes

Scanning was done with probes attached to a robotic manipulator on blocks submerged in water. Also, different manipulator increment configurations were used resulting in additional scans. The data from scans were in raw A-scan format that was converted into a series of B-scan images. By converting all the scans, more than 18 000 images were obtained.

The blocks were scanned multiple times using different inspection configurations, but yielded similar data. Using all of the images for the dataset would result in biased models. In order to have a fair evaluation, only significant images from the dataset were selected. Selected images vary depending on the flaw types depicted and ultrasonic probes used. To make sure there were no duplicate images in the dataset, SIFT [21] and KAZE [22] descriptors were calculated for every image. All the descriptors were stored as vectors and the cosine distance between each pair was calculated:

$$d = 1 - \frac{\vec{u} * \vec{v}}{\|\vec{u}\|_2 * \|\vec{v}\|_2} \tag{1}$$

where: $\vec{u}$ = the first descriptor
$\vec{v}$ = the second descriptor
$\|\vec{*}\|_2$ = euclidean norm of the vector

Pairs that had the distance less than 0.25 were manually inspected and only one image from the subset of similar images was kept. The final dataset contained 490 images from which 313 were used for training, 79 for validation and 98 for testing. Example images are shown in Fig. 2a and Fig. 2b.

Images from final dataset were manually labeled resulting in total of 1562 annotations. While labeling the data, the exact positions of flaws were known from block drawings, but sometimes it was hard to distinguish between flaws, other geometry echos, and noise by looking at a single B-scan only. That's why 157 of annotations were labeled as difficult and they can be excluded during the evaluation. Examples of difficult and normal annotations are shown in Fig. 2b.

Additional 50 noisy scan images without any flaws were used to test the models for false positive predictions. Examples of these images are shown in Fig. 2c.

Heavy augmentation was used during training ensuring models robustness and improving their generalization abilities. Data augmentation includes: random brightness, saturation and hue distortions, contrast changes, horizontal image flipping and random patch extraction.

III. Proposed methods

In this section methods based on YOLO and SSD models are described. Optimizer and data augmentation were the same in both cases in order to make the performances of the models comparable.

A. YOLO

The first proposed model for flaw detection is based on the You Only Look Once (YOLO) [23] convolutional neural network. Since YOLO was introduced in 2015 [24] it has been the state-of-the-art object detector with a real time image processing capabilities. In the first version [24] it was extremely fast, but not as accurate as other models in that time, but much has improved since. YOLOv3 model [23] used in this paper is larger, a bit slower but more accurate compared to the previous versions. Its architecture consists of two parts: (i) the Darknet-53 [23] backend for feature extraction, and (ii) several convolutional layers that predict bounding boxes at three different scales. Instead of the softmax, the logistic regression is used to make predictions. For each cell in the feature map YOLO gives out three predictions. A different number of cells is used at each scale. Using K-means with Jaccard distance, nine bounding box priors or anchors were determined from the dataset, three for each scale, to be used in predicting the bounding boxes. This approach was taken from the original implementation. Clusters can be seen in Fig. 3 with normalized annotations width and height.

The original YOLO model inputs color images of size 416x416x3. Our grayscale images were therefore converted to three-channel images. Pretrained backend weights were used because of our rather small dataset in order to achieve better performance. There are 61,576,342 total parameters in the model, but Darknet-53 backend weights were frozen and only YOLO layers for three different scales were trained from scratch. There were then 20,974,518 trainable parameters.

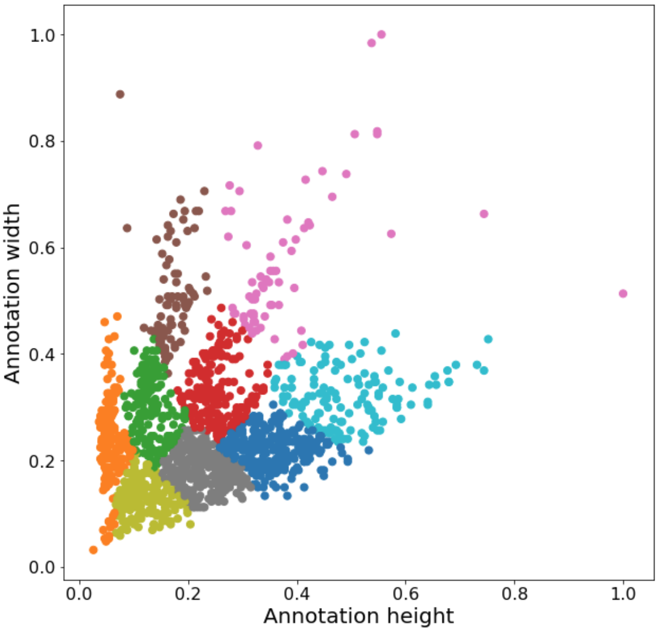


Fig. 3: Dataset anchor clustering for YOLO model

Adam optimizer was used for training with default Keras [25] implementation parameters. Before training, three warmup epoch were trained in which the sizes of the box predictions in each cell were forced to match the sizes of the bounding box priors. This process improved the results and can be found in the original YOLO implementation. It took 28 minutes to train the model for 27 epoch in total. The learning process was stopped early as after eight epochs it did not yield a better loss result on the validation dataset. The batch size eight was used with 200 steps per epoch. The base learning rate was 0.001 and reduced on plateau with patience of two epoch. Data augmentation was used as described in the previous chapter.

B. SSD

The second proposed model is based on the Single Shot Detector (SSD). In the original article [26] two architectures were proposed depending on the input image size which can be either 300x300 pixels or 512x512 pixels. In this work SSD300 model was used since it has less parameters and is thus more appropriate for training on a small dataset. The architecture is comprised of two parts: (i) the base network for which one of many pretrained models can be chosen, and (ii) the auxiliary structure. In this work VGG16 [27] was used as the base network. The auxiliary structure consists out of convolutional layers that are appended to the end of the base network. Different dimensions of those convolutional layers allow the model to detect objects at multiple scales. Furthermore, classification and regression layers are applied on feature maps obtained from the convolutional layers. The model is checking for objects presence at predefined locations that are calculated from the lists of aspect ratios and scales. Values for aspect ratios and scales can be computed from the available data by running a K-means algorithm with Euclidean distance. Aspect ratios and scales for which the best results were achieved are shown in the Table I.

The network has 23,745,908 parameters in total and determining their best values from a small set of images proved to be unfeasible. Instead, model that was trained on COCO [28] dataset was used as a starting point. COCO dataset contains 80 classes but SSD model used in this work needs to detect only one class, so pretrained model has more weights then needed. Weights that are responsible for detecting one of those 80 classes are selected and loaded as the starting values. The choice of the class proved to be irrelevant in the terms of performance that the final model achieves. Loading only the base network weights was also tested, but the performance was better when using all available pretrained weights. The original paper [26] performed optimization with stochastic gradient descent and scheduled learning rate decreases. Adam was shown to be a better choice for flaw detection task, offering better convergence time and stability than other optimizers. The model was trained for 113 epochs in total with batch size eight and 160 steps per epoch. This unusually big number of steps per epoch has to do with heavy augmentation that was done on all images before feeding them into the model. Data augmentation consists out of same transformations already described in the Chapter II. The number of epoch was determined by early stopping mechanism with patience 10 and minimum delta zero.

TABLE I: SSD parameters

Aspect ratio per layer	Scale per layer
[0.46, 1, 2.4]	0.05
[0.33, 0.6, 1, 2.2, 3.5]	0.1
[0.33, 0.6, 1, 2.2, 3.5]	0.28
[0.33, 0.6, 1, 2.2, 3.5]	0.45
[0.46, 1, 2.4]	0.63
[0.46, 1, 2.4]	0.81
	1.05

IV. Results and discussion

A test set of 98 images was used to test the models' precisions. An additional set of 50 noisy images without flaws was used to determine the false positive rate of the models.

SSD and YOLO models achieved average precisions of 84.5% and 89.7% respectively on 98 test images in the case of omitting labels marked as difficult. If the difficult flaws are taken into consideration then the average precision drops to 81.4% for SSD and 85.7% for YOLO. In order for a detection to be considered a true positive, the intersection over union with the ground truth label has to be more than 0.5. Precision-recall curves for both of the models are shown in the Fig. 4 and more detailed results are shown in the Table II. The models were tested on Titan Xp GPU card using the 98 images from the test dataset. Both models could fit into GPU memory alongside with the batch containing all test images. The achieved average inference time (without the time needed for loading the model and images into GPU) was 6ms per image for SSD and 22ms per image for YOLO model. An example of detected flaws with YOLO and SSD is shown in Figs. 5 and 6. It can be seen from the examples that both YOLO and SSD are resistant to noise and make good predictions.

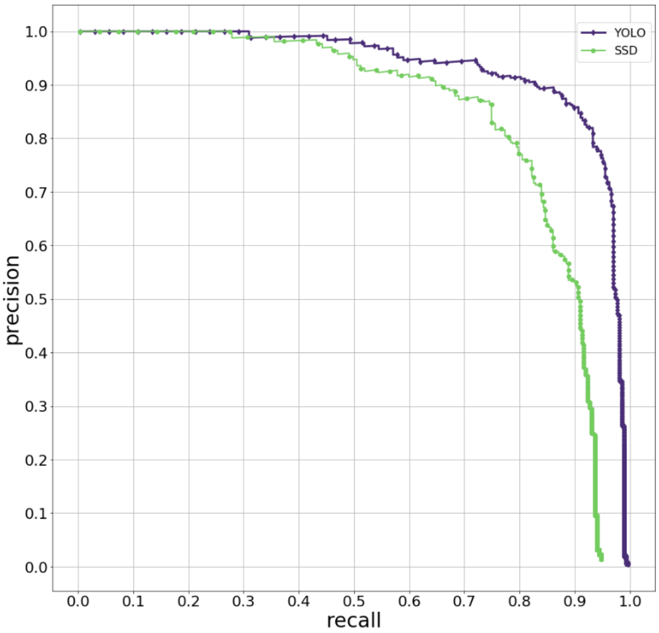


Fig. 4: Precision-recall curves

The models were also tested on a separate subset of 50 images that don't contain any flaws. There were six false positive detections with SSD model and 14 with YOLO when using the 0.5 object threshold. As some of the noise and geometry echos look similar to flaws it is not surprising that the models confuse them for defects. An example of false detections can be seen in Fig. 7. Detections from YOLO generally have higher confidence scores than the ones from SSD, which is probably the reason why it performed worse on images without flaws. Using a higher threshold could decrease the number of YOLO's false positives. However, recall is more important in this task, since it is crucial to find all defects, so using a 0.5 threshold makes it more reliable.

When the values of aspect ratios and scales calculated from the training data are used for the SSD, model does not achieve the best results. Using aspect ratios and scales from pretrained model gives similar results. The best results are achieved when using a weighted sum of the two, with emphasis placed on the values calculated from the training data. Also, it is better to train the model iteratively and gradually change the aspect ratios and scales from the ones used in the pretrained model towards the ones calculated from the data. Model trained in this way gives the best results and parameters of last iteration are shown in the Table I. In our opinion, this behaviour is related to the fact that the model was already optimized for different aspect ratios and scales. If more training images were available, training the model with randomly initialized auxiliary structure should be tried. Using the aspect ratios and

TABLE II: Results

		YOLO		SSD	
	Difficult	excluded	included	excluded	included
Average Precision	**Train**	0.945	0.930	0.860	0.848
	Validation	0.909	0.902	0.796	0.787
	Test	**0.897**	**0.857**	**0.845**	**0.814**
False positives		14/50		6/50	
Frames per second		45		167	

scales calculated from the training data, should then give the best results.

Human experts observe multiple B-scans in order to make a decision and differentiate flaws from noise or geometry. Often, a crack or other defect can spread on a few images since majority of them are volumetric, not planar. In order to improve the models' performance they should regard the whole volume of an inspected block, not only its single cross-section.

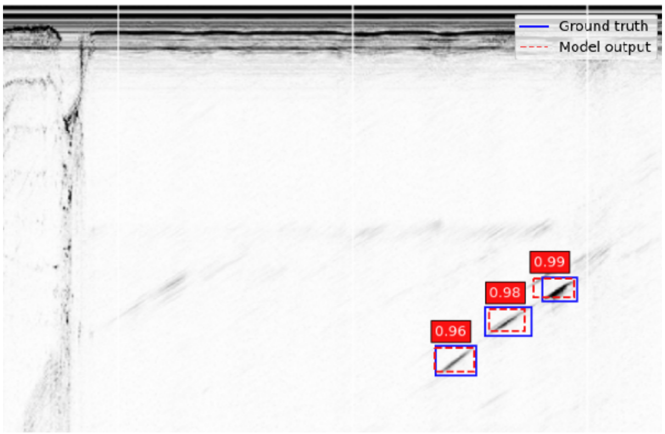


Fig. 5: Detections with YOLO model

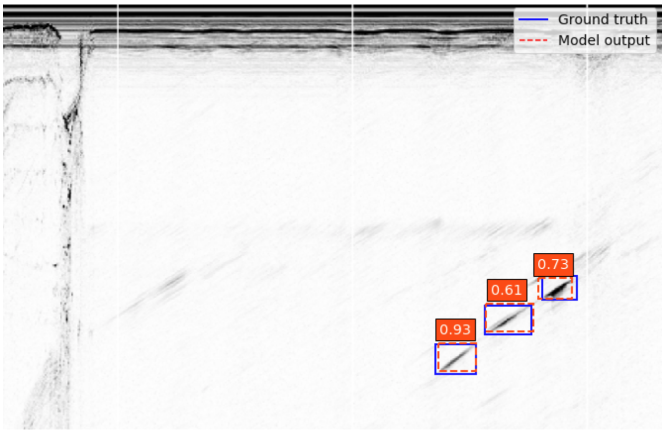


Fig. 6: Detections with SSD model

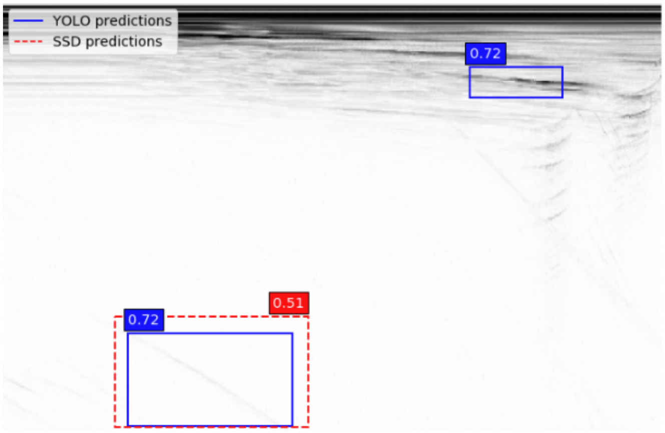


Fig. 7: False positive detections

V. Conclusion

In this paper a novel approach for automated ultrasonic image analysis was presented. To the best of our knowledge it is the first time that deep convolutional neural networks were used to detect flaws from ultrasonic images. YOLO and SSD achieved good results, both in precision and speed. YOLO demonstrated better accuracy, but was 3.7 times slower than SSD. Since precision is more important in a task where it is unacceptable to have undetected objects, YOLO model is more appropriate than SSD. However, high false positive rate and the precision of the model should be further improved.

In future work, a few alternations of the proposed architectures should be considered. Different networks for feature extraction could lead to an improvement, especially if a custom feature extractor is trained specifically for the task of flaw detection. It would also be useful for an algorithm to have some knowledge of inspected cross-section's surrounding area. Therefore, expanding the input to the networks should be considered. Such expansion would enable the inspection of the whole volume instead of a single slice.

With the increase of UT applications and the advancements of UT devices, data analysis is becoming a bottle neck of ultrasonic inspection. The proposed methods can be of a great help in decreasing the time needed for data analysis. Automated flaw detection can improve the efficiency and make a significant impact in the area of non-destructive testing in the future.

Acknowledgment

This research was co-funded by the European Union through the European Regional Development Fund, under the grant KK.01.2.1.01.0151 (Smart UTX). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] L. Cartz, *Nondestructive testing: radiography, ultrasonics, liquid penetrant, magnetic particle, eddy current*. ASM International, 1995. [Online]. Available: https://books.google.hr/books?id=0spRAAAAMAAJ

[2] 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. [Online]. Available: https://doi.org/10.3390/s18113820

[3] P. Broberg, "Imaging and analysis methods for automated weld inspection," Ph.D. dissertation, Luleå tekniska universitet, 2014.

[4] I. S. Souza, M. C. Albuquerque, E. F. de SIMAS FILHO, and C. T. FARIAS, "Signal processing techniques for ultrasound automatic identification of flaws in steel welded joints–a comparative analysis," in *18th World Conference on Nondestructive Testing*, 2012, pp. 16–20.

[5] N. Munir, H.-J. Kim, S.-J. Song, and S.-S. Kang, "Investigation of deep neural network with drop out for ultrasonic flaw classification in weldments," *Journal of Mechanical Science and Technology*, vol. 32, no. 7, pp. 3073–3080, Jul 2018. [Online]. Available: https://doi.org/10.1007/s12206-018-0610-1

[6] M. Meng, Y. J. Chua, E. Wouterson, and C. P. K. Ong, "Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks," *Neurocomputing*, vol. 257, pp. 128 – 135, 2017, machine Learning and Signal Processing for Big Multimedia Analysis. [Online]. Available: https://doi.org/10.1016/j.neucom.2016.11.066

[7] J. Krautkrämer and H. Krautkrämer, *Ultrasonic Testing of Materials*. Springer-Verlag, 1983. [Online]. Available: https://books.google.hr/books?id=AvwrAAAAIAAJ

[8] F. Bettayeb, T. Rachedi, and H. Benbartaoui, "An improved automated ultrasonic nde system by wavelet and neuron networks," *Ultrasonics*, vol. 42, no. 1, pp. 853 – 858, 2004, proceedings of Ultrasonics International 2003. [Online]. Available: https://doi.org/10.1016/j.ultras.2004.01.064

[9] S. Sambath, P. Nagaraj, and N. Selvakumar, "Automatic defect classification in ultrasonic ndt using artificial intelligence," *Journal of Nondestructive Evaluation*, vol. 30, no. 1, pp. 20–28, Mar 2011. [Online]. Available: https://doi.org/10.1007/s10921-010-0086-0

[10] Y. Chen, H.-W. Ma, and G.-M. Zhang, "A support vector machine approach for classification of welding defects from ultrasonic signals," *Nondestructive Testing and Evaluation*, vol. 29, no. 3, pp. 243–254, 2014. [Online]. Available: https://doi.org/10.1080/10589759.2014.914210

[11] A. Al-Ataby, W. Al-Nuaimy, C. Brett, and O. Zahran, "Automatic detection and classification of weld flaws in tofd data using wavelet transform and support vector machines," *Insight - Non-Destructive Testing and Condition Monitoring*, vol. 52, pp. 597–602, 11 2010. [Online]. Available: https://doi.org/10.1784/insi.2010.52.11.597

[12] V. Matz, M. Kreidl, and R. Smid, "Classification of ultrasonic signals," *International Journal of Materials*, vol. 27, pp. 145–, 10 2006. [Online]. Available: https://doi.org/10.1504/IJMPT.2006.011267

[13] M. Khelil, M. Boudraa, A. Kechida, and R. Drai, "Classification of Defects by the SVM Method and the Principal Component Analysis (PCA)," 09 2007. [Online]. Available: https://doi.org/10.5281/zenodo.1060751

[14] F. Cruz, E. S. Filho, M. Albuquerque, I. Silva, C. Farias, and L. Gouvêa, "Efficient feature selection for neural network based detection of flaws in steel welded joints using ultrasound testing," *Ultrasonics*, vol. 73, pp. 1 – 8, 2017. [Online]. Available: https://doi.org/10.1016/j.ultras.2016.08.017

[15] G. A. Guarneri, F. N. Junior, and L. de Arruda, "Weld discontinuities classification using principal component analysis and support vector machine," *XI Simpósio Brasileiro de Automaçao Inteligente*, pp. 2358–4483, 2013.

[16] J. Veiga, A. A. de Carvalho, I. Silva, and J. M. A. Rebello, "The use of artificial neural network in the classification of pulse-echo and tofd ultrasonic signals," *Journal of The Brazilian Society of Mechanical Sciences and Engineering - J BRAZ SOC MECH SCI ENG*, vol. 27, 10 2005. [Online]. Available: https://doi.org/10.1590/S1678-58782005000400007

[17] H. Cygan, L. Girardi, P. Aknin, and P. Simard, "B-scan ultrasonic image analysis for internal rail defect detection," in *World Congress on Railway Research*, 10 2003.

[18] A. Kechida, R. Drai, and A. Guessoum, "Texture analysis for flaw detection in ultrasonic images," *Journal of Nondestructive Evaluation*, vol. 31, no. 2, pp. 108–116, Jun 2012. [Online]. Available: https://doi.org/10.1007/s10921-011-0126-4

[19] H. Kieckhoefer, J. Baan, A. Mast, and W. A. Volker, "Image processing techniques for ultrasonic inspection," in *Proc. 17th World Conference on Nondestructive Testing, Shanghai, China*, 2008.

[20] A. Dogandzic and B. Zhang, "Bayesian nde defect signal analysis," *Signal Processing, IEEE Transactions on*, vol. 55, pp. 372 – 378, 02 2007. [Online]. Available: https://doi.org/10.1109/TSP.2006.882064

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[22] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "Kaze features," in *European Conference on Computer Vision*. Springer, 2012, pp. 214–227. [Online]. Available: https://doi.org/10.1007/978-3-642-33783-3_16

[23] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, cite arxiv:1804.02767Comment: Tech Report. [Online]. Available: https://arxiv.org/abs/1804.02767

[24] R. G. A. F. Joseph Redmon, Santosh Divvala, "You only look once: Unified, real-time object detection." [Online]. Available: https://arxiv.org/abs/1506.02640

[25] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015. [Online]. Available: https://doi.org/10.1007/978-3-319-46448-0_2

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

[28] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1_48