Image Data Assessment Approach for Deep Learning-Based Metal Surface Defect-Detection Systems

HSIEN-I. LIN (Member, IEEE), AND FAUZY SATRIO WIBOWO
Graduate Institute of Automation Technology, National Taipei University of Technology, Taipei 10608, Taiwan
Corresponding author: Hsien-I. Lin (sofin@ntut.edu.tw)

This work was supported in part by the Ministry of Science and Technology under Grant MOST 108-2221-E-027-115-MY3.

ABSTRACT The current trend in automated optical inspection (AOI) systems employs deep learning models to detect defects on a metal surface. The setback of deep learning models is that they are time-consuming because the images obtained after every lighting adjustment must be used to train the deep learning models again and confirm whether the detection results have improved. To save the time spent using datasets to train deep networks, we proposed a comprehensive assessment score that combines defect visibility, visibility distribution, and overexposure based on the operation principles of convolution neural networks. It can be used to assess whether the training image dataset can improve the defect detection rate of the deep learning model such as You Only Look Once (YOLO), Single Shot MultiBox Detector (SSD), and Faster Region-based Convolutional Neural Network (Faster R-CNN) without training defect image datasets. We collected all of the weight combinations with the right prediction results and used linear regression to obtain the optimal weight coefficients. We found that visibility and overexposure had a greater impact on the comprehensive assessment score. We compared the proposed approach with existing image quality assessment methods, including mean absolute error (MAE), peak signal-to-noise ratio (PSNR), universal image quality index (UIQI), natural image quality evaluator (NIQE), perception-based quality evaluator (PIQE), and blind/referenceless image spatial quality evaluator (BRISQUE). The experiment results indicated that our proposed comprehensive assessment score is more correlated to the F2-score of the detection models than the IQA methods by the verification methods of Spearman Rank Correlation Coefficient (SRCC), Pearson Correlation, and Kendall Correlation. Thus, referring to this index during the collection of image data and choosing datasets with the highest score to train the model will produce better detection accuracy.

INDEX TERMS Deep learning model, defect detection, metal surface, comprehensive assessment score, image quality assessment.

I. INTRODUCTION
Metal products such as the casings of electronics are common in the consumer market, causing the quality and yield in the metal manufacturing industry to become increasingly important. The surface quality of metal pieces has a direct impact on how customers view the final product. Metal surface defect detection, which is a simple but highly repetitive and labor-intensive task, is usually performed manually by human eyes [1]. Due to slow detection speed, high labor costs, and visual acuity limitations, manual defect detection can no longer meet today’s industries demands. However, swift progress in image processing, machine vision, artificial intelligence, and other related fields [2], [3] have significantly enhanced the capabilities of visual inspection technology, thereby promoting the development of vision-based automated optical inspection (AOI) systems.

Before the emergence of deep learning, most AOI systems were based on image processing algorithms. Deep learning has been incorporated into the AOI systems of factories in recent years, with massive databases to train deep learning models. To achieve even better inspection results, the image capture and deep learning systems in AOI complement each other, and neither is dispensable. Even with a robust deep
learning detection system, poor image quality of the object or detection target would prevent the detection system from being effective.

Metal defects are one of the most difficult to detect. This is mainly because the defects that need to be detected on metal surfaces are relatively small and because there are many types of metal defects, such as holes, stains, scratches, dents, burs, and other defects of unknown origin in the production process. Different types of defects also show up very differently in images, so we narrowed down the scope of this study to scratches only. Even so, detecting scratches on reflective metal surfaces is still very challenging because even scratches by themselves have a wide variety of features, such as direction, size, and depth. These features make it challenging to design a suitable lighting system to augment each scratch feature. Consequently, the characteristics of features in images from industrial cameras may lack clarity, thereby increasing the difficulty of automated defect detection.

The current AOI system employs deep learning models to recognize defects. The imaging quality of defect datasets also exerts a direct impact on the recognition capabilities of the model [4], thereby making the acquisition of training sets fairly important. Using poor datasets for training would just be a waste of time. To evaluate the image quality of image datasets was conventionally performed by human subjective assessments. This study, therefore, focused on the means of conducting the pre-training objective assessment of defect image quality in the training sets for deep network recognition models so that time will not be wasted on training with poor datasets. Figure 1 shows that the role of image quality assessment helps deep learning models to learn effectively by choosing high-quality training datasets.

This paper is a continuation and extension of earlier work on image assessment for metallic surface [5] where a comprehensive assessment method based on (1) average visibility; (2) visibility distribution; (3) overexposure was proposed. However, this paper conducted the further analysis of our proposed method in different deep learning models such as You Only Look Once (YOLO), Single Shot MultiBox Detector (SSD), and Faster Region-based Convolutional Neural Network (Faster R-CNN) and validated that the proposed method was more correlated to the F2-score of the deep learning models than the non-reference and full-reference traditional image assessment algorithms.

The remainder of this paper is structured as follows. Section II presents the literature review. Section III introduces the hardware of the AOI, including robotic arm control, the imaging method, and the lighting configurations. Section IV explains the proposed approach, including the deep network model used and the proposed image assessment indices. Section V presents our experiment and an analysis of the results, and Section VI contains the conclusion and future directions of improvement.

II. LITERATURE REVIEW

Based on existing literature involving defect detection systems, the subjects of our investigation were divided into the lighting method and the recognition method, and existing image quality assessment methods were examined.

A. LIGHTING METHOD

The lighting equipment is a crucial component in machine vision systems. To ensure the consistency and reliability of machine vision systems when they perform measurements and recognition tasks, choosing the right lighting equipment based on different application needs is essential. Improper lighting methods will cause the image features to be unclear or incomplete, in which case post-processing will not be able improve or recover them. Obtaining the optimal lighting solution means considering several lighting control parameters, such as the geometry of the lighting (e.g., type, angle, and direction), the type of the light source (e.g., wavelength, color temperature, and luminous intensity), the characteristics of the detection surface (e.g., color, reflectance, and roughness), mechanical limitations (e.g., volume, convenience of installation, and service life), and installation costs.

There are numerous parameters to determine the lighting equipment, and these parameters also influence one another, increasing the difficulty of selecting and setting the lighting system. As a result, most light source types and settings in the industry are decided based on experience and the technical documents provided by the lighting manufacturer [6] to highlight the object features of interest. Thus, for any machine vision application, the lighting objectives can be divided into the following three:

- Maximizing the contrast of target features.
- Minimizing the contrast of features not of interest.
• Minimizing unpredictable changes caused by external environmental factors (such as ambient light).

B. DEFECT RECOGNITION METHOD

We discuss three different methods for defect recognition: image processing algorithms, machine learning, and deep learning.

1) DEFECT RECOGNITION BASED ON IMAGE ALGORITHMS

Xu et al. [7] employed an image filter and threshold value to detect defects on metal cylindrical surfaces. Out of 600 samples, their error rate was 2.3%; however, this low error rate was accompanied by much overkill, which was later improved in their study. To detect defects on non-ferrous metal surfaces, Senthikumar et al. [8] used spatial filtering to reduce noise and an iterative thresholding technique to obtain binary images of the defects. Any white pixels in the binary images were determined to be defects. They used three sample images to demonstrate that their method could accurately detect the shape of defects on non-ferrous metal surfaces. However, without a greater number and variety of samples to verify this, their method can only detect certain types of defects. To detect and analyze the defects on the surfaces of metal spoons, Li et al. [9] employed image filtering to eliminate noise, histogram equalization and the Laplace operator to enhance defect features, and Canny, Sobel, and Log edge detection algorithms to obtain the images of the spoon and defect edges. This study merely demonstrated that the spoon and defect edges could be detected in a few certain samples. They did not use a large number and variety of samples to test the accuracy of their approach, so the recognition capabilities of their approach with regard to other forms of defects are not yet confirmed.

Li and Ren [10] utilized local normalization and defect contour projection to detect surface defects on steel rails, and their proposed system can perform detection in real time on a train moving at high speeds. This study detected two types of defects with recalls of 93.1% and 80.41%, respectively, and lower accuracy at 72.78% and 81%. Yun et al. [11] used several line scan cameras to scan the surface defects of hot steel rods and employed an adaptive local binarization method and a discrete wavelet transform to filter out the background and separate the defects. Experiments with 1,550 defect images produced an accuracy of 84.7% and a false alarm rate of 1.5%, which are still somewhat inadequate for practical applications.

In conclusion, the image processing-based methods of defect detection, appropriate algorithms and parameters must be selected for the defects being detected, and the more general approach of classification is binarization.

2) DEFECT RECOGNITION BASED ON MACHINE LEARNING

Riaz et al. [12] proposed a K-means-based image segmentation technique to detect surface defects on castings, including cracks, blowholes, and pinholes. With images of these three types of defects on five different scales, the proposed method successfully separated defects from the background, but they did not go on to verify the recognition capabilities of the algorithm. Xue-Wu et al. [13] used a diffuse bright-field back light illumination for defect imaging and added a diffuser to reduce metal surface reflection. In addition, they used a more complex wavelet transform to filter out noise and extract features during image preprocessing and finally used a multi-class support vector machine (SVM) to classify various defects with highly reflective properties on metal surfaces, including scratches, oil stains, cracks, and holes. The accuracy of their approach for the test set reached 85%. They stated that the recognition capabilities of their approach were superior to those of back-propagation neural network (BPNN). They also pointed out that the test sample size was too small and needed further verification from a production line.

Kurokawa et al. [14] used a line scan camera to scan the surface of metal cylinders and then experimented with and compared various feature extraction methods, such as the discrete Fourier transform (DFT) and the discrete cosine transform (DCT). For classification, they used a one-class SVM to separate defects from the background. Their results indicated that with DCT as the feature extraction method, they could obtain a false alarm rate less than 1%; however, the miss rate was 40%. Choi et al. [15] developed a defect detection method based on a multi-class SVM for rolled steel strips and used a line scan camera and a linear light source to scan continuously rolling steel. They used image processing for positioning and obtained the geometric features of defects such as length, area, direction and grayscale features such as the maximum, minimum, and median for SVM inputs. This study considered six types of defects, and the accuracy of their classification ranged from 87% to 94%. Shanmugamani et al. [16] proposed an SVM-based approach to detect rust spots, corrosion, and normal wear on the surfaces of gun barrels. They employed histograms and gray-level co-occurrence matrices (GLCMs) to extract textural features and indicated that the accuracy of SVM classification is better than that of artificial neural networks and k nearest neighbors (k-NN). The recognition accuracy of their approach reached 96.67%.

In conclusion, machine learning algorithms have excellent recognition capabilities and require less time and data for training than deep learning does. However, they rely on manually designed feature extraction methods to achieve this effect.

3) DEFECT RECOGNITION BASED ON DEEP LEARNING

Soukup and Huber-Mork [17] used a line scan camera and light sources of two different colors to capture images of defects on rail surfaces. Their recognition method used a classical convolutional neural network (CNN) framework with an error rate of 1.108%. They also pointed out that they used very little data in their study and that multiple data augmentation methods can be employed to prevent overfitting in the model. Masci et al. [18] proposed a max-pooling CNN to detect surface defects in steel on the production line. When applied
to 646 test images, their method presented an error rate of 7%. Martinez et al. [19] employed bright field and dark field illumination for imaging and utilized an image processing algorithm to combine the image features of defects under different light sources so that different types of defects could present their features in images. For classification, a feed-forward artificial neural network with only a single hidden layer was used to determine whether scratches, bumps, dents, or machining scars were present on metal surfaces. The overall recognition accuracy of their model was 90.16%, but the detection rate for bumps was particularly low at 72.3%.

Cha et al. [20] developed a real-time vision detection method based on a faster region-based convolutional neural network (Faster R-CNN) to detect cracks, steel corrosion, and screw corrosion in concrete. The average precision (AP) of this approach with regard to all of the types of defects was 87.8%, and it needed only 0.03 seconds to process an image with 500×375 pixels. Using bright field illumination for imaging, Tao et al. [21] developed a recognition method using an autoencoder (AE) to position defects and determine whether they appear in the images. The local images of the defects are then input into a CNN for defect classification. The process did not need a manually designed image processing or feature extraction method at all. The accuracy of AE positioning was 89.6%, and the accuracy of CNN classification was 86.82%. They also compared their proposed model with other recognition algorithms and found its recognition capabilities to be superior.

Xiao et al. [22] employed a threshold value to extract the regions of interest (ROIs) and input them into fully convolutional networks (FCNs) to recognize surface defects on galvanized stamping parts. The average accuracy of their approach for the four types of defects reached 99.6%. The network architecture that they used is similar to the YOLO model employed in this study. They upscaled images to increase the dimensions of the feature map and combined the shallower feature map for the recognition of objects on different scales. Song et al. [23] developed an FCN based on a U-net framework to recognize minute surface defects with uneven grayscale distributions on metal plates and used two white LED bars as the lighting system. Unlike that of Xiao et al., the framework proposed by Song et al. replaced upsampling with deconvolution so that the model can extract more features. Moreover, it had a different number of layers and different filter sizes in each layer but basically still decreased and then increased the dimensions in the feature map. They derived an accuracy of 83.45% with the training set, and the mean error between prediction results and actual defect sizes was 0.29%.

In conclusion, the deep learning methods mentioned above performed well in defect recognition, but the greatest drawback was that they needed substantial amounts of training data and time, so the image processing methods were generally paired with semi-automated or fully-automated methods to label the training data needed for deep learning. Furthermore, most of the studies focused on improving the recognition capabilities and efficiency of their algorithms. The success of a deep learning model is still determined by the training set.

C. IMAGE QUALITY ASSESSMENT METHODS
1) SUBJECTIVE METHODS
Subjective assessment methods [24] that are widely considered to be accurate in image quality evaluation involve having participants giving opinion scores to image quality without knowing the scores given by others or any image quality information. The opinion scores are divided into five levels, with each level representing a different degree of subjective judgment. The average of the scores given by participants then serves as the basis of reference for image quality. Such methods require opinion scores from people, and results from more people will have greater reference value.

2) OBJECTIVE METHODS
Objective methods for image quality assessment (IQA) use numerical methods to quantify the image quality. Figure 2 shows the categorization of IQA methods. Full reference approaches compare the input image against an original reference image with no distortion; meanwhile, no-reference approaches compare the statistical features of the input image with a set of features acquired from an image database.

FIGURE 2. Categorization of IQA methods.

In the past, statistical error metrics were mostly used to evaluate image quality. These included calculating the mean absolute error (MAE) of a high quality image and the image being evaluated, the peak mean square error (PMSE) [25] of the two images, or the peak signal-to-noise ratio (PSNR) [26]. Image quality assessment methods that are more practical and widely used than statistical error metrics are based on human visual systems (HVSs), such as universal image quality indices (UIQIs) [25], structural similarity index measures (SSIMs) [27], and dissimilarity structural similarity indices (DSSIMs) [28]. HVS-based methods, as their name suggests, come from human observations of images and quantify the human sense of sight to assess image quality, encompassing brightness, contrast, texture, and exposure. However, the aforementioned statistical error metrics and HVS methods were mostly used to assess the quality of natural images in past studies. They are not really suitable
for defect detection because defects in images are presented differently from objects in natural images. Furthermore, these two types of methods require the definition of a high quality image beforehand, which is not possible in most industrial applications.

Some studies focused on objective image quality assessments on natural images. Wu et al. [29] proposed a novel blind image quality assessment (BIQA) method by selecting statistical features extracted from binary patterns of local image structures. This method demonstrated a better performance than a real-time approach on public datasets such as LIVE II and TID2008. They aimed to develop efficient methods for BIQA without any training process. Xue et al. [30] presented a novel general-purpose BIQA approach free of the human subjective scores in learning. Their method converted a partition of the distorted images into overlapped patches and used a percentile pooling strategy to estimate the local quality of each patch. Mittal et al. [31] proposed a natural image quality evaluator (NIQE) based on the construction of a ‘quality-aware’ collection of statistical features on a simple natural scene statistic (NSS) model. It assessed the image quality without the knowledge of anticipated distortions or human opinions. Similarly, Venkatanath et al. [32] proposed perception-based quality evaluator (PIQE) using an unaware opinion method to quantify distortion without training data. Their method extracted local features to predict the quality, especially significant spatial regions. They tested their method on natural image datasets such as LIVE IQA, TID, and CSIQ. Mittal et al. [33] also proposed a method to quantify the loss of the “naturalness” of the image using a locally normalized luminance coefficient instead of distorting-specific features. They showed a better performance than full reference methods such as the peak signal-to-noise ratio and the structural similarity index. Blind/referenceless image spatial quality evaluator (BRISQUE) [33] was also an IQA approach using scene statistics of locally normalized luminance coefficients. However, most of the previous assessments for natural images were not suitable for metallic objects.

Jayant et al. [34] provided a dataset of document images taken with a camera, the images containing varying degrees of blurriness. They used two sharpness indices [35], [36] and an unsupervised feature learning framework [37] to assess the quality of the document images and predict the accuracy of optical character recognition (OCR) and then compared the advantages and disadvantages of the three methods. Li and Pan [38] used standard deviation, entropy, and average gradient as image assessment indices to evaluate the effectiveness of image fusion techniques in improving the quality of remote sensing images. They also used the assessment indices to examine the advantages and disadvantages of the different image fusion techniques and select suitable image fusion techniques for different applications. Hossain et al. [39] developed a support vector regression (SVR) framework to predict the quality of infrared images obtained by unmanned aerial vehicles.

In past quality assessment methods for natural images, the contrast, sharpness, and brightness are usually taken into consideration but they are not sufficient for defect detection on metal surfaces. Also, in previous methods, high quality images must be predetermined and then compared with the target image to gauge image quality. However, high quality images can rarely be predetermined in practical applications. A number of studies chose to extract target features to assess image quality, and their respective experiments on image datasets all obtained better results than conventional image quality assessment methods. Furthermore, conventional assessment methods only considered the errors, means, and standard deviations in images and could not present the image features extracted by convolutional operations. This study therefore used the operational features of CNNs simulating human vision to calculate the visibility of defects in images, the distributions of defect visibility in image datasets, and the overexposure in each image. By quantifying the defect imaging results in the each image database, we proposed a comprehensive assessment score to predict the F2-Score of using the database in question to train a deep network model, thereby saving the time needed for model training. The primary contributions of this study are as follows:

- This study focused on the image dataset assessment for defects on metal surfaces.
- This study used comprehensive assessment scores to evaluate the defect visibility in images and exposure in datasets to predict model recognition results. The dataset with the highest assessment score was used to train the model to obtain optimal recognition capabilities, thereby saving the time needed to repeatedly train and test the recognition model.
- The proposed approach was compared with existing image assessment methods, including full reference and no reference assessments such as MAE, PSNR, UIQI, NIQE, PIQE, and BRISQUE. The experiment results indicated that the proposed assessment method were better predictors for the F2-scores of recognition models.
- This study designed assessment indices by observing the weights and numbers of filters during the network operation process and the resulting feature maps, and allowed us to use another perspective to gain a deeper understanding of what the weights and feature maps in the network represent.
- Inexperienced users will be able to adjust the lighting setting based on the scores of the assessment indices.
III. AOI SYSTEM HARDWARE

Figure 3 displays the configurations of the optical device of the AOI system. In this study, the camera and lighting system are fixed at the end of the robotic arm using a specially designed jig made of rigid and lightweight aluminum. To make the lighting system in this study useful for different samples, we specifically designed the fixtures of the lighting to be adjustable. Two sets of height control knobs allow the heights of the dome and square ring lights to be adjusted separately, the height difference between the highest and lowest settings being 70 mm.

FIGURE 3. Configurations of optical engine of AOI system.

1) Height control knobs of square ring light
2) Height control knobs of dome light
3) Basler industrial camera
4) 12 mm lens
5) Dome light
6) Square ring light
7) Flange connection

A. CAMERA AND LENS

Assessments were made with two objectives: the target defects that needed to be detected in this study were 0.2 mm$^2$ or larger in size, and the scanning motions of the robotic arm had to be as fast as possible. On account of a convolutional deep neural network being the primary recognition algorithm in this study, a high-resolution camera would provide the network model with more defect features and make recognition easier. Another consideration was the scanning time; with the amount of area to be scanned being fixed, a larger field of view (FOV) would mean fewer times moving the robotic arm to extract images of different regions, which implies less time needed to scan the target object. We chose one of Basler’s industrial cameras with 10 MP resolution. The objects that were inspected were metal plates with an area of 349.3 $\times$ 240.7 mm. In consideration of the space needed by the lighting system, the working distance of the camera was set at 154 mm. Our goals were to widen the FOV as much as possible to shorten the time needed by the system to inspect samples and to increase the numbers of pixels in the defects as much as possible to facilitate defect feature extraction. However, we cannot have both, as widening the FOV would reduce the number of pixels in the defects, and vice versa. We used a 12 mm lens, and the FOV was 82.1 $\times$ 59 mm. Thus, 36 images were needed to scan an entire target.

B. MOTION CONTROL OF ROBOTIC ARM

The robotic arm used in this study was Yaskawa’s GP7, a six-axis industrial robotic arm. The robotic arm was required to move the camera to different regions of each sample for image capture and recognition, and the local images would ultimately form a global image. Preliminary motion control of sample scanning was performed on the X-Y plane of the Cartesian coordinate system. Figure 4 shows the relationship between the image capture locations and the images when the robotic arm scans the samples. The dashed rectangle represents the inspection sample, the size of which is 349.3 $\times$ 240.7 mm. The red square boxes present the sample images that are taken. Each square box is 1,600 $\times$ 1,600 pixels, which is actually 37 $\times$ 37 mm. The red circle in each box indicate the center of said box, and the blue areas show where the captured images overlap. In addition, as the figure shows, a fixed margin was kept between the image boundaries and the sample edges to prevent the edges of the sample from falling outside of the images in case the sample is not placed in the exact center. The width of the margin can be adjusted; a wider margin means a larger error tolerance, whereas a narrower margin means a smaller error tolerance. In this study, we set the margin width as 4 mm.

FIGURE 4. Diagram of locations of sample images.

Figure 5 shows that a traditional snake scan or a spiral scan can be used for the scanning path of the robotic arm; whichever scanning path is used does not affect the detection speed. We adopted the traditional snake scan path because it is more intuitive and makes it easy to observe the scanning conditions.
C. LIGHTING CONFIGURATIONS

Two types of light sources were used in this study. One was a blue LED dome light, which can provide uniform and indirect light and is thus also known as overcast lighting. The other was a blue LED square ring light, which comprises four light bars of the same length and width. As the light bars provide direct light, we added a diffuser to each bar to reduce the glare of light reflecting off the metal samples. These two lights were coordinated with a six-channel light controller. We chose these two types of lights for the lighting system because the effective illumination area of the dome light is smaller than the image FOV, which means that the dome light cannot provide uniform light at the edges of the images, thereby creating dark regions. If the defect is located near the edges of the images, its features would not show up on the images. For this reason, we used a square ring light to make up for the inadequacies here and enable the features of defects anywhere in the images to become visible.

IV. PROPOSED METHOD

A. DEEP NETWORK RECOGNITION MODEL

The entire recognition process of the AOI system is divided into two parts: data collection and actual inspection. The input images of the system comprise several images obtained from scanning the sample using the robotic arm. The number of images is determined based on the size of the samples and the FOV of the camera. During data collection, the original images are first preprocessed. In this case, preprocessing merely involves cropping the images to reduce the size of the original images. This reduces the number of parameters in the model, shortens training time, and also increases the number of images in the training set. Subsequently, we can label the dataset, perform configuration training, test the dataset, and then train the recognition model. For the actual inspection part, several input images undergo defect detection using the recognition model, and then they undergo the final image post-processing. The local images of the sample are patched together into a complete image, and the defect detection results are drawn on the image. The complete image is then displayed on the screen for user reference.

We employed a YOLO algorithm framework [40] as a reference. Although this algorithm is less accurate than the other object recognition algorithms when it comes to natural images, its faster calculation speed meets the real-time calculation needs of industrial applications. Furthermore, an open source C++ development platform that is available can be easily integrated with the industrial camera, robotic arm, and lighting system in this study. YOLO can predict and position objects in images more swiftly than other object detection algorithms because it inputs entire images into fully convolutional networks (FCN) at a time for feature extraction and directly uses the feature map output by the last layer of the network to predict the location, size, and classification of object bounding boxes. The model framework that was actually used in this study was the YOLOv3 lightweight model with the input and output dimensions revised to meet the application needs of the defect detection task. The \(1 \times 1\) convolutional layer was used to reduce the dimensions of the feature map and decrease model calculations without severely affecting recognition accuracy [41].

In this study, the assessment of model performance in defect recognition applications is mainly based on precision and recall. Precision is defined as the proportion of model defect predictions that are actually defects, whereas recall is defined as the proportion of all labeled defects that are also predicted as defects by the model. During the calculations of precision and recall, two conditions determine the standard of model defect recognition. One is that the Intersection over Union (IoU) of the bounding box of defects predicted by the model and the bounding box of labeled defects must be greater than 0.5. The other is that the confidence value of the bounding box of defects predicted by the model must be greater than 0.5. When the output results of the model meet both of these conditions, then the model is considered to be accurate. Due to the inverse relationship between precision and recall, we used the F2-score as a more intuitive way of assessing and comparing model performance.

B. IMAGE QUALITY ASSESSMENT INDICES

This study proposed an assessment score to predict the recognition capabilities of the model when training with different datasets. In this section, we will describe the calculation methods of the three indices in detail and explain why the indices are used to assess model performance. Finally, we will explain the calculation method of the comprehensive assessment score combining the three indices.

1) DEFECT VISIBILITY

Based on the perspectives of HVSs, we assume that humans can recognize defects more easily when the defects differ more greatly from the background. Thus, a CNN developed to simulate the human sense of sight should be the same. Each convolution layer in neural networks have several filters to detect different shapes, colors, and textures in images, and deeper convolution layers can detect more complex features in images. Take the convolution operations in Fig. 6 as an example. On the left are black-and-white images with an X. In the middle are three \(3 \times 3\) filters respectively used to detect whether there are diagonal lines from the upper left to the lower right, intersections of diagonal lines, and diagonal lines from the upper right to the lower left in the input
represents the number of defects in each image; \( N \) where \( M \) is the number of images in the dataset; \( p_{\text{max}} \) is the maximum value of pixel intensity; \( \bar{x}_{ij} \) denotes the mean pixel intensity of defects; \( x'_{ij} \) is the pixel intensity of the background surrounding the pixels; \( \bar{x} \) and \( \bar{x}' \) are respectively defined as:

\[
\begin{align*}
\bar{x} &= \frac{1}{WH} \sum_{i=x_{\text{defect}}}^{W} \sum_{j=y_{\text{defect}}}^{H} I_{ij} \times M_{ij} \\
\bar{x}' &= \frac{1}{4WH} \sum_{i=x_{\text{defect}}-2W}^{2W} \sum_{j=y_{\text{defect}}-2H}^{2H} I_{ij} \times M'_{ij}
\end{align*}
\]

In Eqs. (2), \( W \) and \( H \) denote the width and height of the bounding boxes of the defects; \((c_x, c_y)\) are the coordinates of the center point of the defect; \( I \) represents the original image, and \( M \) is the pixel-level label image, as shown in Figs. 7(a) and (b); \( M' \) is the inverse image of \( M \). We used the regions encompassed by the bounding boxes of the defects during model training to calculate the mean pixel intensity of the defects, such as the regions in the blue boxes in Fig. 7(a). For the mean value of the background, we only considered the areas around the center points of the defects that were four times the area of the bounding boxes of the defects rather than the entire image, such as the regions in the red boxes in Fig. 7(a). This can prevent the black background areas that are not of the sample in Fig. 7(c) from lowering the mean intensities of the pixels at the edges of the sample, which would increase the pixel intensity differences between the defects and the background and affect their visibility.

2) DISTRIBUTION OF DEFECT VISIBILITY
This study employed the lightweight version of the YOLO model as the reference predictor of dataset quality. The framework of this model is simpler, and when the features of the defects in the training set are more concentrated, the model will more easily learn and detect defects with better accuracy. In addition, the SSD and Faster R-CNN were also used to validate our proposed assessment approach. We therefore included the discreteness of visibility distributions as one of the assessment considerations. The mathematical definition of visibility discreteness is presented in Eq. (3), where \( n \) is the total number of defects in the dataset; \( v_a \) is the mean visibility of the defects in the sample dataset shown in the previous section, and \( x_i \) represents the visibility of defect \( i \) in the dataset.

\[
s = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - v_a)^2} \in [0, 1].
\]

The visibility distributions of the defects in each dataset are as shown in Figs. 8(a) through 8(f), in which Dataset 1 to Dataset 6 were obtained using dark to bright light settings. From Dataset 1 to Dataset 6, the visibility distributions become more dispersed; in other words, brighter light settings make visibility distributions more dispersed. This is because the pixel intensity differences between deeper defects and the background remain with brighter light settings, whereas the
3) IMAGE OVEREXPOSURE

In the assessment of defect image quality, image overexposure is a crucial index. This is because metal is a material with high reflection coefficients. Illuminating its surface with light generally creates a large amount of reflected light, and overexposure can cause the defect features that we need to be unclear or even disappear from the image. Figures 9(a) through 9(d) display sample images with low to high overexposure. The blue box in the upper left shows a deeper scratch defect, whereas the one in the lower right shows a shallower scratch defect.

A comparison of the sample images with varying degrees of overexposure shows that with the high overexposure conditions in Fig. 9, the features of the shallower defect are eroded by the reflected light and thus do not appear in the image. The deeper defect also is visible in the image but also presents erosion from the reflected light, which may cause the scratch to become fragment in shape or even cut into two defects. In the other conditions with lower degrees of overexposure, the features of the defects are completely visible. Although the shallower defect does not stand out significantly from the background, it is still discernible to the naked eye. Furthermore, due to the fact that the light source was set fairly close to the inspection samples in this study, overly bright light from the light source caused partial overexposure in the images. Defects of any depth may appear anywhere in the image, creating uncertainty in defect features, and as a result, overexposure needs to be one of the indices of image quality assessment.

Past studies mostly set a threshold value and then employed binarization to directly divide images into overexposed pixels and non-overexposed pixels and quantify the degree of overexposure in the images [42]. Grayscale images have only one channel, and the maximum pixel intensity is 255, so the threshold value is generally 250 or higher. However, observation of the images in the datasets revealed that such a setting would not be able to quantify the overexposure in Dataset 5, which has a weaker impact on the presentation of defect features in the image. We therefore slightly lowered the threshold value to 245 so that more bright pixels that influenced defect features would be classified as overexposed pixels. The white regions in Figs. 10(a) and 10(b) are the overexposed regions in the images of Dataset 5 and Dataset 6, respectively. The remaining images in the dataset were not overexposed, and the results of their binarization were all pixel intensity differences between shallower defects and the background decrease and even approach 0.
The first step of developing a machine vision-based automated inspection system is assessing the imaging effects, which is also a crucial stage determining the success of the system. Generally, suitable light sources are chosen for the lighting system based on design experience, and the imaging effects are assessed using observations with the naked eye. However, this can be quite difficult and subjective for beginners, which means that an objective numerical approach is needed for assessment. When we used the three indices above separately to assess dataset quality, we cannot effectively describe the recognition results of the model. This is because the datasets indicated as having the best quality by mean defect visibility, visibility distributions, or overexposure individually differ from those indicated by the F2-Score of the model. Thus, any index cannot be used alone for assessment. All three quantified index must be considered at the same time to assess dataset quality. A higher mean visibility is better so that the model is provided clearer features for extraction and recognition, while lower visibility discreteness and image overexposure are better. The comprehensive score combining the three indices was defined as follows:

\[
    \text{score} = \alpha V_d + \beta (1 - s) + \gamma (1 - E_d) \tag{4}
\]

where \(V_d\), \(s\), and \(E_d\) respectively denote the mean visibility, standard deviation, and overexposure of the datasets, and \(\alpha\), \(\beta\), \(\gamma\) are the corresponding weight coefficients. In our experiment, we used \(\alpha = \beta = \gamma = 1/3\). These weight coefficients represent the impact of the indices on the accuracy of the assessment results.

V. EXPERIMENT AND ANALYSIS

A. EXPERIMENT

The first part compared the assessment index and recognition results to understand the explanatory power of each assessment index and the comprehensive assessment score on the recognition results of the model and determine whether the quantified indices have reference value. The second part examined the results of the comprehensive assessment score with regard to the recognition capabilities of detection models for other objects to further understand the applicability of the assessment approach in this study to different recognition methods. The third part focuses on the weight coefficients in the comprehensive assessment score to confirm the influences of the individual weight coefficients on the assessment results and discuss the most suitable weight coefficients for the assessment of model recognition capabilities. The fourth part compares the proposed approach with existing image quality assessment methods to verify its efficacy in the field of image assessment.

To verify the correlation between the quality of different image datasets compared by the proposed assessment approach and model recognition results, we collected datasets of sample images with varying degrees of brightness by moving the camera using the robotic arm and by adjusting the PWM output values (0–255) of the light source controller. We thus obtained datasets of images with varying visibility, visibility distributions, and overexposure. The samples used in this experiment were two anodized aluminum casings. Figure 11 displays an actual sample, and the defects were bumps and scratches manually applied to the surfaces.

The FOV of the camera was 37 \(\times\) 37 mm. Using the image capture method mentioned in Section III, a total of 80 original images, each with 1,600 \(\times\) 1,600 pixels, were captured for each sample. Each dataset in this study contained 160 original images, and the defects in each image varied in form and number, including scratches of different depths, sizes, and direction. Image labeling was performed manually using a
labeling tool with pixels as the unit, and a program written before hand was used to convert the labels into the bounding box information needed for YOLO. Another matter of particular note was that the placing of the sample for each image capture had to fit the coordinate positioning of the robotic arm; otherwise, positioning errors would cause defects to be labeled at the wrong places.

We used the NVIDIA RTX2080 Ti GPU to accelerate the training of the deep learning model. Even so, training a detection model in which each input image had a resolution of 1,600 × 1,600 pixels required 16 hours to complete the 10,000 iterative calculations of a model, and each single image required roughly 80 ms to process. However, the training and recognition time for such a large model was too long, so we divided each 1,600 × 1,600-pixel image into four 800 × 800-pixel images for training. This shortened the training time of a model by approximately 4 hours, and each single image required roughly 20 ms to process. Another advantage of this was that it increased the number original images, increasing the number of images in each dataset from 160 to 640. We used 80% of all of the original images as the training set, and 20% as the test set. This 80-20 (train-test) ratio was suggested from the investigation of Gholamy et al. [43] to avoid overfitting from the model.

In conclusion, the training data comprised 512 images that contained 2,162 defects, and the test data included 128 images that contained 513 defects.

When training the recognition model of each dataset, we used rotations at random angles between -90 and 90 degrees and three types of mirror effects (horizontal flips, vertical flips, and both horizontal and vertical flips) to augment the data. With a batch size of 64 and 10,000 iterations, each model was trained using 640,000 images. Table 1 lists the final average loss of the models after training with our six datasets. The average loss can be used as a preliminary assessment of the training results. However, the assessment is only applicable to the computation results of the training sets and does not represent the final test results. Theoretically speaking, a lower average loss means that the model has better recognition capabilities. Nevertheless, overfitting may occur in the neural network, which means that we still have to refer to the recognition capabilities of the model with regard to the test set to obtain the optimal training results.

**TABLE 1. Average loss after training with six datasets.**

| Dataset 1 | Dataset 2 | Dataset 3 |
|-----------|-----------|-----------|
| Ave. Loss | 1.023     | 0.966     | 0.922     |

| Dataset 4 | Dataset 5 | Dataset 6 |
|-----------|-----------|-----------|
| Ave. Loss | 0.932     | 1.632     | 2.581     |

Spearman rank correlation coefficient (SRCC) was applied to IQA approaches to investigate whether they had a relationship with the F2-score of the deep learning model using a monotonic function. SRCC could be any value from −1 to 1 where 1 was a perfect positive correlation and vice versa. SRCC is express as:

\[
\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}
\]

where \(d_i\) is the difference between a pair of rankings and \(n\) is the number of observations. To deal with the tied rankings, the full version of SRCC to be applied is described as:

\[
\rho = \frac{\frac{1}{n} \sum_{i=1}^{n} [R(x_i) - R(\bar{x})] \cdot [R(y_i) - R(\bar{y})]}{\sqrt{\left[ \frac{1}{n} \sum_{i=1}^{n} (R(x_i) - R(\bar{x}))^2 \right] \cdot \left[ \frac{1}{n} \sum_{i=1}^{n} (R(y_i) - R(\bar{y}))^2 \right]}}
\]

where \(R(x)\) and \(R(y)\) are the rankings of the x and y variables and \(R(\bar{x})\) and \(R(\bar{y})\) are the mean rankings.

In addition, the Kendall relationship was also used to examine the ordinal association between two measured quantities. Kendall correlation is expressed as:

\[
\tau = \frac{\# \text{ concordant pairs} - \# \text{ discordant pairs}}{\frac{n(n-1)}{2}}
\]

where the output is close to 1 when the two variables have high similarity.

**B. COMPARISON OF ASSESSMENT INDICES REGARDING RECOGNITION RESULTS OF DIFFERENT DEEP LEARNING NETWORK MODELS**

1) YOLO

Under different brightness settings, we obtained six datasets to train six recognition models. Figure 12 presents the precision, recall, and F2-score of each model with regard to their respective test sets. As can be seen, the highest recall among the models was only about 78% and did not reach required standards. This is because the lighting equipment and setup used could not show a portion of the defects in the images even without overexposure. Consequently, the recognition capabilities of the models were not very good. The sample defects that did not show up in the images were mostly fairly minor scratches or bumps. On metal surfaces, they hardly had any depth and were imperceptible by touch, and some could even be described as faint changes in color. As a result, the imaging of these defects was difficult. Table 2 shows the quantification results of visibility, visibility distribution, over-exposure, and comprehensive assessment score described in Section IV with regard to the image quality in the datasets.

To compare the explanatory powers of the assessment indices with regard to the model recognition results, Table 2 ranks the six datasets based on the four indices and the F2-scores of the models. Based on the F2-score, the datasets with the best and poorest quality were Dataset 3 and Dataset 6, respectively. If only visibility is used to assess dataset quality, the datasets with the best and poorest quality would be Dataset 5 and Dataset 1, whereas Dataset 3 and Dataset 6 would be ranked third and fourth. This is clearly inconsistent with actual model recognition results, which means that defect visibility cannot be used alone for...
Fig. 12. Recognition performance of YOLO models for various datasets.

Table 2. Quantification results of indices with regard to dataset quality.

|       | Dataset 1 | Dataset 2 | Dataset 3 |
|-------|-----------|-----------|-----------|
| \( v_a \) | 0.053     | 0.103     | 0.157     |
| \( s \)  | 0.025     | 0.046     | 0.069     |
| \( E_a \) | 0         | 0         | 0         |
| Score  | 67.6      | 68.6      | 69.6      |

Table 3. Three correlation coefficients on YOLO, SSD, and faster R-CNN models.

|       | YOLO | Faster R-CNN | SSD |
|-------|------|--------------|-----|
| Spearman | 0.717 | 0.717       | 0.717 |
| Person  | 0.928 | 0.748       | 0.943 |
| Kendall | 0.866 | 0.866       | 0.866 |

Table 4. Pearson correlation coefficient on the proposed method and F2-YOLO.

|       | D1 | D2 | D3 | D4 | D5 | D6 |
|-------|----|----|----|----|----|----|
| F2    | 64.9 | 76 | 80 | 79.4 | 75.4 | 47.3 |
| \( va \) | 0.053 | 0.103 | 0.157 | 0.212 | 0.264 | 0.133 |
| \( s \)  | 0.025 | 0.046 | 0.069 | 0.093 | 0.115 | 0.12 |
| \( E_a \) | 0 | 0 | 0 | 0.107 | 0.904 |
| Score  | 67.6 | 68.5 | 69.5 | 70.3 | 67.6 | 37.0 |

Table 5. Spearman correlation coefficient ranking on the proposed method and F2-YOLO.

|       | D1 | D2 | D3 | D4 | D5 | D6 |
|-------|----|----|----|----|----|----|
| F2-YOLO | 5 | 3 | 1 | 2 | 4 | 6 |
| \( va \) | 6 | 5 | 3 | 2 | 1 | 4 |
| \( s \)  | 1 | 2 | 3 | 4 | 5 | 6 |
| \( E_a \) | 4.5 | 4.5 | 4.5 | 4.5 | 4.5 | 2 |
| Proposed | 5 | 3 | 2 | 1 | 4 | 6 |

SRCC value of F2 (YOLO) - Proposed = 0.717

This is also different from human instinctive assertions; generally, when a human is recognizing objects, the clearer the object, the easier it is to recognize it. The recognition capabilities of a deep learning model should also follow this principle, but the comparison results did not match expectations.

Similarly, using visibility distributions or overexposure alone to assess data and quality is even less feasible because these two indices indicate that Dataset 1 and Dataset 2 rank first and second in quality; however, the models actually trained using these two datasets only ranked fifth and third in recognition capability. Because using any of these three indices alone cannot represent the recognition capabilities of the models, we used a comprehensive assessment score that combines these three indices to evaluate datasets.

Table 3 shows that the Spearman, Person, and Kendall correlation coefficients between the proposed score and the F2-score are high in YOLO, SSD, and Faster R-CNN models. As shown in Table 4 and 5, the Pearson and Spearman correlations of the proposed method are 0.928 and 0.717, respectively and the dataset rank based on the proposed comprehensive assessment score is almost entirely identical to that of the recognition results. In Table 5, we also ranked the performance of our assessment method and the F2-score using the SRCC. From these Tables, our proposed assessment score is highly correlated to the F2-score of deep learning prediction.

In addition, we conducted a T-test to determine whether the difference between these F2-scores was statistically significant. When the T-value of the two samples is greater than the critical value \( T_{\alpha, \ df} \), then a significant difference is believed to exist between the two samples. The critical value is determined by the significance level \( \alpha \) and the degrees of freedom of the samples \( \df \). The 0.05 level of significance is generally used, and the samples in this study had 1024 degrees of freedom. Thus, the critical value \( T_{0.05, 1024} \) was 1.645. Table 6 shows the T-values of each dataset pair. As can be seen, the T-value of Dataset 3 and Dataset 4 is less than the critical value, which means that the reverse ranking of these two datasets does not affect the accuracy of the assessment results. We could therefore use this index to assess dataset quality and choose the best dataset for model training.
TABLE 6. T-Test results of YOLO model where D1 TO D6 represent dataset 1 to dataset 6, respectively.

| Defect No. | D1  | D2  | D3  | D4  | D5  | D6  |
|------------|-----|-----|-----|-----|-----|-----|
| True Pos.  | 513 | 513 | 513 | 513 | 513 | 513 |
| D1         | X   | 4.35| 6.13| 5.82| 4.21| 4.21|
| D2         | 4.35| X   | 1.75| 1.45| 0.14| 10.65|
| D3         | 6.13| 1.75| X   | 0.30| 1.89| 12.59|
| D4         | 5.82| 1.45| 0.30| X   | 1.59| 12.25|
| D5         | 4.21| 0.14| 1.89| 1.59| X   | 10.49|
| D6         | 4.21| 10.65| 12.59| 12.25| 10.49| X   |

2) SSD AND FASTER R-CNN MODELS

The experiment in the previous section demonstrated that our comprehensive assessment score is effective in assessing the accuracy of YOLO models. However, in the field of machine vision inspection, the recognition model that is used varies with the application. We therefore focused on comparing the more commonly used Single Shot MultiBox Detector (SSD) [44] and faster region-based convolutional neural network (Faster R-CNN) [45] object detection algorithms in the experiment in this section. This gave us an understanding on whether the comprehensive assessment score proposed based on the principals of neural networks can still identify datasets that can enable models to have the best or worst recognition capabilities with the recognition algorithms of different network architectures. For the SSD model, we used SSD-MobileNet V2, whereas for the Faster R-CNN model, we employed Faster R-CNN Inception V2. The same six datasets and data augmentation methods as those in the experiment in the previous section were used to train the models. Figures 13 and 14 compare the recognition results of the two models. As shown in these two figures, the recognition capabilities of SSD and Faster R-CNN with regard to the various datasets presented trends similar to those of YOLO. Datasets collected when the lighting was too bright or too dark resulted in poorer recognition results, whereas moderate lighting contributed to better recognition results.

Table 7 and 8 rank the six datasets based on the actual recognition results of the YOLO, SSD, and Faster R-CNN models after they were trained using one of the six datasets. As can be seen, the dataset ranked third in the two models was ranked fourth by our comprehensive assessment score, whereas the dataset that was ranked fourth in the two models was ranked third by our comprehensive assessment score. The rankings of the remaining datasets were identical, Dataset 4 being the best dataset and Dataset 6 being the worst. Also, it is shown that SRCCs in Table 8 are all positively correlated.

Similarly, we conducted a T-test to determine whether the F2-score difference was statistically significant and understand the reference value of assessment index accuracy. Tables 9 and 10 display the T-values of each dataset pair in the two models. The critical value was $(T_{0.05}, 1024) = 1.645$, and the T-values of Datasets 2~5 in the two models were all less than the critical value. Thus, the assessment results of the comprehensive assessment score were consistent with the actual results of these two models.

The comparison of these two models show a positive correlation between the rankings based on the comprehensive assessment index accuracy.
assessment score and those based on the recognition capabilities of the models. However, in terms of the values, some of the F2-scores of the three models presented no significant differences. In the SSD model, the T-values of Dataset 2 and Dataset 5 and of Dataset 3 and Dataset 4 were less than the critical value. In the Faster R-CNN model, no significant differences existed among Dataset 3, Dataset 4, and Dataset 5. Thus, a higher value in the comprehensive assessment score proposed in this study merely indicates that the F2-score of the model will be higher but does not guarantee that its recognition capabilities will reach a certain level. In other words, there is no absolute mapping relationship between the comprehensive assessment score and the F2-score; our assessment approach can only be used to compare dataset quality.

### C. OPTIMAL WEIGHT COEFFICIENTS OF COMPREHENSIVE ASSESSMENT SCORE

Regarding the weight coefficients in the comprehensive assessment index mentioned in Section III, we used identical weights for the three indices in the previous experiments, i.e., $\alpha = \beta = \gamma = 1/3$. These two experiments demonstrated that using the comprehensive assessment score to assess dataset quality is effective; however, these weight coefficients may not make the assessment results the most accurate. We adjusted the values of the weight coefficients, observed the resulting dataset quality rankings, collected all of the weight combinations with the right rankings, and used linear regression to obtain the optimal weight coefficients.

This also gave us an understanding of the importance of each quantified index in image quality assessment. Further analysis was conducted by increasing the sample points of weight coefficient settings, which were $\alpha = [0, 0.02, 0.04, \ldots, 1]$, $\beta = [0, 0.02, 0.04, \ldots, 1]$, and $\gamma = [0, 0.02, 0.04, \ldots, 1]$. Figure 15 displays the visualized results in three-dimensional space. The triangular plane formed by the red and blue dots satisfies $\alpha + \beta + \gamma = 1$. The three corner points show the assessment score results when defect visibility ($\alpha = 1$, $\beta = \gamma = 0$), visibility distribution ($\beta = 1, \alpha = \gamma = 0$), or overexposure ($\gamma = 1, \alpha = \beta = 0$) is used alone to calculate the assessment score. The red dots indicate the weight coefficient values where the assessment results completely match the recognition results of the model, whereas the blue dots indicate the weight coefficients where the assessment results are incorrect. Next, we used linear regression to obtain the optimal weight coefficients for the defect visibility ($v_d$), visibility distribution ($s$), and overexposure ($E_d$) of the datasets, which were $\alpha = 0.378$, $\beta = 0.229$, and $\gamma = 0.393$, i.e., the black cross in Fig. 15. As can be seen, visibility and overexposure are roughly the same in importance, whereas visibility distribution exerts a smaller impact on the comprehensive assessment score. During actual application, this optimal weight coefficient combination was used to calculate the comprehensive assessment scores of different image datasets, and then the dataset with highest score was used to train the models. Once they were applied to the recognition algorithms under different architectures, the datasets could then be compared.

### D. COMPARISON OF COMPREHENSIVE ASSESSMENT SCORE AND EXISTING IMAGE QUALITY ASSESSMENT METHODS

To compare the image quality assessment methods mentioned in Section II and the proposed approach, we used $\alpha = 0.378$, $\beta = 0.229$, and $\gamma = 0.393$ as the weight coefficient settings. We applied the compared methods to the six datasets.
Firstly, we compared our proposed method with full reference assessments including MAE, PSNR, and UIQI. The MAE calculates the absolute error between the target image (predefined high-quality image) and the test image (the image to be tested). A lower value indicates better image quality. Its mathematical expression is as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |x_i - y_i|$$

where $x = \{x_i | i = 1, 2, 3, \ldots, N\}$ and $y = \{y_i | i = 1, 2, 3, \ldots, N\}$ are the pixels of the target and test images. The UIQI used in this study adopts the mean, variance, and covariance of the target and test images. The UIQI is described as follows:

$$\text{UIQI} = \frac{4\sigma_x\sigma_y \bar{x} \bar{y}}{(\sigma_x^2 + \sigma_y^2)(\bar{x}^2 + \bar{y}^2)}$$

where $\bar{x}$ and $\bar{y}$ denote the pixel means of the two images; $\sigma_x$ and $\sigma_y$ are the variances of the two images, and $\sigma_{xy}$ is the covariance between the two images.

Secondly, we compared our proposed method with no reference assessments including NIQE, BRISQUE, and PIQE. NIQE [31] computes the 36 identical natural scene statistics (NSS) features from the patches of the image, fits them with a multivariate Gaussian (MVG) model, and compares its MVG to the natural MVG model. Thus, NIQE model is described as follows:

$$\text{NIQE}(v_1, v_2, \Sigma_1, \Sigma_2) = \sqrt{(v_1 - v_2)^T (\frac{\Sigma_1 + \Sigma_2}{2})^{-1} (v_1 - v_2)}$$

where $v_1$ and $v_2$ are the mean vectors and $\Sigma_1, \Sigma_2$ are the covariance matrices of the natural MVG and the distorted MVG model.

BRISQUE adopts the transformed luminances $\hat{l}(I, J)$ as a mean subtracted contrast normalized coefficient (MSCN). Later, a generalized Gaussian distribution (GGD) is applied to capture the broader spectrum of the distorted image statistics and fit them into the MSCN empirical distribution from the distorted images and undistorted ones. The GGD model is expressed as follows:

$$\text{GGD}(x; \alpha, \sigma^2) = \frac{\alpha}{2\beta \Gamma\left(\frac{1}{2}\right)} \exp\left(-\frac{|x|}{\beta}\right)$$

where

$$\beta = \sigma \sqrt{\frac{\Gamma\left(\frac{1}{\alpha}\right)}{\Gamma\left(\frac{3}{2}\right)}}$$

and $\Gamma$ represents the gamma function. After that, asymmetric generalized Gaussian distribution (AGGD) and fast matching method [46] are applied to estimate parameters for each direction. To calculate the image score, BRISQUE used Support Vector Machine (SVM).

PIQE uses local mean removal and divisive normalization to the input image as the initial steps. It adopts the BRISQUE model [33] to transformed the luminance value by using the MSCN method. PIQE is described as follows:

$$\text{PIQE} = \frac{(\sum_{k=1}^{N_{sa}} D_{sk}) + C_1}{(N_{sa} + C_1)}$$

where $N_{sa}$ is the number of spatially active blocks in a given image, $D_{sk}$ is a constant that does not exceed value of 1, and $C_1$ is a positive constant to prevent numerical instability when the denominator goes to zero.

However, in most industrial applications, a high-quality image cannot be defined beforehand. Thus, to compare the proposed approach with past methods, we chose three defect images that we believed to have excellent quality from the six datasets and defined them as high-quality images to calculate the quantified results of the assessment indices. Figures 16 through 18 display the results of the four assessment methods. The blue, orange, and gray curves represent the calculation results based on the three predefined images with different image quality. These three images were obtained from Dataset 3, Dataset 4, and Dataset 5, respectively. Figure 19 presents the assessment results of this study. As the method used in this study does not need a predefined image to perform assessments, only a single curve was used to display the results.

**FIGURE 16. MAE result.**

The previous experiment results indicate that Dataset 3 and Dataset 4 result in no significant differences in recognition capabilities, and neither do Dataset 2 and Dataset 5. Thus, in the dataset rankings, switching Dataset 3 with Dataset 4 or Dataset 2 with Dataset 5 does not affect the accuracy of the
Table 13 presents the dataset rankings based on the quantified values of the assessment indices from Fig. 16 to Fig. 19. Among the MAE methods, only the first two datasets in the rankings established by MAE-2 were correct; the others were incorrect and even ranked the dataset with the worst quality as the second best. The PSNR methods produced the same results as the MAE methods, so the same issues exist in the results. Among the UIQI methods, the UIQI-1 method produced results closer to the correct results; nevertheless, Dataset 1 and Dataset 6 were still ranked incorrectly. The UIQI-2 and UIQI-3 methods gave higher rankings to the datasets with the poorest quality (Dataset 5 and Dataset 6).

To visualize the performance and correlation of IQA, no-reference method, and F2-score, we present their SRCCs in Table 12. From Spearman and Pearson correlation in Tables 13 and 14, it clearly shows that the proposed method is positively correlated with F2-score of YOLO than full reference and no-reference assessment approaches.

| Table 11: No-Reference performance on dataset 1-6. |
|---------|---------|---------|---------|---------|---------|---------|
|         | D1      | D2      | D3      | D4      | D5      | D6      |
| NIQE    | 8.4     | 9.7     | 11.9    | 10.3    | 9.2     | 7.9     |
| BRISQUE | 35.6    | 37.2    | 40.4    | 40.6    | 42.3    | 44.7    |
| PIQE    | 13.0    | 42.2    | 54.1    | 57.2    | 54.1    | 55.2    |
| BRISQUE Fit | 40.9  | 51.5    | 34.0    | 57.6    | 43.7    | 59.0    |
| NTQE Fit | 0.8     | 0.9     | 0.9     | 0.9     | 0.8     | 9.8     |
| Proposed | 67.6    | 68.6    | 69.6    | 70.7    | 68.2    | 37.0    |

| Table 12: Spearman ranking on no-reference assessments with regard to F2-Score. |
|---------|---------|---------|---------|---------|---------|---------|
|         | D1      | D2      | D3      | D4      | D5      | D6      |
| NIQE    | 2       | 4       | 6       | 5       | 3       | 1       |
| BRISQUE | 1       | 2       | 3       | 4       | 5       | 6       |
| PIQE    | 1       | 2       | 4       | 6       | 3       | 5       |
| Proposed | 2       | 3       | 2       | 1       | 4       | 6       |
| Brisque train | 2   | 4       | 1       | 5       | 3       | 6       |
| NTQE train | 1     | 5       | 3       | 4       | 2       | 6       |
| F2-YOLO  | 5       | 3       | 1       | 2       | 4       | 6       |
| F2-FRCNN | 5       | 4       | 2       | 1       | 3       | 6       |
| F2-SSD   | 5       | 4       | 2       | 1       | 3       | 6       |

| Table 13: Spearman ranking on full-assessments and no-reference assessments. |
|---------|---------|---------|---------|---------|---------|---------|
|         | D1      | D2      | D3      | D4      | D5      | D6      |
| MAE-1   | 4       | 2       | 1       | 3       | 5       | 6       |
| MAE-2   | 6       | 5       | 2       | 1       | 3       | 4       |
| MAE-3   | 6       | 5       | 4       | 3       | 1       | 2       |
| PSNR-1  | 4       | 2       | 1       | 3       | 5       | 6       |
| PSNR-2  | 6       | 5       | 2.5     | 1       | 2.5     | 4       |
| PSNR-3  | 6       | 5       | 4       | 3       | 1       | 2       |
| UIQI-1  | 6       | 4       | 2.5     | 1       | 2.5     | 5       |
| UIQI-2  | 6       | 5       | 3       | 2       | 1       | 4       |
| UIQI-3  | 6       | 5       | 3       | 2       | 1       | 4       |
| NIQE    | 2       | 4       | 6       | 5       | 3       | 1       |
| BRISQUE | 1       | 2       | 3       | 4       | 5       | 6       |
| PIQE    | 1       | 2       | 4       | 6       | 3       | 5       |
| Proposed | 2       | 3       | 2       | 1       | 4       | 6       |
| Brisque train | 2 | 4       | 1       | 5       | 3       | 6       |
| NTQE train | 1     | 5       | 3       | 4       | 2       | 6       |
| F2-YOLO  | 5       | 3       | 1       | 2       | 4       | 6       |
| F2-FRCNN | 5       | 4       | 2       | 1       | 3       | 6       |
| F2-SSD   | 5       | 4       | 2       | 1       | 3       | 6       |
VI. CONCLUSION AND FUTURE WORK
The current trend in automated vision inspection systems is using deep learning models to recognize defects. This is due to the significant improvements in the recognition effects of deep models. As a result, overly long training time is one of the bottlenecks in using deep learning models to develop automated vision inspection systems. To save the time spent on using poor datasets to train deep networks, we proposed a comprehensive assessment score that combines defect visibility, visibility distribution, and overexposure and can be used to assess whether the training image dataset can improve the defect recognition rate of the deep learning model without training with defect image datasets.

Applying the proposed comprehensive assessment score first requires a dataset to train the model. The F2-Score used to predict the dataset serves as the basis for comparison. Then, the light settings are adjusted to obtain a dataset with a higher comprehensive assessment score than that of the benchmark dataset. The new dataset is used to train the model to obtain higher recognition accuracy than the benchmark dataset. Experiment results demonstrated that the comprehensive assessment score can effectively predict the recognition accuracy of YOLO, SSD, and Faster R-CNN models and that choosing the dataset with the highest assessment score to train models could indeed obtain better accuracy than other datasets could. In addition, we adjusted the values of the weight coefficients, observed the resulting dataset quality rankings, collected all of the weight combinations with the right rankings, and used linear regression to obtain the optimal weight coefficients. We found that visibility and overexposure had a greater impact on the comprehensive assessment score.

We compared the proposed comprehensive assessment score with full reference and no-reference image quality assessments, including MAE, PSNR, UIQI, NIQE, BRISQUE, and PIQE. Only our proposed comprehensive assessment score could predict the dataset rankings derived using the F2-score in deep networks with better accuracy. This is because we designed exclusively our assessment approach for defects on metal surfaces. The proposed assessment method is not without limitations and shortcomings. The first is the relationship between the comprehensive assessment score and model recognition accuracy. Based on the experiment results, a higher value in the comprehensive assessment score indicates a relatively higher recognition accuracy in the model, but no absolute relationship that is clear for users exists between the two. It is only known that when the score is above a certain threshold value, a certain degree of recognition accuracy can be achieved.

The second is the applicability of the inspection target. The assessment approach proposed in this study is aimed at scratch and bump defects in grayscale images. In the images, these two types of defects are darker, whereas the background is lighter. If the type of defect that needs to be detected does not have a dark-light relationship with the background, then the proposed approach is not applicable. For instance, if color must be considered in the determination of defects, then the other quantified indices mentioned in Section II, such as color or saturation, will be needed for assessment, rather than the indices used in this study.

ACKNOWLEDGMENT
The authors would like to thank B. Y. Lin for preparing the experimental work.

REFERENCES
[1] F. C. Campbell, Inspection of Metals: Understanding the Basics. Geauga County, Ohio: ASM International, 2013.
[2] B. Ganowska, M. Moloritis, A. Hosovsky, J. Pitel, J. B. Krolczyk, A. Ruggierio, G. M. Krolczyk, and S. Hloch, “Design of the model for the on-line control of the AWJ technology based on neural networks,” Indian J. Eng. Mater. Sci., vol. 23, pp. 279–287, Aug. 2016.
[3] F. Jia, Y. Lei, N. Lu, and X. Xing, “Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization,” Mech. Syst. Signal Process., vol. 110, pp. 349–367, Sep. 2018.
[4] S. Dodge and L. Karam, “Understanding how image quality affects deep neural networks,” in Proc. 8th Int. Conf. Qual. Multimedia Exper. (QoMEX), Jun. 2016, pp. 1–6.
[5] H.-I. Lin and P.-Y. Lin, “An image quality assessment method for surface defect inspection,” in Proc. IEEE Int. Conf. Artif. Intell. Test. (AITest), Aug. 2020, pp. 1–6.
[6] I. Jahr, “Lighting in machine vision,” in Handbook of Machine and Computer Vision: The Guide for Developers and Users. Weinheim, Germany: Wiley, 2017, pp. 63–178.
[7] L. M. Xu, Z. Q. Yang, Z. H. Jiang, and Y. Chen, “Light source optimization for automatic visual inspection of piston surface defects,” Int. J. Adv. Manuf. Technol., vol. 91, nos. 5–8, pp. 2245–2256, Jul. 2017.
[8] M. Senthukumar, V. Palanisamy, and J. Jaya, “Metal surface defect detection using iterative thresholding technique,” in Proc. 2nd Int. Conf. Current Trends Eng. Technol. (ICTET), Jul. 2014, pp. 561–564.
[9] Z. Li, J. Zhang, T. Zhuang, and Q. Wang, “Metal surface defect detection based on MATLAB,” in Proc. IEEE 3rd Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC), Oct. 2018, pp. 2365–2371.
[10] Q. Li and S. Ren, “A real-time visual inspection system for discrete surface defects of rail heads,” IEEE Trans. Instrum. Meas., vol. 61, no. 8, pp. 2189–2199, Aug. 2012.
A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment of camera captured document images,” in Proc. Int. Workshop Camera-Based Document Anal. Recognit. Cham, Switzerland: Springer, 2013, pp. 113–125.

J. Kumar, P. Ye, and D. Doerrmann, “A dataset for quality assessment of blind image quality metrics for IR images for UAV applications,” in Proc. Data Compress. Conf. (DCC), Mar. 2019, p. 578.

J. Masci, U. Meier, D. Ciresan, J. Schmidhuber, and G. Fricout, “Steel surface defect detection and classification with max-pooling convolutional neural networks,” in Proc. Int. Conf. Pattern Recognit. (ICPR), 2012, pp. 3292–3295.

M. Lin, Q. Chen, and S. Yan, “Network in network,” in Proc. Int. Conf. Pattern Recognit. (ICPR), 2012, pp. 3204–3207.

Y. Li and S. Pan, “The quality assessment of fusion effect for remote sensing image,” in Proc. Int. Conf. Remote Sens., Environ. Transp. Eng., Jun. 2011, pp. 2724–2727.

K. Hossain, C. Mantel, and S. Forchhammer, “Evaluation of prediction of quality metrics for IR images for UAV applications,” in Proc. Data Compress. Conf. (DCC), Mar. 2019, p. 578.

Z. Xiao, Y. Leng, L. Geng, and J. Xi, “Defect detection and classification of galvanized stamping parts based on fully convolution neural network,” in Proc. Int. Conf. Pattern Recognit. (ICPR), 2012, pp. 515–521.

K. Choi, K. Koo, and J. Lee, “Development of defect classification algorithm for POSCO rolling strip surface inspection system,” in Proc. SICE-ICASE Int. Joint Conf., 2006, pp. 2499–2502.

S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Eur. Conf. Comput. Vis., Sep. 2015, pp. 339–343.

L. Song, W. Lin, Y.-G. Yang, X. Zhu, Q. Guo, and J. Xi, “Weak micro-scratch detection based on deep convolutional neural network,” IEEE Access, vol. 7, pp. 27547–27554, 2019.

R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, “Comparison of four subjective methods for image quality assessment,” in Computer Graphics Forum, vol. 31, no. 8 Hooboken, NJ, USA: Wiley, 2012, pp. 2478–2491.

A. M. Eskicioglu and P. S. Fisher, “Image quality measures and their performance,” IEEE Trans. Consum. Electron., vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

A. C. Bovik, Synthesis Lectures on Image, Video and Multimedia Processing, Morgan & Claypool, 2006.

A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in Proc. 20th Int. Conf. Pattern Recognit., Aug. 2010, pp. 2366–2369.

K. Gu, G. Zhai, X. Yang, W. Zhang, and M. Liu, “Subjective and objective quality assessment for images with contrast change,” in Proc. IEEE Int. Conf. Image Process., Sep. 2013, pp. 383–387.

Q. Wu, Z. Wang, and H. Li, “A highly efficient method for blind image quality assessment,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2015, pp. 339–343.

W. Xue, L. Zhang, and X. Mou, “Learning without human scores for blind image quality assessment,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog. Pattern Recognit., Jun. 2013, pp. 995–1002.

A. Mittal, R. Soundararajan, and A. Bovik, “Making a ‘completely blind’ image quality analyzer,” IEEE Signal Process. Lett., vol. 20, no. 3, pp. 209–212, Mar. 2013.

N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, “Blind image quality evaluation using perception based features,” in Proc. 21st Nat. Conf. Commun. (NCC), Feb. 2015, pp. 1–6.

A. Mittal, A. K. Moothy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” IEEE Trans. Image Process., vol. 21, no. 12, pp. 4695–4708, Dec. 2012.