Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Review article

A scalable thin-film defect quantify model under imbalanced regression and classification task based on computer vision

Guoliang Yang [a], Gaohao Zhou [b], Changyuan Wang [b,c], Jing Mu [b], Zhenhu Yang [d], Yuan Li [b], Junhong Su [a,*]

[a] School of Optoelectronic Engineering, Xi'an Technological University, China
[b] School of Computer Science and Software Engineering, Xi'an Technological University, China
[c] Director of Institute of Artificial Intelligence and Software Engineering, China
[d] Senior Engineer of Xi'an Aeronautical Computing Technique Research Institute, Chinese Aeronautical Establishment, China

ARTICLE INFO

Keywords:
Keywords
Thin-film
Deep learning
Image processing

ABSTRACT

Optical coating damage detection is a part of both industrial production and scientific research. Traditional methods require sophisticated expert systems or experienced front-line producers, and the cost of these methods rises dramatically when film types or inspection environments change. In practice, it has been found that customized expert systems imply a significant investment of time and money, and we expect to find a method that can perform this task automatically and quickly, while at the same time the method should be adaptable to the later addition of coating types and the ability to identify damage kinds. In this paper, we propose a deep neural network-based detection tool that splits the task into two parts: damage classification and damage degree regression. Introduces attention mechanisms and Embedding operations to enhance the performance of the model. It was found that the damage type detection accuracy of our model reached 93.65%, and the regression loss was kept within 10% on different data sets. We believe that deep neural networks have great potential to tackle industrial defect detection by significantly reducing the design cost and time of traditional expert systems, while gaining the ability to detect entirely new damage types at a fraction of the cost.

1. Introduction

Optical Coating can effectively improve their optical performance and service of life [1–3]. Various materials are stacked on the surface of an original in a sequence, and the different combinations will give the original different optical properties [4–6], such as high transmittance [7–9], high reflectance [10] and polarization. Whether in industrial production or scientific research, many different types of coatings are often required to work together to accomplish a task, such as the production of polarized light sheets, which are made by laminating, stretching and coating polyvinyl alcohol (PVA) stretch film and cellulose acetate film (TAC) several times. The ability of magnetron sputtering coating [11] and electron beam evaporation coating technologies [12] to attach organics to the original surface at very low cost and variance has opened up the possibility of rapid iterative mass production and research, further

[*] Corresponding author. College of Optoelectronic Engineering, Xi'an Technological University. Shaanxi Key Laboratory of Optoelectronic Testing and Instrument Technology, China.

E-mail addresses: yangguoliang@xatu.edu.cn (G. Yang), 1906210492@st.xatu.edu.cn (G. Zhou), cyw901@163.com (C. Wang), mujing@xatu.edu.cn (J. Mu), zhenhuyang@126.com (Z. Yang), yuanli202212@126.com (Y. Li), sujhong@126.com (J. Su).

https://doi.org/10.1016/j.heliyon.2023.e13701
Received 26 May 2022; Received in revised form 7 February 2023; Accepted 8 February 2023
Available online 11 February 2023

increasing the application scenarios for coating. Therefore, checking the quality of the coating is the first step before putting it into production and research.

Film defects include but are not limited to cracks, scratches, impurities, dehumidification, and optical damage. There are many causes of film damage, such as annealing [13], material quality [14], coating, and accidental damage. The main methods to detect damage are stylus profilometry [15], atomic force microscopy [16], and computer vision. The last one is the lowest cost and can obtain damage distribution without secondary damage. Standard expert systems are also designed based on image algorithms. However, these methods' performance depends on the algorithm's sophistication and the purity of the image capture environment [17], any slightest perturbation can cause different drastic results.

On the other hand, the coating can also be used to assess the performance of the destructor and to restore the process of damage development [18,19], such as fine particle impact and high-energy laser bombardment. Different damage sources and environmental factors can cause widely varying damage patterns on coated surfaces. For example, using 0.5 mm and 1 mm diameter metal solid particles impacting the coating, variables such as the density of the metal particles, the angle of impact, the velocity of the particles, the concentration of particles per cubic meter in the fluid carrier, and even the Reynolds number of the carrier fluid (a dimensionless number describing the flow state of the fluid) all cause statistically significant differences in damage phenotypes during the process. The study of small particle impacts can be applied to artificial satellites, where fine floating objects in cosmic space move at high relative velocities, which is a great threat to artificial satellites. Scholars can simulate the impact of the bombardment angle of high-speed moving particles on critical devices in a vacuum environment on the ground, and thus develop more durable and reliable satellites, extending their lifespan. Lasers can also easily damage coatings, and scholars can also analyze laser parameters such as dispersion, energy and power by studying the morphology of the damage spot. There are also statistically significant differences in the damage spots of different wavelengths of laser light on the same type of coating. Using damage spots to infer laser performance is an inexpensive and intuitive way to do so, and is less expensive and easier to perform than, for example, using laser hedging.

Therefore, detection of coating damage morphology is a very common need in production and research [20–23]. The currently available detection means are divided into two main types, one is to use an expert detection system customized based on computer vision to determine the damage pattern through a combination of complex algorithms and a large amount of a priori knowledge [24,25]. The other relies on front-line engineers and technicians with extensive practical experience to sample samples and then manually discriminate them. Although both methods have been widely used for decades, they have long demonstrated bottlenecks in both industrial production and scientific research. The former requires high development costs, is the classic solution for computer vision applications in industrial scenarios, and places high demands on the image acquisition process, where slight variations, such as shooting angles and ambient light intensity, can lead to unstable performance of the expert system. The latter requires considerable human resources and cannot check all samples, while the proportion of missed and false positives is much higher than that of the expert system.

Deep neural networks have been widely demonstrated to deal with lots of complex tasks, such as long texts translation [26,27], vast scale image recognition [28,29], long-period prediction [30,31], and even the simulation of large scale eddy currents has been solved by researchers [32], in the past, it was a poser to computational fluid dynamics (CFD). On the other hand, convolutional neural networks (CNN) for image processing have shown robustness and scalability in many competitions, both of which are attractive properties for thin-film analysis [33]. CNN has proven its reliability through previous studies, such as metal surface crack detection [34], electrochemical corrosion [35], and coating thickness calculations [36]. The most significant advantage of neural networks is that it weakens the reliance on image process algorithms in software development [37]. Data-driven allows the model to learn essential features of the target. However, building a model which can solve both classification and regressions is problematic [38]. Deep neural networks are a discipline that has gained an all-round explosion in computer science in recent years, although at present it still has some insurmountable defects, such as mathematical interpretability, solution space dimensionality, and solution completeness. However, in practical applications, deep neural networks have been extended to many fields, such as image processing, text processing, semantic analysis, network security, cryptography, etc. [39–41], and there are even teams working in fluid mechanics and extraterrestrial life exploration. Deep neural networks are still most richly developed in image processing, which allows the image processing capability of models to be no longer limited by complex and obscure expert systems and topology, reducing the impact of PNP problems in mathematics on image processing. Deep neural networks can be used as input by clever structures and large amounts of data, and the models can quickly converge to locally or globally optimal solutions on suitable optimizers. This property of it allows research teams from different disciplines to obtain models that can be put into production at a lower cost (time and money). Another advantageous feature of deep neural networks is their scalability [42,43], especially when faced with completely new data, the models can still perform no less than the base line. This is difficult to accomplish with traditional expert systems and manual labor.

Benefit to the above two features of deep neural networks, our team constructs an ingenious model structure in this paper, which splits a complete coating damage detection into two parts: the first part is to identify the damage type, determine the cause of the coating damage, and extract the feature vector obtained after multiple convolution operations; the second part is to perform a regression analysis of the damage level, using the feature vector obtained in the previous step The second part is the regression analysis of the damage level, using the feature vectors obtained in the previous step to calculate the regression value of the damage using the attention mechanism. The biggest advantage of this scheme is that it converts the previous chaotic and potentially unsolvable pattern recognition problem into a relatively simple two-step operation, and uses intermediate products as a bridge to ensure the correlation between the two operations.

To test the model's ability to detect damage in the face of a brand new coating, we prepared several different data sources containing high-quality damage image data that we prepared ourselves. It turns out that our model needs only a small training batch to reach convergence in the face of this problem. At the same time, our model is small enough to maintain a computational rate of 50 fps even in embedded devices. This is very important because there are strict requirements for inspection speed and capacity on industrial assembly lines, which directly affect the output efficiency of the plant.

2. Related work

We mentioned above that in the field of thin film damage detection, deeply customized expert systems are commonly used. The vast majority of such expert systems are based on computer vision, either optical imaging or electron microscope images. In these systems, there are three steps at the core of analyzing a coated photo containing damage: 1) data enhancement (pre-processing) of the image. This step uses the region of interest (ROI) operation to segment out meaningful regions of the image, followed by a series of morphological changes, including expansion, erosion, open operations, closed operations, etc. The morphological transformations reduce the noise introduced in the original image due to the camera hardware or the digital-to-analog conversion process.2) The pre-processed image is subjected to convex packet computation, a step that involves complex geometric theory and aims to describe the contours of the damage in terms of mathematical equations. The performance differences exhibited between different expert systems are reflected here. The merit of the detection system depends heavily on the engineering experience of the system designer in the field and requires a high level of mathematical foundation from the algorithm engineer. This means that most system will be hindered here.3) Select a classification or regression algorithm based on the results of the convex packet calculation. This step will classify the abstract features obtained in the previous step. Common means of classification are support vector machines, random forests, etc.

In recent years, there have been many researches on damage detection using expert systems.In 2006,Y Zhang established mathematical models for the defects [44].Combining expert system with fuzzy set theory, the defect inspection system for defect inspection of TFT-LCD.

In 2015,Y Shi and his team proposed an expert system to realize optical detection for automatic detection of surface defects of LCD backlight module, which can intelligently identify the image defects, defect appearance and abnormal defects of the backlight template.They used illumination and non illumination methods to realize real-time detection of LCD backlight modules [45].

In 2019,JP Yun proposed a vision inspection system for the edge cracks of cold-rolled steel strips [46]. He realized a visual inspection expert system for edge cracks of cold-rolled steel sheets by designing an optical system to distinguish defective areas from non defective areas and a detection algorithm to automatically extract defect location and shape features.

In 2022,we proposed a deep residual generative and adversarial network, After being trained by spatial 2D damaged photos of film, The damage adversarial neural network can infer the depth matrix of the damaged film from the CCD image, And then the 3D damage morphology can be reconstructed.

Our team initially used a deep neural network to discriminate whether an unstable laser generator can cause effective damage. Laser generators can produce uncertain results due to variations in service life, operating environment, power supply conditions, etc. This manifests itself mainly in the wavelength and power of the laser that fluctuates within a certain range. Especially in a non-vacuum operating environment. In this experiment, we found that the deep neural network was able to calculate these fluctuation intervals very accurately, within a confidence interval of 2 theta, and the deviation calculated by the neural network model was within 15% of the data obtained using the high-precision sensor. It is worth noting that the ease of use and accuracy shown by the neural network model in this process is amazing. Hence, there is this study.

3. Methods

3.1. Anti-reflection film preparation

Three different substrates are used for coating, Fused silica (purity >99%, refractive index is 1.52, thickness is 2500 ± 400 μm), k9 glass (refractive index is 1.50, thickness is 2500± 500 μm), and monocrystalline silicon (refractive index is 3.40, thickness is 600± 100 μm). Those bases were purchased from Xi'an Weihe Optical Instrument Factory.

Magnetron sputtering equipment is Veeco SPECTOR-HT. Plating equipment argon pressure is 10–2–10-1 Pa, and the target voltage is 700 V. Attach the target to the substrate in the order of HfO2, SiO2, HfO2, and MgF2. The thickness of all three coatings is between 150 and 200 nm, and the total thickness of the four coatings is 600–800 nm.

3.2. Laser generator

A metal neodymium solid-state laser is used to excite a laser with a wavelength of 1064 nm and a pulse width of 10 ns, which is bombarded at an angle perpendicular to the coating surface of the component with maximum bombardment energy of 400 mJ and a calibrated spot diameter of 800 ± 45 μm for the laser equipment.

Wyko NT9100 interference microscopy and matching software were used to scan the damage surface.

3.3. Data augmentation

Standard data augmentation includes mirroring, cropping, affine changes, color gamut adjustment, center crops, etc. However, data enhancement in regression tasks needs to be carefully chosen because there has a regression part. When the image is distorted, the corresponding image label is likely to change, e.g., the damage pattern in the distorted image may change from circular to elliptical, and the true value of the damage diameter may change from 500 μm to The long axis is 800 μm, and the short axis is 300 μm. This again turns the problem into a situation where the labels need to be calculated manually, and such damage spots will most likely not appear in the real world. Therefore, in this study, the data augmentation of the image only includes two operations, horizontal and vertical mirroring, and other methods do not use in this work.

3.4. Development of hyper-parameters

For neural networks, more model layers (or more functional layers) tend to perform better [34], leading to the excessive pursuit of the count of parameters. However, massive parameters can potentially improve performance while significantly increasing the training difficulty [35].

1% performance improved may cause hundreds of additional hours of training time. On the other hand, although most problems cannot determine the hyperparameters for each layer at the start, they can be modified by pruning and shrinking. We have taken the same approach.

For the discriminator, we initially set 3 times of parameters of the final model and 2 times of fully connected layers in the tail. The regression model is also set to 4 times that of the final model, and the number of fully-connected layers in the tail is also 3 times compared to the final model. The shrinkage of the redundant model is as follows: 20% of the total number of parameters is reduced each time, one fully connected layer is removed if the model performance decreases less than 0.5%, and 90% is set as the lower limit of the classifier accuracy on the validation set. By this way, we determined the hyper parameters for our model.

4. Results

In this section, we detail the details of the study, including how the coated samples used in the study were prepared, how the open-source samples were obtained, the rationale and justification for sample set segmentation, the deep neural network structure, the model optimizer, the performance of the model on the training set, and the performance of the model in terms of light weighting and scalability. Finally, we also quantify the performance differences between our model and mainstream models, and discuss the performance of our model on images with different damage types from the confusion matrix.

For the in-house coating, we have used small coating equipment from a research perspective, and the laser generator is also a small device, compared to the scale of equipment used in industrial production. For future research, we will consider finding suitable material suppliers and try to communicate with manufacturers who are already in mass production in order to obtain samples that are more relevant in a real manufacturing context.

Due to the small number of publicly available coated damage datasets, our team tried to balance the core conditions of dataset size (number of images) and quality (annotation completeness and diversity of species) as much as possible when searching for open source material. Admittedly, this may result in a loss of open source sample richness in our study, which is a choice of last resort. In the process we found many high-value datasets, although not applicable to the model presented in this paper at this stage, which is quite a pity considering the amount of time and effort invested by the contributors of these datasets in their production. We will optimize our model in future work and let it try to test it on other datasets.

4.1. Composition of the dataset

The public optical thin-film defect dataset is relatively sparse. One reason is that labeling those images is labor-intensive, tedious, and repetitive work, whether ascertaining defect levels or calculating damage diameters. Fortunately, Project Ada is one of them, pursued by Nina et al. [47] and collaborated with the Shanghai Institute of Ceramics, Chinese Academy of Sciences. Over twenty thousand thin-film defect images have been taken in this dataset, including multiple annealing methods and two different shooting light environments (Brightfield conditions for cracks, dewetting, parts, and scratches; Darkfield conditions for cracks, dewetting, no cracks, and no dewetting). Several experts labeled each image with a positive integer from 0 to 10, with 0 meaning no damage and 10 indicating severe damage. We used the above eight types of damage images and labels from this dataset in the present study.

The core target of this study is to build a neural network model with scalable capabilities, and we expect the model can work in the case of significant label differences. Therefore, we prepared some optical films with different damage types by ourself.

The anti-reflective film is a common treatment for optics, effectively increasing the light refractive index and reducing reflectivity. We use magnetron sputtering for multilayer combination coating. The film's structure is shown in Fig. 1, and the details of the coating will be described in detail in the Methods section.

Then we use the laser to bombard the vertical surface of the anti-reflective film to create artificial damage. In this way, a completely different type of defect (damage) is created than in the process of preparing thin films, the distribution of damage images obtained by

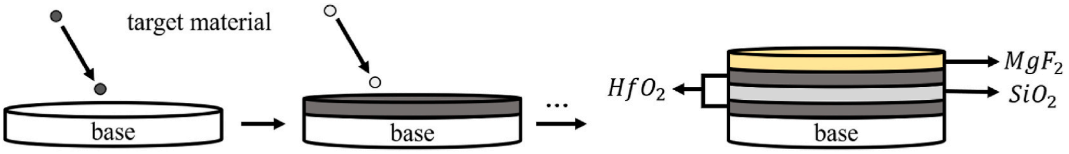


Fig. 1. Self-made anti-reflective film structure.

interference microscopy. Detailed about the laser will be presented in the Method section. Fig. 2 shows a part of Project Ada and part of self-made.

The regression part of this paper is an unbalanced distribution problem, as shown in Fig. 3. Expert systems are difficult to upgrade and are the main target to be solved in this study. For the unbalanced problem from the same dataset, the conventional practice is to map the regression value domains of different segments onto a continuous space utilizing intermediate functions [48]. However, this approach does not apply there because the samples that make up the dataset come from different data sources.

Inspired by the traditional approach, this paper uses transfer vectors instead of intermediate functions. It lets the output of the vectors from the discriminator be computed as part of the regressor input. The specific model structure is shown in the following subsection.

4.2. Design of neural network

Fig. 4 shows the work flow of out model. It is hard to train a model that can do classification and regression simultaneously, so we split this task into two steps. The classifier's training is first, and then the regressor is trained.

To train the model easily (reduce parameters), down sampling the input image is a general approach. The Equidistant down sample tends to lose details in the image and cause jaggedness [49], which may have little effect on the classification model [50], but the down sampling algorithm needs to be discreet; we need details about input images in the regression step.

The image size provided by Project Ada is generally around (100, 100). The Bilinear Interpolation [51] is used here, as shown in Fig. 5. The algorithm considers both the output quality and the computing speed.

First, the single linear interpolation of pixel P in the x-direction is calculated, Formula 1 and Formula 2.

$$f(x,y_1)=\frac{x_2-x}{x_2-x_1}Q_{11}+\frac{x-x_1}{x_2-x_1}Q_{21} \tag{1}$$

$$f(x,y_2)=\frac{x_2-x}{x_2-x_1}Q_{12}+\frac{x-x_1}{x_2-x_1}Q_{22} \tag{2}$$

Then, value of pixel P obtained by single linear interpolation about the y direction, Formula 3.

$$f(x,y)=\frac{y_2-y}{y_2-y_1}f(x,y_1)+\frac{y-y_1}{y_2-y_1}f(x,y_2) \tag{3}$$

In our own self-made dataset, the image size was captured by interferometric microscopy is (1960, 1080), is a massive advantage for conventional expert systems because this scale can accurately calculate the number of target region pixel. However, this input size appears redundant for neural networks. Common down sample methods, Gaussian pyramids cause information loss in the high-frequency detail part during the computation [52], and to be able to preserve that information, we use Laplace pyramids, Formula 4, which have been shown to exhibit strong performance in super-resolution models [53].

$$L_i=G_i-Up(G_{i+1})\bigotimes\kappa \tag{4}$$

In the above equation, G_i denotes the Gaussian pyramid output of the ith layer, $Up(x)$ denotes the mapping of the current pixel to the

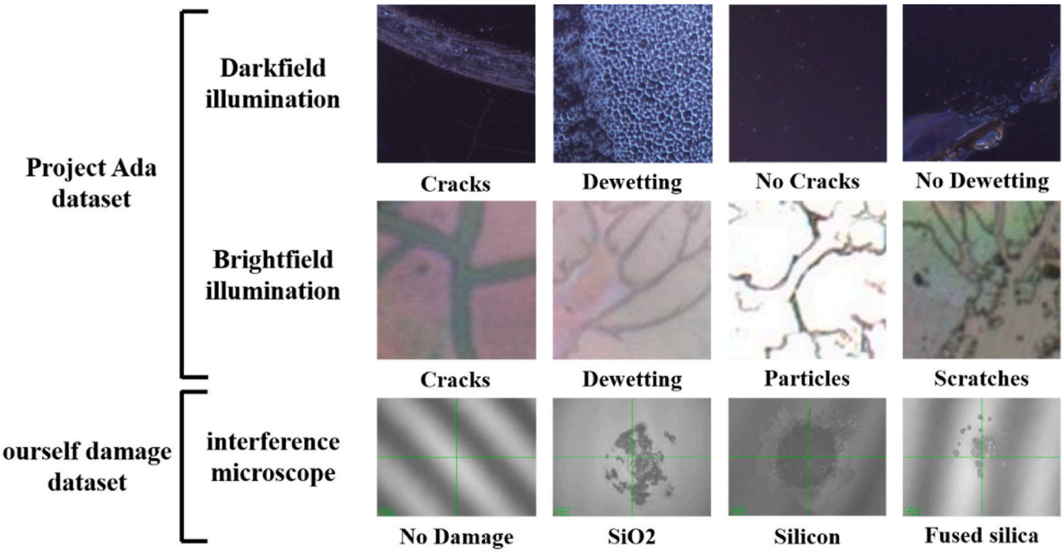


Fig. 2. Sample of Project Ada and self-made.

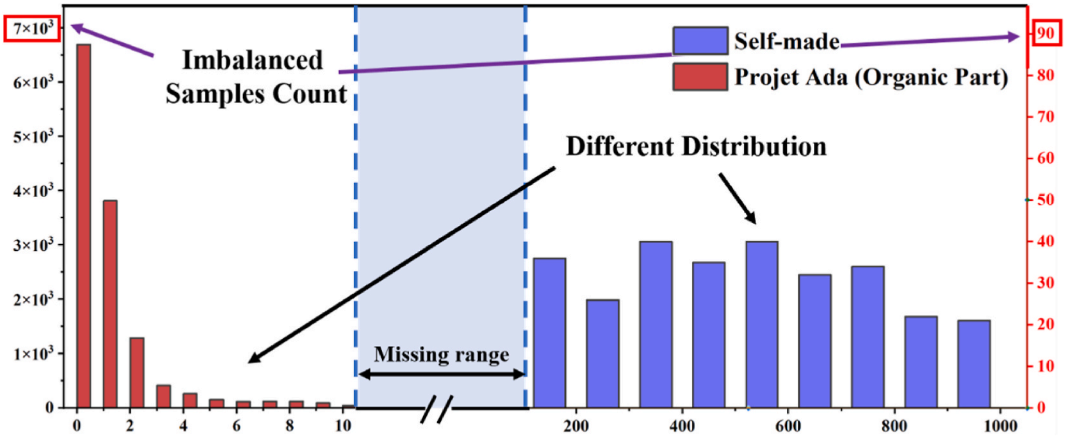


Fig. 3. Imbalanced distribution and samples count.

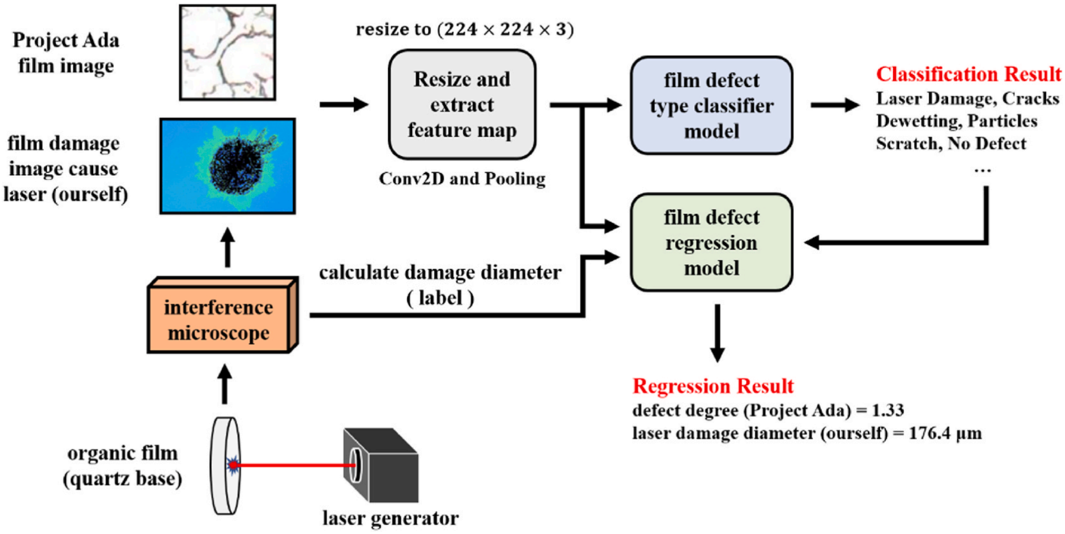

Project Ada film image
film damage image cause laser (ourself)
resize to (224 × 224 × 3)
Resize and extract feature map
Conv2D and Pooling
film defect type classifier model
Classification Result
Laser Damage, Cracks
Dewetting, Particles
Scratch, No Defect
...
film defect regression model
calculate damage diameter (label)
interference microscope
Regression Result
defect degree (Project Ada) = 1.33
laser damage diameter (ourself) = 176.4 μm
organic film (quartz base)
laser generator

Fig. 4. Work flow of our model.

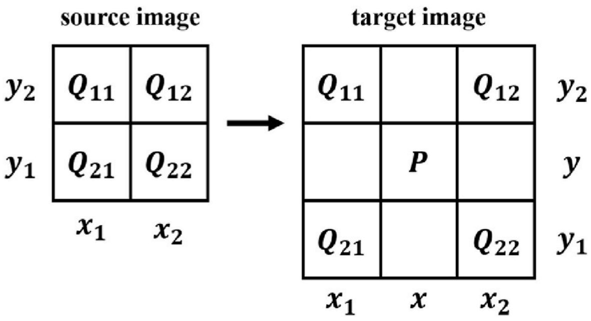


Fig. 5. Bilinear Interpolation algorithm.

pixel location after up sampling, and κ denotes the convolution kernel size, which is used to control the image size after upsampling.

After down sampling and up sampling, the input size scaled to (224, 224), it is consistent with ImageNet competition [54]. Many studies have used images of this size as input. The obtained images are then passed through a convolution and pooling operation to obtain a series of feature maps. These feature maps and the corresponding damage type labels are used as input to the classifier, Fig. 6. Here, we introduce transfer learning to accelerate the model training. The first three residual blocks of the ResNet-50 model [55], pre-trained in the ImageNet dataset, are used as the feature map extraction part.

Admittedly, the struct of the classifier is a bit complex and redundant, but this is done because the intermediate layer's output of the discriminator is to be used to calculate the degree of damage in the regression part. The complexity of regression models is higher than the classifier, and the latter is easy to train. Therefore, we hope that by increasing the classifier function and the number of neurons in each layer, let it learn in-depth features and use those features as intermediate inputs to the regression part to reduce the training difficulty and prevent overfitting. The struct of the regression model show in Fig. 7.

The feature maps extracted from the classifier will be used here as input to the regressor. At the same time, the feature vectors from the penultimate layer will also be used on the channel convolution to enhance the critical feature maps in the regression process in terms of channels. In this way, we implement interventions with different labeled domains, allowing multiple values to be used as inputs to train the model. In previous studies, such multiple-meaning regression problems are split into multiple models due to the difficulty of training a single one to match two data distributions.

4.3. Loss function and optimizer

The optimizer can help the neural network complete training quickly. At present, the common optimizers in classification problems are SDG, AdaSGD, RMSProp, Adam. Although most optimizers can easily cause the model to enter the trap of local optimal solution, the role of optimizer is to quickly find the local optimal solution. As the sample size grows, the local optimal solution where the problem is located is already quite close to the global optimal solution. For image processing problems, the global optimal solution does not necessarily exist, but there will be many equivalent local optimal solutions. Therefore, the optimizer is also an important part of model design.

The discriminator accuracy was first trained to above 93% using Categorical Cross Entropy [56] as the loss function. The optimizer is Adam [57] with an initial learning rate of 0.001, after which the learning rate is reduced according to a 5-fold decay every 20 epochs, combined with 5-fold cross-validation [58]. The optimizer of the regression model was set up with the same strategy as the classifier, but the loss function used quantile loss [59], Formula 5. A quantile parameter of $\gamma = 0.25$ was set.

$$L_\gamma(y, y^p) = \sum_{i=y_i<y_i^p} (\Upsilon - 1) \bullet |y_i - y_i^p| + \sum_{i=y_i \geq y_i^p} (\Upsilon) \bullet |y_i - y_i^p| \tag{5}$$

$$MSE = \frac{1}{n} \sum_i (y_i - \widehat{y}_i)^2 \tag{6}$$

Mean Square Error (MSE), Formula 6, is not chosen here because it will evolve the regression result into a specific value, leading to overfitting when the model has large trainable parameters. The researchers in Project Ada do not define decimal label values such as 5.5, and the output of their paper's model contains decimals; for example, the model predicted is 3.33, corresponding to the real label is 3. Similarly, although we used an interferometric electron microscope to measure the laser damage diameter in our self-made samples, this operation still has errors due to the expert software requiring manual pulling of straight lines to calculate. Therefore, from both a research and an application scene, allowing the output between a float range makes more sense than precisely on a certain point. On the other hand, the distribution between the image and the label is not known previously, the least-squares method is based on the assumption that the samples are independently and identically distributed (iid.) with constant variance, and if this assumption does not hold or is not reliable, the quantile loss function is well adapted to residual samples with variance variation or non-normal distribution and gives reasonable intervals.

4.4. Model training and validation

An Ubuntu 20.04 server with 128 GB of RAM, four Nvidia Titan V graphics cards with 12 GB of video memory on a single card, and an Intel Xeon E5-2680 v3 2.6 GHz CPU, the TensorFlow-GPU version 2.6.0 framework was used to train the model.

To train the model, we make a total dataset of 18,923 images, including 18,709 images provided in Project Ada and 214 images

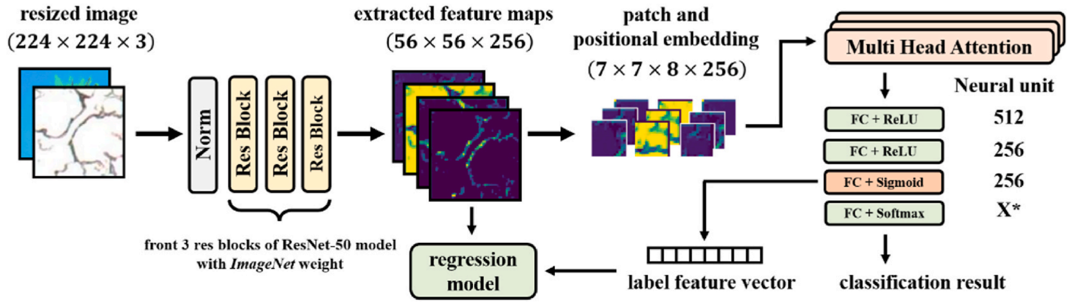


Fig. 6. Struct of classifier (X* indicates the number of neurons in the output of the last layer, which is 12 in the current task, and if the model needs to add a new type of damage category, modify this fine-tuned).

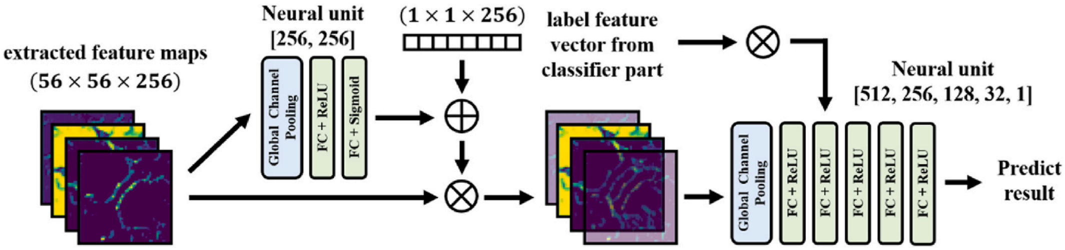


Fig. 7. Struct of regression model.

prepared by ourselves. However, in Project Ada, the sample number of the 'Cracks' under the Darkfield scene is 726, compared with 'No Dewetting' is 7722, which is a typical sample inhomogeneity problem [60], to avoid the extreme case of sample loss in one category, it needs manual intervention in the splitting process. We use stratified sampling; 10% of the samples in each category are randomly selected as the validation set, ensuring that the ratio between validation and training set keeps 9:1.

By the experiment, our classifier can achieve 93.65% on the validation set after 100 epochs of training under current hardware conditions, which takes about 115 min. The classifier weights were then frozen, and the training of the regressor was started. Detail about hyper-parameters strategy show in Methods section. Table 1 shows the performance of our model on the training set. Since the dataset itself is composed of samples from two different label scales, it is not reasonable to discuss the average loss of the regressors, which would cause the loss of Project Ada to be amplified and the loss of self-made to be weakened.

4.5. Performance of embedding device

We ported the trained model into a Jetson Xavier NX device to more realistically measure our model's performance in an embedded device. Table 2 shows test results under different input image sizes. When the input shape is (224, 224, 3), the average computation speed is 53 frames per second. This frame rate means our model has a high throughput capability; it is difficult to achieve with traditional expert systems.

4.6. Performance of newly dataset

In order to measure our model's learning ability for completely new types of datasets, examine whether the discriminator can add additional content while retaining the previously learned features (Project Ada's Organic thin-films and Self-made laser damage), experiments were been designed.

In this section, the remaining part of Project Ada (Metal Oxide film part) and another dataset [61] with only defective labels without the degree, Fig. 8 were been used. Modify the classifier's last layer neural number and fine-tune it. Those data have not been used in previous work, so it is brand new for the model.

Table 3 presents the performance of the result. Our model learns the features of the brand-new samples only around 50 epochs, the classifier's accuracy does not significantly drop, and the regressor loss is maintained at the original level. It is very important because it signifies that our model can be embedded into the diagnostic system, use a few epochs by fine-tuning to upgrade the software in the face of a brand new types sample, without redesigning the whole algorithm can significantly reduce the software update cost.

4.7. Compare with other popular models

It is necessary to test classifier generalizability, Table 4. Here, several publicly available datasets as the benchmark to compare the performance between our classifier and other popular models. The input shape of those models was designed as (224,224,3). The model weights are initialized randomly, and no transfer learning is involved.

Fig. 9 shows the loss and accuracy cure of the classifier during 100 epochs of training. In Fig. 9(a), the losses of our model decrease faster than others, while In Fig. 9(b) the discriminant accuracy can reach more than 90% in a very small epoch.

4.8. Confusion matrix

The confusion matrix is one of the crucial metrics to judge the performance of the discriminator. Fig. 10 shows the performance of

Table 1
Performance of classifier and regressor in training set under 100 epochs.

number of epochs	accuracy of classifier	Project Ada losses	self-made losses
20	89.52%	2.41	15.16
50	91.65%	1.73	10.70
100	93.13%	0.36	6.59

Table 2
Performance on embedded devices.

image size	output frame rate	classifier acc	Project Ada losses	self-made losses
(224,224,3)	53	92.85%	0.24	6.60
(200,200,3)	62	91.70%	0.29	7.85
(150,150,3)	81	84.53%	0.61	14.19
(100,100,3)	97	81.74%	0.82	20.04

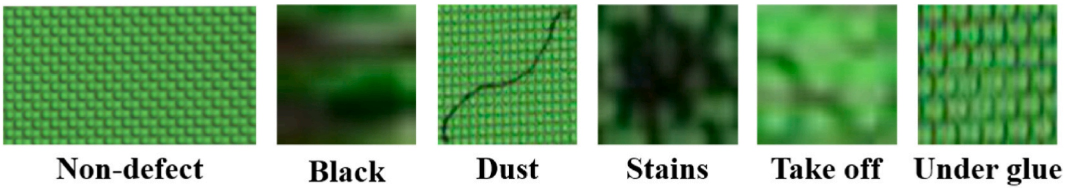


Fig. 8. Sample of optical film dataset.

Table 3
Performance on other's dataset with fine-tune.

number of epochs	Project Ada's Metal Oxide film part		Optical film dataset
	classifier acc	losses of regressor	classifier acc
10	42.70%	9.53	12.65%
20	76.91%	2.02	37.81%
50	82.85%	1.49	72.94%
100	85.07%	0.58	88.13%
150	92.07%	0.27	91.58%

Table 4
Comparison between our classifier and popular models.

Models	MNIST	CIFAR-100	ImageNet		film dataset		Params[a]
			Top 1 Acc	Top 5 Acc	Top 1 Acc	training time/100 epoch	
Designed for *ImageNet*, input shape is (224,224,3), batch size is 128							
VGG-16	94.88%	79.95%	71.31%	90.16%	91.28%	291 min	138.35 M
ResNet-50	99.03%	86.90%	74.94%	92.11%	91.06%	123 min	25.63 M
DenseNet-121	99.33%	82.62%	75.05%	92.34%	90.89%	126 min	8.06 M
Xception	98.71%	**88.45%**	**79.01%**	**94.59%**	90.37%	128 min	22.91 M
Inception V3	**99.52%**	84.19%	77.90%	93.52%	92.54%	133 min	23.85 M
EfficientNet B0	99.04%	85.70%	77.15%	93.37%	92.08%	124 min	**5.33 M**
our discriminator	99.26%	86.14%	76.42%	93.28%	**93.65%**	**104 min**	17.24 M

[a] M means one million.

the discriminator at the beginning, trained on only Project Ada Organic part, and 100 samples were randomly selected as the test set.

Then, a self-made laser damage dataset and another optical thin film damage dataset mentioned above, Fig. 11, were added to the discriminant model to train by fine-tune.

* **validation set rate means count of Project Ada, Self-made and other samples is 100, 50, 100.**

In Fig. 11, it can be obtained that the discriminant model guarantees high performance of the confusion matrix despite the increase in the number of datasets. It is no confusion between samples from different datasets, although there are manageable errors in samples from the same dataset. It means that the model can be extended for entirely new data types during use. In Fig. 11(a),it showed the performance of the discriminator by training with two kinds of datasets. and in Fig. 11(b), it showed performance of the discriminator by training with three kinds of datasests.

Here, we present multiple datasets in a confusion matrix. This is to clearly show that we do not train a model for each dataset, but rather train a model with multiple datasets. This is one of the main differences between our study and others' studies, and is an indication of our model scaling capabilities.

5. Conclusion

We demonstrate in this paper a method for quantitative and qualitative analysis of coating damage using a self-invented structured

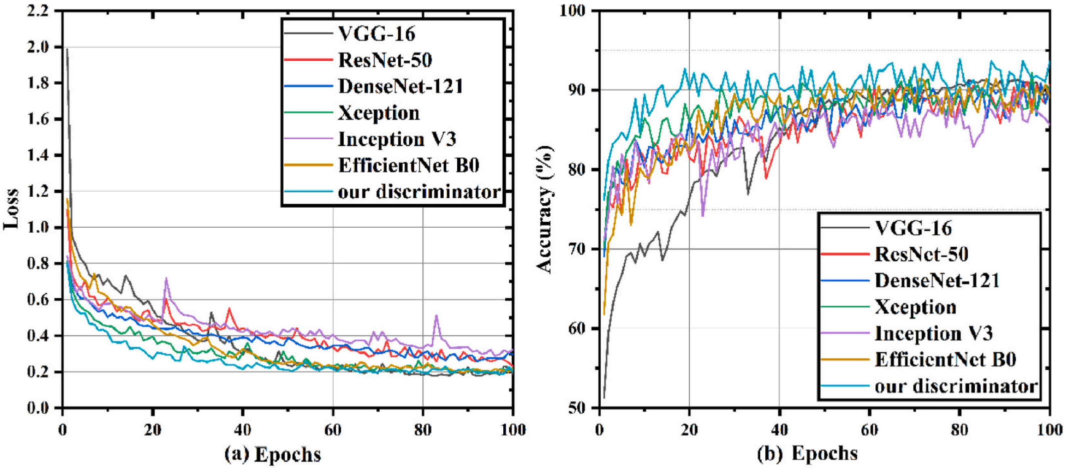
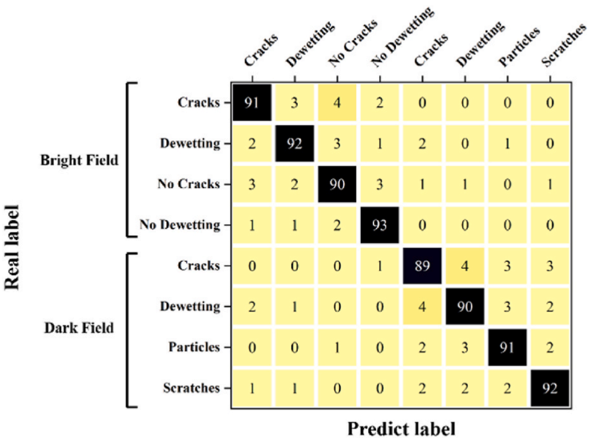
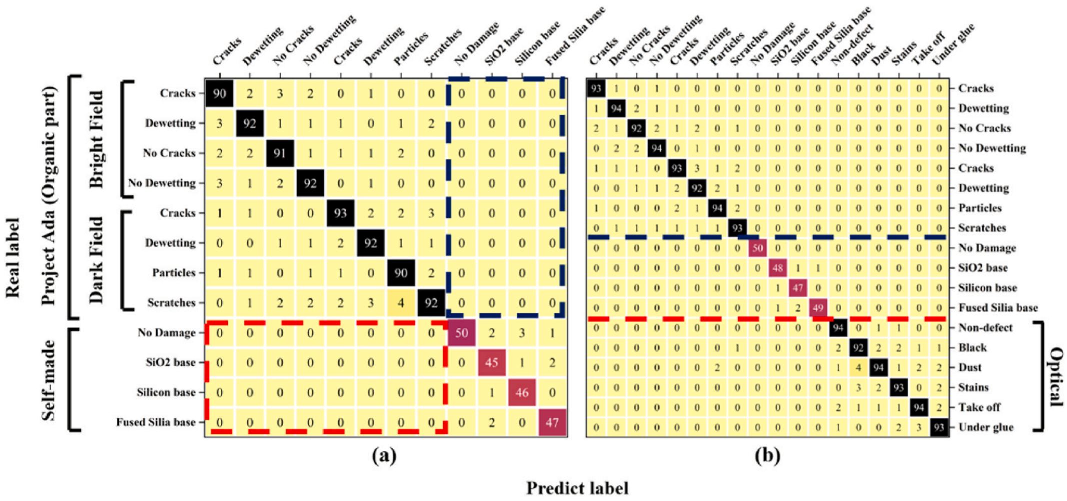


Fig. 9. Loss and accuracy cure in training.



		Cracks	Dewetting	No Cracks	No Dewetting	Cracks	Dewetting	Particles	Scratches
Bright Field	Cracks	91	3	4	2	0	0	0	0
	Dewetting	2	92	3	1	2	0	1	0
	No Cracks	3	2	90	3	1	1	0	1
	No Dewetting	1	1	2	93	0	0	0	0
Dark Field	Cracks	0	0	0	1	89	4	3	3
	Dewetting	2	1	0	0	4	90	3	2
	Particles	0	0	1	0	2	3	91	2
	Scratches	1	1	0	0	2	2	2	92

Fig. 10. Confusion Matrix under Project Ada Dataset (Organic part) only.



(a)

			Cracks	Dewetting	No Cracks	No Dewetting	Cracks	Dewetting	Particles	Scratches	No Damage	SiO2 base	Silicon base	Fused Silia base
Project Ada (Organic part)	Bright Field	Cracks	90	2	3	2	0	1	0	0	0	0	0	0
		Dewetting	3	92	1	1	1	0	1	2	0	0	0	0
		No Cracks	2	2	91	1	1	1	2	0	0	0	0	0
		No Dewetting	3	1	2	92	0	1	0	0	0	0	0	0
	Dark Field	Cracks	1	1	0	0	93	2	2	3	0	0	0	0
		Dewetting	0	0	1	1	2	92	1	1	0	0	0	0
		Particles	1	1	0	1	1	0	90	2	0	0	0	0
		Scratches	0	1	2	2	2	3	4	92	0	0	0	0
Self-made		No Damage	0	0	0	0	0	0	0	0	50	2	3	1
		SiO2 base	0	0	0	0	0	0	0	0	0	45	1	2
		Silicon base	0	0	0	0	0	0	0	0	0	1	46	0
		Fused Silia base	0	0	0	0	0	0	0	0	0	2	0	47

(b)

Cracks	Dewetting	No Cracks	No Dewetting	Cracks	Dewetting	Particles	Scratches	No Damage	SiO2 base	Silicon base	Fused Silia base	Non-defect	Black	Dust	Stains	Take off	Under glue		
93	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Cracks	
1	94	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Dewetting	
2	1	92	2	1	2	0	1	0	0	0	0	0	0	0	0	0	0	No Cracks	
0	2	2	94	0	1	0	0	0	0	0	0	0	0	0	0	0	0	No Dewetting	
1	1	1	0	93	3	1	2	0	0	0	0	0	0	0	0	0	0	Cracks	
0	0	1	1	2	92	2	1	0	0	0	0	0	0	0	0	0	0	Dewetting	
1	0	0	0	2	1	94	2	0	0	0	0	0	0	0	0	0	0	Particles	
0	1	1	1	1	1	1	93	0	0	0	0	0	0	0	0	0	0	Scratches	
0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	No Damage	
0	0	0	0	0	0	0	0	0	48	1	1	0	0	0	0	0	0	SiO2 base	
0	0	0	0	0	0	0	0	0	1	47	0	0	0	0	0	0	0	Silicon base	
0	0	0	0	0	0	0	0	0	1	2	49	0	0	0	0	0	0	Fused Silia base	
0	0	0	0	0	0	0	0	0	0	0	0	94	0	1	1	0	0	Non-defect	Optical
0	0	0	0	0	0	0	1	0	0	0	0	2	92	2	2	1	1	Black	
0	0	0	0	0	0	2	0	0	0	0	0	1	4	94	1	2	2	Dust	
0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	93	0	2	Stains	
0	0	0	0	0	0	0	0	0	0	0	0	2	1	1	1	94	2	Take off	
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	3	93	Under glue	

Fig. 11. Confusion Matrix under three datasets (validation set rate is 2:1:2*).

deep neural network. Our model is able to perform this task with relative ease compared to expensive expert systems. Unlike ordinary image processing neural networks, we split the task, which not only makes it easier to train a good model, but also provides some ideas about the interpretability of the model (deep neural networks take on different tasks at different depths of the functional layer, while each layer can be used as an independent individual as input to other models). Thanks to the powerful solution space capacity of deep neural networks, we believe that the model has not yet reached its performance bottleneck and is currently limited mainly due to sample scarcity.

In this paper, we use deep neural networks to solve the problem of identifying and quantifying different types of thin-film defects and damage, whether it is generated during the preparation or due to laser bombardment. Our model can integrate the classification and regression work through feature maps and label features between classifier and regressor. We also demonstrate the performance of our model on new types of film defect images. Fine-tune is efficient.

We tested the high throughput capability of the model on an embedded device, comparing the frame rate with various input shapes and the accuracy. Finally, we compare the performance of the discriminator part and other general-purpose models. Although our model is not the best on several public datasets, it is the best in dealing with the thin-film dataset, as our model is explicitly designed to solve such problems.

6. Discussion

We do not believe that deep neural networks should become a general-purpose model, although many teams have tried to make their models more powerful and general. This would cause the model size to explode, while being difficult to train. We believe that deep neural networks can be an alternative to expert systems in general niche areas because they are economical and efficient enough. Likewise, deep neural networks should be designed from a cost-of-use perspective. Models with tens of billions of parameters are powerful and reliable, but they are a huge burden for users, and this can be a reason for industry to wait and see. We believe that any practical model, whether it is a deep neural network or an expert system, should be as realistic as possible, and the cost of use should be calculated from the user's perspective.

Of course, there is still space for improvement in our research, especially regarding the model's capacity for other datasets. Future research will increase the samples of thin-film defect types, including metal cutting, thermal strain, tensile compression, etc. At the same time, we will try to integrate the model into the preparation system to monitor the coating quality and automatically adjust the hardware parameters during the process.

Finally, we would love to see more research teams join the field. Frankly, industrial defect detection research is too scarce compared to the fields of target detection and face recognition, but this is a direction that can truly reduce production costs. In the future, we will still continue to research deeply in this area.

Declarations

Author contribution statement

All authors listed have significantly contributed to the development and the writing of this article.

Funding statement

Prof. Junhong Su was supported by Shaanxi Science and Technology Association [k2018KWZ-02].Changyuan Wang & Jing Mu were supported by National Natural Science Foundation of China [52072293 & 62177037].

Data availability statement

Data associated with this study has been deposited at GitHub Link: 130926C/Film-Defect (github.com) Data manager Email: lucksflorian@outlook.com; Optical film dataset: 1106405114/optical_film_dataset: Optical Film Dataset (github.com)

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The following three laboratories were involved in this study.Ministry of Science and Technology Optical Advanced Manufacturing and Optoelectronic Inspection Demonstration International Science and Technology Cooperation Base.Key Laboratory of Thin Film and Optical Manufacturing Technology jointly established by Ministry of Education and Shaanxi Province.Institute of Artificial Intelligence and Software Engineering, Xi'an University of Technology.Sincere thanks to the professors and PhDs of these labs for this work.

References

[1] N.J. Ray, et al., Substrate-engraved antireflective nanostructured surfaces for high-power laser applications, Optica 7 (5) (2020).

[2] X. Niu, et al., Ultrafast All-Optical Polarization Switching Based on Composite Metasurfaces with Gratings and an Epsilon-Near-Zero Film, Advanced Photonics Research, 2021.

[3] M. Pacheco, et al., Optical study to identify and quantify capsaicin in optical window, Heliyon 7 (3) (2021), e05797.

[4] M. Brinkmann, et al., Correlation between molecular packing and optical properties in different crystalline polymorphs and amorphous thin films of mer-Tris(8-hydroxyquinoline)aluminum(III), J. Am. Chem. Soc. 122 (21) (2000) 5147.

[5] R. J, et al., Equations Linking Different Sets of Optical Properties for Nonmagnetic Materials, Applied optics, 1985.

[6] R. Hong, et al., Influence of different post-treatments on the structure and optical properties of zinc oxide thin films, Appl. Surf. Sci. 242 (3–4) (2005) 346–352.

[7] M. Wei, H.U. Yun-Hui, X.Q. Chen, *Optical Properties of high-transmittance and flexible low-emissivity coatings constituted by ITO-Ag stack*, Vacuum 5 (3) (2002) 2.

[8] E. Eby, R. O'Shaughnessy, R. Bond, High Transmittance, Low Emissivity Coatings for Substrates, 2006.

[9] F. Giovannetti, et al., High transmittance, low emissivity glass covers for flat plate collectors: applications and performance, Sol. Energy 104 (104) (2014) 52–59.

[10] R Saxena, L Dhar, D K Mohanty, et al., Study and Preparation of High Reflection Optical Coatings, Proc. IWPSD 2017 (2019).

[11] S.H. Jeong, et al., Characterization of SiO2 and TiO2 films prepared using rf magnetron sputtering and their application to anti-reflection coating, Vacuum 76 (4) (2004) 507–515.

[12] J. Robert, D.S.I. Crase, A.L.G. Dr, Margareta Hamel, *Optical Thin-Film Coating Methods*, Photonics Spectra, 2008.

[13] J. Liu, X. Ling, X. Liu, Mechanism of annealing effect on damage threshold enhancement of HfO2 films in vacuum, Vacuum 189 (2021), 110266.

[14] D.J. Barber, Radiation damage in ion-milled specimens: characteristics, effects and methods of damage limitation, Ultramicroscopy 52 (1) (1993) 101–125.

[15] K.J. Kim, C.S. Jung, T.E. Hong, A new method for the calibration of the vertical scale of a stylus profilometer using multiple delta-layer films, Meas. Sci. Technol. 18 (9) (2007) 2750.

[16] R.A. McAloney, et al., Atomic force microscopy studies of salt effects on polyelectrolyte multilayer film morphology, Langmuir 17 (21) (2001) 6655–6663.

[17] R. Kalish, Ion-implantation in diamond and diamond films: doping, damage effects and their applications, Appl. Surf. Sci. 117 (1997) 558–569.

[18] C. Yong, et al., Researches on laser damage resistance of optical films, High Power Laser Part Beams 28 (7) (2016).

[19] J.L. Fisher, P.K. Shukla, Detection of Coating Defects on Buried Pipelines Using Magnetic Field Variations within the Pipeline, 2017.

[20] K. Mikami, et al., Theoretical analysis for temperature dependence of laser- induced damage threshold of optical thin films, Journal of Physics Conference 688 (2016), 012065.

[21] G. Jinman, et al., Diagnosing laser-induced damage to optical thin films using peak sound pressure of shock waves, Laser Part. Beams 35 (2) (2017) 1–6.

[22] K. Mikami, et al., Laser-induced Damage Thresholds at Different Temperature for Optical Devices, International Society for Optics and Photonics, 2013.

[23] F. Elsen, et al., Demonstration of a 100-mJ OPO/OPA for future lidar applications and laser-induced damage threshold testing of optical components for MERLIN, Opt. Eng. 57 (2) (2018) 21205.1–21205.4.

[24] A. Shimada, et al., Development of integrated damage detection system for international America's Cup class yacht structures using a fiber optic distributed sensor, Technical Report of Ieice Oft 99 (2000) 7–10.

[25] M. Scheerer, Z. Djinovic, M. Tomic, Development, analyses and verification testing of a hybrid fiber optic system for deflection and damage detection of morphing wing structures, in: IMTC 2015, 2015.

[26] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[27] G. Zhou, C. Wang, Q. Mei, Using graph attention network to predicte urban traffic flow, in: 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM), IEEE, 2021.

[28] M. Tan, Q. Le Efficientnet, Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019.

[29] R. Chen, M. Wang, Y. Lai, Analysis of the role and robustness of artificial intelligence in commodity image recognition under deep learning neural network, PLoS One 15 (7) (2020) e0235783.

[30] H.B. Azad, S. Mekhilef, V.G. Ganapathy, Long-term wind speed forecasting and general pattern recognition using neural networks, IEEE Trans. Sustain. Energy 5 (2) (2014) 546–553.

[31] C. Ji, et al., Behavior inference based on joint node motion under the low quality and small-scale sample size, in: 2021 International Conference on Networking, Communications and Information Technology (NetCIT), IEEE, 2021.

[32] F. Sarghini, G. De Felice, S. Santini, Neural networks based subgrid scale modeling in large eddy simulations, Comput. Fluid 32 (1) (2003) 97–108.

[33] C. Zuo, et al., Deep learning in optical metrology:a review, 光:科学与应用(英文版) 11 (4) (2022) 54.

[34] Y.J. Cha, W. Choi, O. Büyüköztürk, Deep learning-based crack damage detection using convolutional neural networks, Comput. Aided Civ. Infrastruct. Eng. 32 (5) (2017) 361–378.

[35] L. Jian, et al., Determination of corrosion types from electrochemical noise by artificial neural networks, Int. J. Electrochem. Sci. 8 (2) (2013) 2365–2377.

[36] H. Wang, et al., Non-metallic coating thickness prediction using artificial neural network and support vector machine with time resolved thermography, Infrared Phys. Technol. 77 (2016) 316–324.

[37] V.S. Dave, K. Dutta, Neural network based models for software effort estimation: a review, Artif. Intell. Rev. 42 (2) (2014) 295–307.

[38] R. DeVore, B. Hanin, G. Petrova, Neural network approximation, Acta Numer. 30 (2021) 327–444.

[39] Oludare, et al., State-of-the-art in Artificial Neural Network Applications: A Survey, Heliyon, 2018.

[40] H.C. Zhang, S.H. Huang, Applications of neural networks in manufacturing: a state-of-the-art survey, Int. J. Prod. Res. 33 (3) (1995) 13.

[41] A. Sc, B. Pm, Part-of-Speech Tagging Enhancement to Natural Language Processing for Thai Wh-Question Classification with Deep Learning, 2021.

[42] M. Rhu, et al., vDNN: virtualized deep neural networks for scalable, memory-efficient neural network design, in: 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO). 2016, 2016.

[43] Z. Wu, et al., A comprehensive survey on graph neural networks, IEEE Transact. Neural Networks Learn. Syst. 32 (1) (2021) 4–24.

[44] Y. Zhang, J. Zhang, Application of fuzzy expert system in defect inspection of TFT-LCD, J. Optoelectron. - Laser 17 (6) (2006) 719–723.

[45] Y. Shi, et al., Defect inspection system design based on the automated optical inspection technique for LCD backlight modules, Chin. J. Sensors Actuat. 28 (5) (2015) 5.

[46] J.P. Yun, et al., Automatic defect inspection system for steel products with exhaustive dynamic encoding algorithm for searches, Opt. Eng. 58 (2) (2019) 23107.1–23107.9.

[47] N. Taherimakhsousi, et al., Quantifying defects in thin films using machine vision, npj Computational Materials 6 (1) (2020) 1–6.

[48] Y. Yang, et al., Delving into deep imbalanced regression, in: International Conference on Machine Learning, PMLR, 2021.

[49] J.A. Parker, R.V. Kenyon, D.E. Troxel, Comparison of interpolating methods for image resampling, IEEE Trans. Med. Imag. 2 (1) (1983) 31–39.

[50] S. Albawi, T.A. Mohammed, S. Al-Zawi, Understanding of a convolutional neural network, in: 2017 International Conference on Engineering and Technology (ICET), Ieee, 2017.

[51] K.T. Gribbon, D.G. Bailey, A novel approach to real-time bilinear interpolation, in: Proceedings. DELTA, Second IEEE International Workshop on Electronic Design, Test and Applications. 2004. IEEE, 2004.

[52] E.H. Adelson, et al., Pyramid methods in image processing, RCA engineer 29 (6) (1984) 33–41.

[53] W.-S. Lai, et al., Deep laplacian pyramid networks for fast and accurate super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[54] J. Deng, et al., Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009.

[55] K. He, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[56] Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, Adv. Neural Inf. Process. Syst. (2018) 31.

[57] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2014 arXiv preprint arXiv.
[58] T. Fushiki, Estimation of prediction error by using K-fold cross-validation, Stat. Comput. 21 (2) (2011) 137–146.
[59] R. Koenker, Quantile regression for longitudinal data, J. Multivariate Anal. 91 (1) (2004) 74–89.
[60] A. Mikołajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 International Interdisciplinary PhD Workshop (IIPhDW), IEEE, 2018.
[61] UnKnow, Intelligent Defect Detection System, Github, 2020.