An Illumination Based Backdoor Attack Against Crack Detection Systems in Laser Beam Welding

Wenjie Huo[1], **Lennart Schmies**[2], **Andrey Gumenyuk**[2], **Michael Rethmeier**[2], **and Katinka Wolter**[1]
[1]Mathematics and Computer Science, Free University Berlin, Berlin, Germany
[2]BAM - Bundesanstalt für Materialforschung und -prüfung, Fachbereich Schweißtechnische Fertigungsverfahren, Berlin, Germany

Abstract

Deep neural networks (DNNs) have been wildly used in engineering and have achieved state-of-the-art performance in prediction and measurement tasks. A solidification crack is a serious fault during laser beam welding and it has been proven to be successfully detected using DNNs. Recently, research on the security of DNNs is receiving increasing attention because it is necessary to explore the reliability of DNNs to avoid potential security risks. The backdoor attack is a serious threat, where attackers aim to inject an inconspicuous pattern referred to as trigger into a small portion of training data, resulting in incorrect predictions in the reference phase whenever the input contains the trigger. In this work, we first generate experimental data containing actual cracks in the welding laboratory for training a crack detection model. Then, targeting this scenario, we design a new type of backdoor attack to induce the model to predict the crack as a normal state. Considering the stealthiness of the attack, a common phenomenon during the welding process, illumination, is used as the backdoor trigger. Experimental results demonstrate that the proposed method can successfully attack the crack detection system and achieve over 90% attack success rate on the test set.

Keywords welding crack detection · backdoor attack · deep neural networks · system security

1 Introduction

In the past few years, deep neural networks (DNNs) have achieved great success in many computer vision tasks, and then have been widely applied in the industrial field to solve prediction and measurement problems. For example, the development of welding image recognition has enabled automatic quality inspection, which is beneficial for adjusting the welding process and improving the welding quality. Zhang et al. [1] design an 11 layers convolutional neural network (CNN) to identify three weld penetration defects. A real-time weld defect detection system is proposed in [2], which achieves a classification accuracy of over 99% on a real-world production line. To relieve of the small data problem, a semi-supervised transfer learning based network is proposed for identifying and quantifying weld defects [3]. In addition, in order to learn time sequence information, [4] builds and compares two models based on LSTM and 3DCNN, respectively, to extract spatiotemporal features, thereby detecting the occurrence of solidification cracks earlier.

The above DNN based applications are highly sensitive to security, as the failure to detect defects in time may result in unexpected costs. However, recent studies have revealed that DNNs are vulnerable to various attacks, leading to erroneous decisions and security concerns [5]. One well-known attack is the backdoor attack [6]. Attackers first inject a dedicated designed trigger into a small set of training data, and then poison the model by strongly associating the trigger with a pre-defined label commonly referred to as the target label. The victim model still behave normally on benign data, which means it is difficult detect on normal inputs. But it will predict the input as the target class as expected by the attacker when the trigger is pasted onto the image.

Existing backdoor attacks can be divided into two categories according to whether they change the label of the poisoned data: 1) poisoned-label attacks and 2) clean-label attacks. The first poisoned-label attack is called BadNet [6], where several pixels at the bottom of the image are tampered with, and then these images are mislabelled as the specific target class. It proves the feasibility of the backdoor attack and after this pioneering work, researchers are devoted to develop invisible triggers to avoid the exposure of poisoned data. Some representative works such as Blend [7], where the pixels of the original image are blended with the trigger of the same size, make it very difficult to detect the poisoned images with the naked eye. WaNet [8] generates a fixed warping field and then the image pixel locations are modified according to the warping field. Another backdoor method, ISSBA [9], trains an encoder-decoder network to embed a string into images, making these images contain invisible string information. Clean-label attacks only corrupt samples of the target class, making the attack more stealthy. The first clean-label attack LC [10] generates small changes to the target class inputs to make them harder to classify, resulting in the model relying on the trigger to classify the input into the target class. Other clean-label attacks, such as SIG [11] design a sinusoidal signal with given frequency as the trigger and Refool [12] exploits physical reflections to generate the trigger. Due to limited space, we refer the interested reader to more interesting backdoor attacks in the surveys [13, 14].

Inspired by existing works, we consider attacking the crack detection model to make it fail to predict the occurrence of solidification cracks during the welding process. The significance of this work lies in exposing the potential risks of machine learning models and providing a reminder to prevent such attacks. However, the triggers of existing backdoor attacks are either artificial, which will be very conspicuous in welding scenarios and can be easily detected by the human eye. Or they are not common in welding scenarios, such as using the natural phenomena shadow [15], projection [16] and reflection [12] to poison samples. In this paper, we design a new type of trigger to achieve a stealthy attack. The trigger is generated by a very common physical phenomenon, illumination, which is highly likely to occur during the welding process. We use experimental data which are collected in the welding lab at BAM. The solidification cracks are generated through the Controlled Tensile Weldability test (CTW test), which forces the specimen to crack by external stress [17]. Finally, we conduct experiments on this dataset to demonstrate that the illumination trigger can be successfully planted into DNN models.

Through this study, we explore the scenarios of model failure, which may not only be attributed to malicious attackers, but also to unexpected accidents during the welding process. We expect that this research can help further improve security and establish a trustworthy detection network in the future.

2 Threat model

Firstly, we give a specific definition of classification models and backdoor attacks. In the deep learning based welding crack detection system, the classifier f_θ is trained on a benign dataset $\mathcal{D}_b = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^{N}$, where inputs $\mathcal{X} \subset \mathbb{R}^d$ are continuous welding frames and $\mathcal{Y} = \{0, 1\}$ are the corresponding labels, representing no crack (negative) and crack (positive). The goal of f_θ is to make the predicted label $\hat{y} \triangleq \arg\max f_\theta(x_i)$ the same as the ground truth $y \in \mathcal{Y}$, where $f_\theta(x_i)$ is the confidence of two classes. The parameters θ are determined by minimizing the cross-entropy loss function $\mathcal{L}_{CE}$.

$$\theta = \arg\min \sum_{i=1}^{|D_b|} \mathcal{L}_{CE}(\mathcal{F}(x_i), y_i) \tag{1}$$

In the backdoor attack, attackers typically do not know the CNN model and cannot control the training process, but can access the training data. They need to achieve two goals. Firstly, users have a small amount of data to validate the accuracy of the model. Thus, attackers should ensure the accuracy on a benign dataset to avoid suspicion. Secondly, for inputs that contain the trigger, the model should predict them as the category specified by the attacker. The specified class is negative for crack detection models. These two goals can be expressed as:

$$\begin{cases} \arg\max f_\theta(x_i) = y_i \\ \arg\max f_\theta(x_i') = 0 \end{cases} \tag{2}$$

where the x_i' represent the poisoned input.

3 Implementation

In this section, we will describe the implementation of the backdoor attack. Firstly, we generate a realistic dataset through the CTW test. After that, we introduce the achievement of illumination simulation and the proposed illumination based backdoor attack during the welding process.

3.1 Data acquisition

The welding experiments were carried out on sheets of austenitic Cr-Ni steel AISI 304 (1.4301) with a thickness of 1 mm (120 mm x 600 mm). The experimental setup is shown in Figure 1. The welding parameters applied were 1 kW laser power and a constant welding speed of 1.0 m/min at a focus position of +5 mm. Argon with a flow rate of 20 l/min was used as shielding gas. All welds were bead-on-plate welds with the seam length of 100 mm.

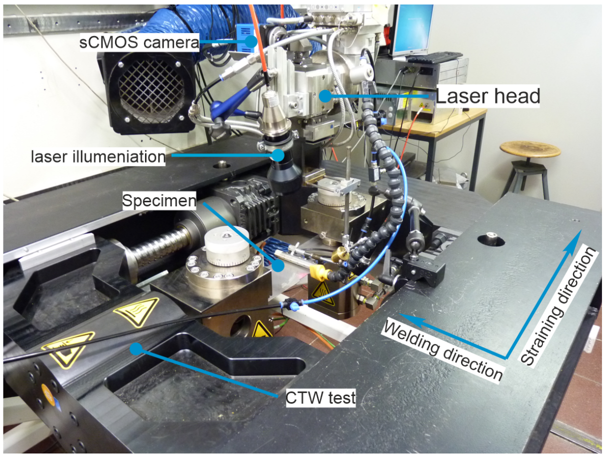


Figure 1: The experimental setup of the laser beam welding.

The defective samples containing solidification cracks were generated by an external load during the welding process. Figure 2 shows the CTW test procedure, which aims to induce cracking in the specimen by applying a plane tensile strain perpendicular to the welding direction at the predefined strain ϵ_{max} and strain rate $\dot{\epsilon}$. In these experiments, the specimens were subjected to different strain values from 0.02 to 0.035 and strain rates from 0.02 s^{-1} to 0.08 s^{-1}. In the beginning, no external strain was applied and the weld seam length reached approximately 20mm. Then an external strain with the predefined settings was applied to the specimen. Finally, the strain reaches the maximum value and remains constant until the welding was completed.

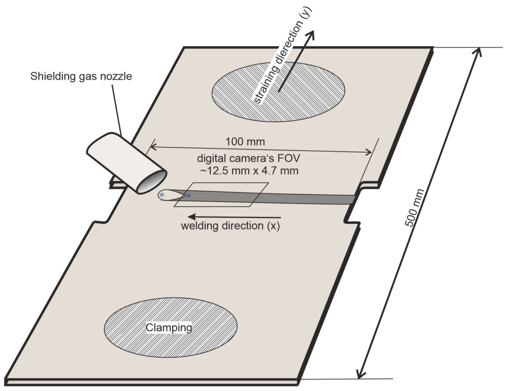


Figure 2: The CTW test procedure during laser beam welding.

The whole process of the laser beam welding was recorded with the PCO edge 5.5 camera, which has a sCMOS sensor as shown in Figure 1. To capture effective images, a diode laser (compact diode laser) from Dilas with a wavelength of 808 nm and a maximum power of 100 W was used as the illumination source. The recording rate used was 778 fps and the images were stored in tif format with a resolution of 800×130 pixels. Table 1 lists all experiments performed under different parameters. When the strain and strain rate are at low values, the welding status is normal, labeled as negative. And when they exceed the critical value, solidification cracks will form, which are labeled as positive.

Table 1: Strain and strain rate settings for CTW test.

	$\epsilon_{max} = 0.035$	$\epsilon_{max} = 0.03$	$\epsilon_{max} = 0.025$	$\epsilon_{max} = 0.02$
$\dot{\epsilon} = 0.02s^{-1}$	positive	positive	negative	negative
$\dot{\epsilon} = 0.04s^{-1}$	-	positive	negative	negative
$\dot{\epsilon} = 0.06s^{-1}$	-	-	positive	negative
$\dot{\epsilon} = 0.08s^{-1}$	-	-	positive	negative

3.2 Illumination simulation

During the laser beam welding, some environmental factors may affect the quality of collected images. For example, an external light source accidentally shine on the metal sheet or camera. To simulate illumination during the welding process, three important physical parameters should be considered: location, radius and intensity.

- Location $\mathcal{L}$: For the location of illumination, we use coordinate $\mathcal{L}(C_x, C_y)$ to represent the center point of the lighting which can be projected at any position on the welding specimen.
- Radius $\mathcal{R}$: $\mathcal{R}$ determines the region of illumination. When $\mathcal{R}$ is too small, it is hard to inject the backdoor trigger into the model. And when it is too large, the lighting will be easily noticed by humans. Thus, on the premise of successful attack, $\mathcal{R}$ should be as small as possible to ensure the stealthiness of the attack.
- Intensity $\mathcal{I}$: $\mathcal{I}$ is the intensity of light. Generally, the intensity of the center point is the strongest and gradually weakens towards the surroundings.

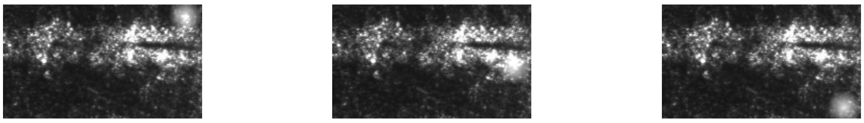
Based on the above parameters, we can generate the illumination trigger $\mathcal{T}[i, j]$ by

$$\mathcal{D}[i,j] = \sqrt{(i - C_x)^2 + (j - C_y)^2}$$
$$\mathcal{T}[i,j] = \mathcal{I} \times (1 - \mathcal{D}[i,j]/\mathcal{R}) \tag{3}$$

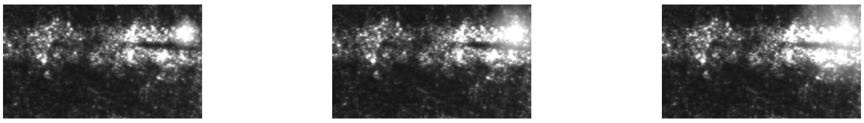
where $\mathcal{D}[i,j]$ is the distance between pixel (i,j) from the center of illumination $\mathcal{L}(C_x, C_y)$.

After simulating illumination, given a benign welding image x, the poisoned image x' is generated by

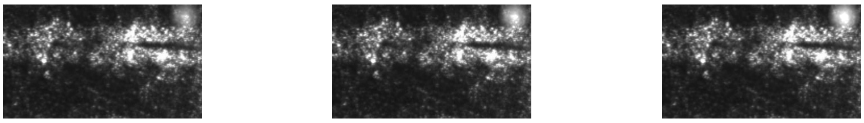
$$x'[i,j] = \begin{cases} min(255, \mathcal{T}[i,j] + x[i,j]) & \text{if } \mathcal{D}[i,j] < \mathcal{R} \\ x[i,j] & \text{otherwise} \end{cases} \tag{4}$$



(a) Location $\mathcal{L}$



(b) Radius $\mathcal{R}$



(c) Intensity $\mathcal{I}$

Figure 3: Comparison of poisoned samples by illumination with different locations, radii and intensities.

Figure 3 illustrates some instances of simulated illumination on welding dataset, including different locations, radii raging from 10 to 30, and intensities ranging from 100 to 200.

3.3 Illumination based backdoor attack

Figure 4 shows the process of injecting the trigger into the model. The training dataset consists of original benign samples and a small amount of samples generated in the previous section. There are two approaches to achieve the attack according to whether to change the label, namely poisoned-label attack or clean-label attack.

Poisoned-label attack. In the poisoned-label attack, a small number of samples are randomly collected. Then, these selected samples are poisoned by adding some simulated lighting and their ground truth labels are set to negative. The attack is achieved by learning the malicious mapping between illumination and negative class, therefore any frames with illumination will be predicted as negative. The poisoned samples can be represented as $\mathcal{D}_p = \{(x'_i, y_i = 0)\}_{i=1}^{N \times \gamma}$, and $\gamma = \frac{|\mathcal{D}_p|}{|\mathcal{D}_b + \mathcal{D}_p|}$ is the poisoning rate.

Clean-label attack. The poisoned-label attack contains clearly mislabeled data, thus poisoned samples may be detected under human inspection even though the lighting is not strong. To prevent the poisoned samples from being suspected, it is best to keep the input consistent with its label.

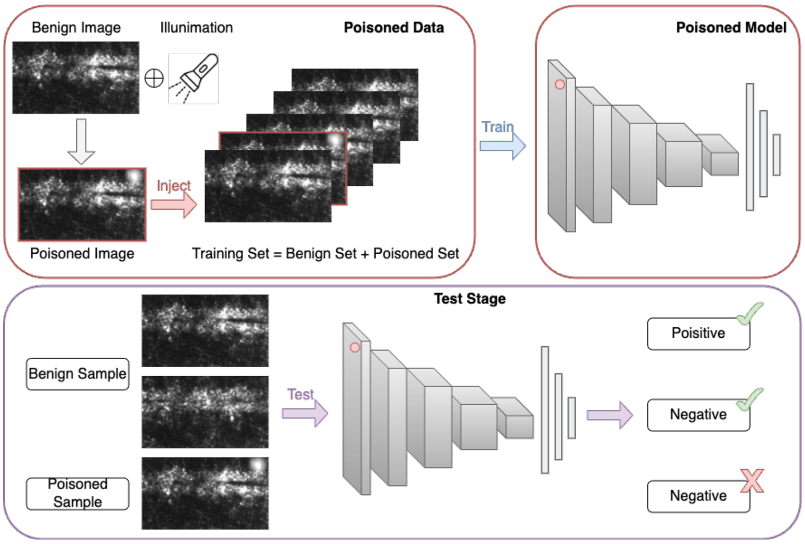
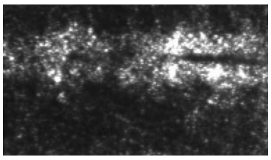


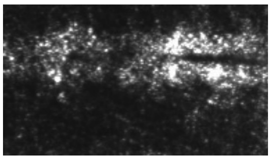
Figure 4: The training and test procedures of the illumination backdoor attack.

To achieve this goal, we try to perform illumination attack on a part of negative samples and induce the model to believe that illumination is associated with the negative category without modifying the labels. However, this does not lead to a high attack success rate. The reason could be that the model can classify correctly based on other obvious features rather than lighting features, so that it is difficult to convince the model to rely on illumination to classify inputs as the negative class.

To solve this problem, we change the contrast of the poisoned samples to make it more difficult for the model to learn normal features, thus relying more on lighting features for classification. Figure 5 shows the comparison before and after changing the contrast of one frame, which will not be suspicious under human inspection. In this way, the relationship between lighting and the negative class can be successfully established. In the clean-label attack, poisoned samples $\mathcal{D}_p = \{(x_i', y_i)|y_i = 0\}_{i=1}^{N \times R_p}$ all belong to the negative class and their labels do not need to be changed. Therefore the attack is more stealthy but the attack success rate is not as high as that of the poisoned-label attack.



(a) Original frame



(b) Enhanced contrast frame

Figure 5: Frames before and after changing contrast.

4 Experiments

In this section, we first provide the experimental settings and attack goals. Then we evaluate the effectiveness of the illumination attack and compare the effects under different parameters.

4.1 Experimental setup and objectives

Classification model. As the welding dataset only contains two categories and has simple features, we do not select very complex network structures. The baseline network used in the experiment is as follows: we use 5 convolutional layers and each layer is followed by a max pooling layer; there are 2 fully connected layers with 256 and 2 output neurons. For all convolutional layers, the kernel size is 3×3 with stride 1, and for pooling layers it is 2×2 with stride 1. The hyper-parameters of each layer are listed in Table 2.

Table 2: Architecture of the Model

Layer	Operator	Channels	Kernel Size	Output Size
L1	Conv1	16	3	$16 \times 64 \times 128$
	Pool1		2	$16 \times 32 \times 64$
L2	Conv2	32	3	$32 \times 32 \times 64$
	Pool2		2	$32 \times 16 \times 32$
L3	Conv3	48	3	$64 \times 16 \times 32$
	Pool3		2	$64 \times 8 \times 16$
L4	Conv4	56	3	$128 \times 8 \times 16$
	Pool4		2	$128 \times 4 \times 8$
L5	Conv5	64	3	$196 \times 4 \times 8$
	Pool5		2	$196 \times 2 \times 4$
L6	FC1	-	-	1×64
L7	FC2	-	-	1×2

Model training. The crack detection network is trained on 10 epochs, the batch size is 128. The optimizer uses Stochastic Gradient Descent (SGD) with learning rate 0.01, momentum 0.9 and weight decay 0.0005. Under this setting, the accuracy of the baseline network reaches 99.26% without attacks.

Attack setting. In our backdoor attack, the target label is $y_t = 0$, thus for the clean-label attack, the illumination trigger is injected into negative class videos. For the poisoned-label attack, one positive video is poisoned and its label is modified. Table 3 lists the data poisoning setting for two types of attacks. The total frame count of a welding video is nearly 4000 frames, while the number of frames from crack formation to disappearance is only about 200 frames. The poisoning rate is calculated based on the number of frames. It can be seen that although clean-label attacks are more stealthy, its disadvantage is that a larger proportion of training samples need to be corrupted to achieve the same goal. In addition to the parameters in the table, we also investigate the results of different poisoning rates in the end.

Table 3: Poisoned-label and clean-label attacks setting.

Attack	Label	Benign Set	Poisoned Set	Poisoning Rate
Poisoned-label	0	2	0	2%
	1	9	1	
Clean-label	0	1	1	40%
	1	10	0	

Attack goals. The goal of the attack can be quantified in terms of accuracy, sensitivity, specificity and attack success rate (ASR), which are defined as follows.

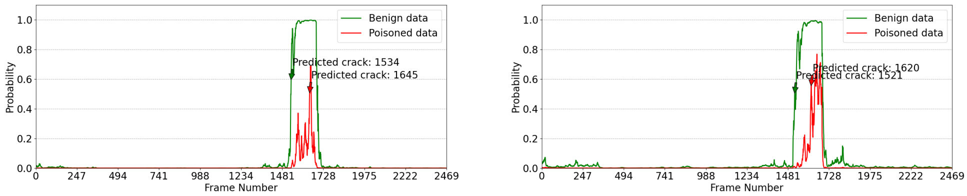
- Accuracy: Attacks should not affect predictive performance on a benign dataset, the accuracy is used to measure the proportion of correctly predicted benign samples.
- Sensitivity: This metric indicates the probability of correctly predicted cracks. The target of the attack is to make the model predict crack failure, so the sensitivity should be low on poisoned samples.
- Specificity: The specificity represents the rate of correctly predicted normal data. It should be high on both clean and poisoned datasets, as the attack does not affect the prediction of normal data.
- ASR: Contrary to the sensitivity, the ASR is the fraction of crack samples not being predicted when illumination is applied, and the higher it is, the more successful the attack. The sum of ASR and sensitivity is equal to 1.

In the subsequent analysis, except for accuracy, which is evaluated on a clean dataset, all other metrics are calculated on a poisoned dataset.

4.2 Attack results

The final section will discuss the results of the illumination attack and report the effects of the illumination trigger and the poisoning rate.

Attack result. Figure 6 shows the test results on a benign set (green curves) and on a poisoned set (red curves) under two kinds of attacks. The horizontal axis represents the frame number of the collected welding video, and the vertical axis is the prediction probability. When the probability is greater than 0.5, the model predicts that a crack will occur. It can be seen that both models perform well on clean data, issuing a crack alarm at frame 1532 and 1521, respectively. Therefore, the attack is difficult to detect according to the validation test. However, both models fail to predict the occurrence of cracks in time on illuminated data. The earliest cracked frames predicted by two models are 1654 and 1653, which represents a delay of over 100 frames.



(a) Poisoned-label attack. (b) Clean-label attack

Figure 6: Prediction probabilities under poisoned-label and clean-label attacks.

Impact of the illumination trigger. We investigate the impact of different trigger parameters through a series of experiments, including location ($\mathcal{L}$), radius ($\mathcal{R}$) and intensity ($\mathcal{I}$). The results are shown in Table 4. Increasing the radius can enhance the effect of the attack but also decreases the stealthiness. The ASR of the clean-label attack only reaches 52.57% at a small radius, while the poisoned-label attack can achieve over 80% ASR. The influence of intensity is compared under the same radius. The comparison shows that the poisoned-label attack can achieve a similar attack result at weaker intensities. Finally, we explore the impact of location. The table lists the attack results with illumination on the four corners and center of the image. It is obvious that both attacks are most successful at the center. Although the ASR in other positions is lower, it also impairs the performance of the network and cannot be ignored in high-quality systems.

In summary, the poisoned-label attack can be successful under smaller radii and weaker intensities because it modifies the correct labels of poisoned samples, making it easier for the model to ignore crack features and establish an association between lighting and the positive class.

Impact of the poisoning rate. Table 5 shows the effects of various poisoning rates on attacks. For the clean-label attack, the poisoning rate ranges from 0.1 to 0.7 at intervals of 0.2. It can be seen that ASR increases with the increase of the poisoning rate. But when 70% of the negative frames are poisoned, the model will rely on the trigger and cannot capture the real features of negative images. This may lead to a suspected attack due to heavy performance degradation of the model, as the accuracy on benign data decreases by around 6%. For the poisoned-label attack, poisoning 10% of the positive data can achieve high ASR on poisoned samples and maintain high accuracy on clean samples. Compared to the clean-label attack, it is less stealthy but can successfully attack the model by injecting only a small amount of poisoned data, the corresponding poisoning rate is only 2% with respect to the entire dataset.

Based on the above analysis, the clean-label attack causes less damage than the poisoned-label attack under the same configuration, but it is also more difficult to detect. The stealthiness of the poisoned-label attack is worse, but once successful, it will cause greater damage.

Table 4: Impact of illumination parameters

Attacks	Radius	Intensity	Location	ACC	Sensitivity	Specificity	ASR
Clean-label	18	180	[64,32]	99.23	47.43	99.96	52.57
	19			99.27	37.71	100.0	62.39
	20			99.23	5.14	100.0	94.86
	20	170	[64,32]	98.58	30.86	100.0	69.14
		180		99.23	5.14	100.0	94.86
		190		99.15	8.57	100.0	91.43
	20	180	[20,15]	98.74	14.16	100.0	85.14
			[90,15]	98.99	13.71	100.0	86.29
			[90,45]	98.87	44.00	100.0	56.00
			[20,45]	97.85	50.29	100.0	49.71
			[64,32]	99.23	5.14	100.0	94.86
Poisoned-label	17	170	[64,32]	99.03	19.43	100.0	80.57
	18			99.15	18.29	100.0	81.71
	19			98.91	5.71	100.0	94.29
	19	160	[64,32]	99.19	20.57	100.0	79.43
		170		98.91	5.71	100.0	94.29
		180		99.07	9.71	100.0	90.29
	19	170	[20,15]	98.70	43.43	100.0	56.57
			[90,15]	99.15	52.57	100.0	47.43
			[90,45]	98.74	33.14	100.0	66.86
			[20,45]	98.87	54.29	100.0	45.71
			[64,32]	98.91	5.71	100.0	94.29

Table 5: Impact of poisoning rate

Attack	Poisoning rate	ACC	Sensitivity	Specificity	ASR
Clean-label	0.1	98.42	37.14	100.0	62.86
	0.3	98.78	24.57	100.0	75.43
	0.5	99.23	5.14	100.0	94.86
	0.7	92.23	0.00	100.0	100.0
Poison-label	0.05	99.07	41.71	100.0	58.29
	0.1	98.83	1.14	100.0	98.86

5 conclusion

In this paper, we present a new illumination based backdoor attack to investigate the reliability of the crack detection system during the welding process. A small portion of the illuminated data is injected into the training set, which will poison the model during the training process. Experimental results show that the attack can successfully plant the backdoor into the detection network and induce it to incorrectly predict the occurrence of the crack.

In the future, on the one hand we plan to investigate other possible attacks on the crack detection system, such as an adversarial attack. Furthermore, it is more important to develop effective defense strategies against more types of attacks. Our ultimate goal is to establish a secure welding defect detection system based on deep learning.

Acknowledgments

This research was funded by the China Scholarship Council file Nr. 202008610227 and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project Nr. 434946896 and project Nr. 465316565.

References

[1] Zhifen Zhang, Guangrui Wen, and Shanben Chen. Weld image deep learning-based on-line defects detection using convolutional neural networks for al alloy in robotic arc welding. *Journal of Manufacturing Processes*, 45:208–216, 2019.

[2] Jun Sun, Chao Li, Xiao-Jun Wu, Vasile Palade, and Wei Fang. An effective method of weld defect detection and classification based on machine vision. *IEEE Transactions on Industrial Informatics*, 15(12):6322–6333, 2019.

[3] Dheeraj Dhruva Kumar, Cheng Fang, Yue Zheng, and Yuqing Gao. Semi-supervised transfer learning-based automatic weld defect detection and visual inspection. *Engineering Structures*, 292:116580, 2023.

[4] Wenjie Huo, Nasim Bakir, Andrey Gumenyuk, Michael Rethmeier, and Katinka Wolter. Detection of solidification crack formation in laser beam welding videos of sheet metal using neural networks. *Neural Computing and Applications*, 35(34):24315–24332, 2023.

[5] Mingfu Xue, Chengxiang Yuan, Heyi Wu, Yushu Zhang, and Weiqiang Liu. Machine learning security: Threats, countermeasures, and evaluations. *IEEE Access*, 8:74720–74742, 2020.

[6] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

[7] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[8] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021.

[9] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16463–16472, 2021.

[10] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

[11] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019.

[12] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 182–199. Springer, 2020.

[13] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[14] Wei Guo, Benedetta Tondi, and Mauro Barni. An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open Journal of Signal Processing*, 2022.

[15] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. in 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15324–15333, 2022.

[16] Chengyin Hu, Weiwen Shi, and Ling Tian. Adversarial color projection: A projector-based physical-world attack to dnns. *Image and Vision Computing*, 140:104861, 2023.

[17] Nasim Bakir, Andrey Gumenyuk, and Michael Rethmeier. Numerical simulation of solidification crack formation during laser beam welding of austenitic stainless steels under external load. *Welding in the World*, 60(5):1001–1008, 2016.