IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 March 2023, accepted 14 March 2023, date of publication 27 March 2023, date of current version 13 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3262264

RESEARCH ARTICLE

MeDERT: A Metal Surface Defect Detection Model

CHENGLONG WANG AND HENG XIE
School of Electronic Information and Electrical Engineering, Huizhou University, Huizhou, Guangdong 516007, China
Corresponding author: Chenglong Wang (wangchenglong_hzu@163.com)

This work was supported in part by the Guangdong Special Projects in Key Areas for Colleges under Grant 2021ZDZX1073.

ABSTRACT Defects in various products are unavoidable because of measurement errors and equipment accuracy limitations in the production process. Recent advances in metal surface defect detection have focused on optimizing traditional methods, developing new detection techniques, and exploring deep learning-based algorithms, providing technological support to improve metal manufacturing quality and production efficiency. To ensure the highest yield rate and meet production requirements, all products must undergo defect inspections before leaving the factory. However, Traditional methods for detecting metal surface defects require a lot of manual involvement, are difficult to accurately detect small defects, are susceptible to environmental interference, and lack stability and reliability. To address this issue, we propose the MeDERT model for metal surface defect detection. Our approach involves a new Span-sensitive Texture Fusion (STF) module structure that focuses on multi-headed attention modules to recover lost details and boost inspection speed and on top of that use the Jump-sensitive detail recovery feature fusion module to ensure the validity of the extracted textures. Additionally, we introduce singular value decomposition and pretzel noise to model the noise and enhance model robustness through data augmentation. Our MeDERT model achieved state-of-the-art (SOTA) results on a specified dataset, demonstrating its effectiveness in enhancing inspection efficiency and accuracy.

INDEX TERMS Metal profiles, defect detection, deep learning, salt-and-pepper noise, singular value decomposition, DERT.

I. INTRODUCTION

Metal products made of aluminum profiles are widely used in people's daily lives, and in many regions, the production of aluminum profiles is considered a pillar industry. Along with the advancement of technology, the production of aluminum profiles has been automated, which has dramatically increased efficiency. Automation is often introduced into all stages of the aluminum profile production process, from the very beginning of the material screening, automatic configuration of production operations, detection of defective products, and finally, sorting and packaging, all of which require monitoring of product quality. However, most aluminum profile product quality inspections are currently manual and prone to errors. Therefore, monitoring product quality during the automation of aluminum profiles is crucial, and practical and accurate assessment is also urgent for modern manufacturing. In this study, we investigated the detection of surface defects in aluminum profiles during the production process.

The development of computer vision-based technology has greatly advanced the automated inspection technique. The surface defect detection problem is an object detection problem. The detection task is based on the analysis of images taken by cameras, which aims to identify the location and category of the defects in the images. Therefore, detection precision is the main metric to evaluate the detection method. There have been three types of automated inspection techniques based on production images: 1) image processing [1], [2], [3]. The technique focuses on using image filters to highlight defect instances. 2) machine learning (ML) [4], [5], [6]. Machine learning-based technique uses certain image filters to extract features that are fed into ML algorithms, for example, support vector machine (SVM). 3) deep learning [7], [8], [9]. Deep neural networks (DNN) use convolutional operations for feature learning and extract features during training. It has been widely applied in image classification and object detection tasks.

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

Deep learning methods have advanced object detection tasks. However, surface defect detection for aluminum profiles has its challenges. First, there are various categories of defects, including scratches and dents. Second, even for the same type of defects, the severity of defect instances brings another challenge, like length, size, shape, orientation, and depth. To address these challenges, we need a model to investigate and distinguish the possible patterns in-depth. Meanwhile, data augmentation will be necessary to enrich the dataset and enhance its diversity.

In the research, we propose a defect detection model for aluminum materials to detect surface defects more quickly. This study is valuable economically and socially. Our proposed model is mainly based on the modification of the classical detector YOLOv4, which is adapted to process image features of aluminum profiles. The main contributions of this research are as follows:

1) We propose a new network structure called Span-sensitive Texture Fusion (STF) module. It can better recover confidence in the detail loss during downsampling. It also focuses on jump points and improves the efficiency of detection.
2) We simulate noises to further enhance the robustness of the model by using Singular Value Decomposition and salt-and-pepper noise when performing data augmentation. This addresses the issue that the noise with a certain degree of false detection may disturb STF, which is highly sensitive to jump points.
3) Furthermore, we design an area strategy-based noise addition strategy to further enhance the model, with the consideration that the amount of noise added should be controlled by the size and how much of the target is in the graph.
4) Finally, the ''Aluminum profile surface defects'' dataset [10], ''Magnetic tile'' dataset [11], and ''Northeastern University surface defect database for defect detection task (NEU-DET)'' dataset [12] are compared with the mainstream targets detection models such as Faster RCNN, YOLOv4, YOLOv5, and DERT. The performance of the proposed model is verified. The proposed MeDERT model becomes the SOTA model for the above three datasets.

In recent years, most of the methods mentioned in other similar research literature have the disadvantages of not being able to process more complex images and the long learning time required upfront through deep learning methods [13], [14], [15], [16], while the new method mentioned in this study has the following advantages:

1) Our proposed MeDERT model outperforms recent models such as YOLOv5 and DERT, becoming the state-of-the-art (SOTA) in aluminum profile defect detection, our proposed modifications have effectively improved the defect detection performance.
2) The multi-head attention mechanism allows the model to capture different aspects of input for better representation learning.
3) We propose a new noise robust data enhancement strategy and detail recovery network structure based on DERT which enhances the detection networks' robustness.

II. RELATED WORK

Computer vision has played a significant role in the development of artificial intelligence in the past 20 years. Many applications are related to the development of computer vision, which is noticeable in two ways.

First, The field of computer vision has seen tremendous growth in recent years, largely due to the development of various fundamental datasets such as MNIST, Cifar10, ImageNet [17], and others. These datasets provide a large amount of data that is essential for training machine learning models and improving their performance. It is well known that the quality of the dataset is a critical factor in determining the effectiveness of the model. The abundance of new datasets is driving the rapid development of computer vision technology, leading to breakthroughs in this field. These datasets provide a wealth of information for researchers to study and understand the real world, and they allow machine learning models to be trained more effectively, resulting in improved performance on various tasks. In conclusion, the development of new datasets is crucial for advancing the field of computer vision and ensuring that machine learning models can be trained to achieve higher accuracy and better results. The quality of the dataset is directly related to the performance of the model, making it a critical factor in the development of computer vision technology.

Second, various models have been designed for various computer vision tasks, such as AlexNet, VGGNet [18], and ResNet [19] for classification, as well as UNet [20] and DeepLab [21]series in the semantic segmentation field, which have achieved great performance. Especially for challenging detection tasks, as in this study, classical models like SSD [22], Fast RCNN [23], and Faster RCNN [24] adapted for detection tasks have propelled technical improvement.

However, all the above models rely on Convolutional neural networks (CNNs), which have a significant number of downsampling processes that generate several scales of feature maps during forward processing and represent input data in many dimensions. It needs the generation of forecasts for each scale of information to avoid missing important information in the output of the final layer. Currently, the prevailing solution to this issue is to employ multiscale prediction strategies, such as models: SSD [22] and YOLOv3 [25]. The Mask RCNN model [26] uses another strategy called FPN [27] as its backbone to forecast and fuse information from different scales.

After the older YOLO [28] and YOLOv2 [29], the YOLO series detection networks have progressed tremendously, from the YOLOv3 mentioned above [25] to the succeeding

YOLOv4 [30], YOLOv5 [31], and YOLOX [32], all with considerable increases in efficacy. Proposed in 2015, YOLO [28] has DarkNet as its backbone. It splits an image into an SXS grid and reduces a large amount of computation. However, it demonstrates poor performance in the detection of small objects. YOLOv2 [29] uses batch normalization and introduces anchor boxes to detect multiple objects. YOLOv3 [25] improves the detection performance by incorporating residual blocks and skip connections into DarkNet-53. YOLOv4 [30] adds some more methodologies to the model, including mosaic data augmentation, weighted residual connections, and Mish activation. These modifications made YOLOv4 achieve SOTA on the MS COCO dataset. YOLOv5 uses a different structure as the backbone, which is the Focus structure with CSPdarknet53. YOLOX was proposed in 2021 and adopted a decoupled head to boost performance. It is also an anchor-free model.

Computer vision technology has achieved outstanding performance in some fields [33], [34], [35], [36], [37]. Take medical imaging as an example. It can even outperform manual accuracy to a certain extent, reducing patients suffering from misdiagnosis. Many applications in medical services could reduce human input while ensuring or even boosting accuracy to some extent, for example, monitoring patients' daily conditions [38] or using chest X-ray images to identify health conditions [39] and even diabetes diagnosis [40].

However, another issue with computer vision is that its applications in various areas progress quite differently. Especially in the equally important sector of industrial production, no specific models are created. Therefore, we investigate the problem in this research.

III. METHODS

In this study, we propose a new noise-robust data enhancement strategy and detail recovery network structure based on DERT, enhancing the detection network's robustness. Figure 1 illustrates the proposed method, and Figure 2 shows the network design of MeDERT, which is improved from DERT. The method is described in detail in the following sections.

A. DERT

DERT is based on the improved language model BERT, which is more suitable for target detection tasks. As a next-generation neural network architecture, the Transformer architecture has been proven to perform well in various task environments, and the target detection domain. It is currently difficult for academics to explain how it performs well in multiple-task environments. Still, it is broadly attributed to two aspects: spanning links and multi-headed attention mechanisms.

DERT is improved from Transformer, with a few innovations, and dedicated to targeting detection tasks. It is widely believed that Transformer can achieve SOTA not only in the Natural Language Processing (NLP) field but also in other fields, such as semantic segmentation and speech recognition, with some structural changes to achieve higher robustness.

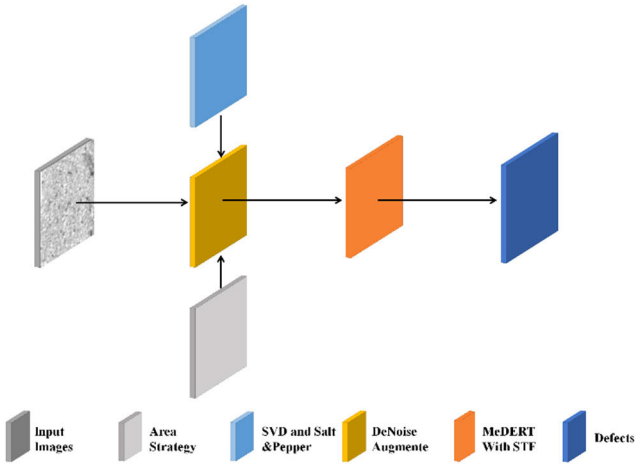


FIGURE 1. The overall workflow of the proposed model.

1) MULTI-HEAD ATTENTION

Multi-Head Attention splits each feature in the feature map into Q(queries), K(keys), and V(values) and achieves a nonlinear addressing process by backpropagation after alternate operations.

The multi-head attention mechanism is a key component in many state-of-the-art NLP models, such as Transformer. It uses three input matrices: Q, K, and V. The goal of the attention mechanism is to compute a weighted sum of values (V) for each query (Q), where the weights are determined based on the similarity between the query and a set of keys (K). The Q matrix is used to compute the attention scores, the K matrix is used for comparison, and the V matrix is used to compute the final attention output.

In the multi-head attention mechanism, multiple attention heads are used to perform multiple attention computations in parallel. This allows the model to capture different aspects of the input, leading to better representation learning. The outputs from each head are concatenated and then projected into a lower-dimensional space to obtain the final attention output. This helps the model to attend to different aspects of the input information, leading to improved performance. This process also manifests Attention to alleviate the spatial complexity of deep learning. Instead of inputting all the features into the subsequent network for computation, some task-relevant information is selected for output. The advantages are as follows:

- *Strengthening the learning ability of a CNN*
  The accuracy of a lightweight CNN could be significantly lowered. Therefore, the learning capability of CNNs needs to be enhanced to achieve good enough accuracy. The structure of CSPNet makes it convenient to be converted to ResNet, ResNeXt, and DenseNet. CSPNet requires 10% to 20% less processing resources when applied to the networks mentioned above while surpassing other classical networks in terms of accuracy.

- *Removing computational bottlenecks*
  When the model structure is too large, inference requires more GPU memory space. Attention aims at the high-quality application of weights such as Q, K, V, etc., thus effectively increasing the utilization of each computational unit and thus reducing wasteful arrangements.
- *Reducing memory costs*
  Manufacturing Dynamic Random-Access Memory (DRAM) wafers are expensive and take up a lot of space. The ASIC (Argus Sour Crude Index) costs can be dramatically reduced when the memory costs can be reduced efficiently. Small-size wafers can also be used in a variety of edge computing devices. CSPNet could reduce memory use on PeleeNet by up to 75% while constructing feature pyramids.

Multi-head Attention is proposed to solve the heavy inference computation of recursive neural networks in computer natural language tasks (e.g., text classification) and establish semantic relations between the preceding and following texts. It is based on the core concepts of nonlinear operations and reuses, replicating the base layer feature map information and reusing a large amount of gradient information. Multi-head Attention maintains the advantage of feature reuse by dividing the feature map of the base layer into three parts but prevents the over-replication of gradient information. One piece will be sent to a dense block and transition layer, while the other part will be combined with the transmitted feature map to the next stage. It successfully allows feature propagation, facilitates the reuse of features by the network, and reduces the number of network parameters. Multi-head Attention is efficiently utilized in the Transformer. In this study, we use DERT as the leading network.

2) DATA AUGMENTATION

For detection tasks, there are a number of methods of data enhancement. The most popular include Mixup, CutOut, and CutMix. Mixup mixes two random samples proportionally, and the classification results are assigned accordingly. CutOut selects a random part of the original image, replaces it with black, and eliminates the original information. cutMix also operates on a pair of images and integrates features from the two previous images. It first generates a random crop frame and crops out the corresponding position in image A. It then uses the ROI (region of interest) of the corresponding position in image B and puts it in the cropped region from image A to form a new sample.

With standard training, small objects are detected considerably less effectively than medium and large objects.

Surrounding features that could be used to recover information are further losses due to the Patch mechanism in the Transformer. The difference in scale and proportion between training and reality also causes some degree of metric degradation.

To address this issue, we modified the enhancement method used by DERT and proposed a new approach to improve detection.

B. JUMP-SENSITIVE DETAIL RECOVERY FEATURE FUSION MODULE

We need to up-sample twice the size of the feature map due to pooling when fusing information from feature maps of different scales, which ensures the correspondence between pixels and a stable shape of the composed tensor during computation. Meanwhile, interpolation is often used to recover the information on missing points. However, the interpolation method (linear or bilinear) assumes a smooth variation of values between different pixels. A relatively obvious jump in the original data will certainly lead to significant differences between the recovered high-resolution map and the real input. The apparent outliers in the image are mainly caused by two situations. One could be the noise during the image acquisition process. During the detection process, there are multiple convolution and pooling operations that the feature maps have undergone, but it is not guaranteed to eliminate all single noise points. However, there is a great chance to improve the situation. Another kind of outlier corresponds to the predominance of a significant change in pixel value or some feature in the image, which is defined as an edge in computer vision. To distinguish foreground and background, edges provide critical information, which greatly impacts the effectiveness of the detection task.

To address this problem, we try to retain the outlier lost during each downsampling to improve the detection effect. In our work, we consider the largest or smallest pixel value in each sliding window as the outlier while not considering other statistical metrics, such as the mean value. In order to preserve as much edge information as possible when upsampling, the input feature map is max-pooled and min-pooled by taking negative numbers to obtain both the maximum and minimum values in the current range. The two obtained feature maps, called F_max and F_min, are the same size as the input image.

For information fusion between different feature maps, we use the concatenation operation instead of the addition operation to alleviate the computational burden. Since different locations in the feature map have different information, the weight parameter in the above concatenation operation has to change corresponding to coordinates and be adaptively adjusted according to the scale between different inputs. In order to fully understand the connection, we design a small network for calculating weights.

We design the model Span-sensitive Texture Fusion (STF) module as follows: Firstly, the $2\times$ upsample feature map, F_min and F_max are concatenated in the channel dimension, and then two 1×1 convolutions are performed to fuse the information from different sources and normalize the weights to obtain the three weights in each coordinate. Figure 3 explains the operation flow in the STF module with $16\times$ downsampled data and $32\times$ downsampled data for fusion.

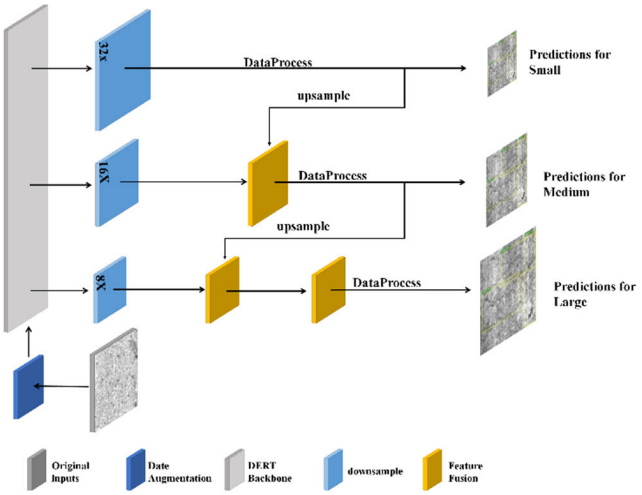


FIGURE 2. MeDERT: an improvement from DERT. Yellow modules represent the improvements we have made.

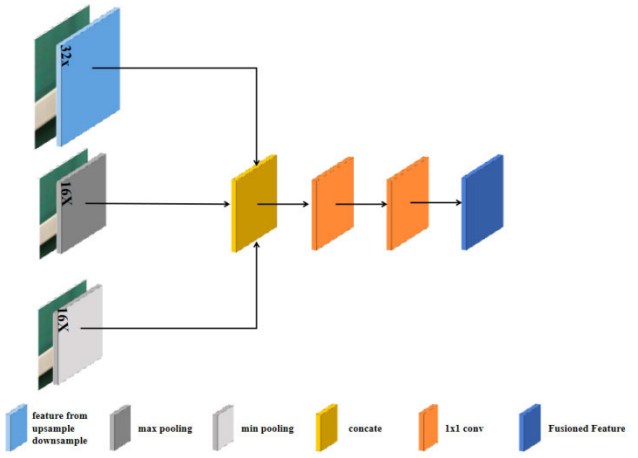


FIGURE 3. STF operation for 16× and 32× downsample feature map.

As illustrated in Figure 2, the prediction head is achieved by combining the above feature fusion structure with the original multiscale prediction head. It is more sensitive to jump points strongly correlated with the possible objects, which can better recover details.

The advantage of the STF fusion method is that it fuses several feature maps of different sizes, giving it the characteristics of pyramidal feature extraction, but with a simpler structure than a pyramid. And this type of feature fusion is more suitable for Transformer architecture.

C. DENOISE DATA AUGMENTATION METHOD

Based on the above strategy, the model is necessarily driven to focus too much on the jump points in the graph, which can cause some problems:

1) The model emphasizes too much the regions with drastic changes and ignores most of the pixels with normal changes, resulting in a certain degree of missed detection.
2) In the image acquisition process, there will be unavoidable noise due to changes in viewpoint, illumination, and shooting tools. The noise has very similar characteristics to the jump location, which becomes an interference for the model to identify and greatly reduces the robustness of the model.
3) It is difficult to define a fixed pattern for the jump points, which can easily lead to missed detection and false alarm.

Therefore, the model is strongly dependent on the data, which needs to be decoupled. We use data augmentation to solve the above problem.

Singular Value Decomposition is used in our design to interfere with high-resolution photos and create low-resolution images for training. It dynamically strips a portion of the information in the graph and reflects it in specific pixel values, which is similar to the effect of noise overlay. We keep the first 200 principal components in each image.

To further simulate noise during data augmentation and fully enhance the robustness of the model, we use salt-and-pepper noise to perform augmentation and add only the following noise in 0.5% of the pixels to not disturb the original image too much.

The data augmentation approach described above provides good simulation capability for noise and enhances the robustness of the model. To a certain extent, it can eliminate the interference of unpredictable noise in the model. Therefore, we call it DeNoise Data Augmentation Method.

There are two strategies for salt-and-pepper noise incorporation. One is to use a fixed proportion. The other uses the proportion of the whole area occupied by the actual target box as the basis for noise incorporation. According to experimental results, the latter achieves a better effect.

We adopt the following strategy for the input image x. The proportion of added salt-and-pepper noise is set to P. Formula 1 explains how the proportion is determined.

$$P(x) = 0.2 \times \sigma(\frac{1}{2}\gamma\ (area_u + area_n + \gamma)) \tag{1}$$

A random perturbation γ was introduced into the calculation as a way of improving the accuracy and stability of the calculation. γ is added to the calculation by randomly generated values which can make the results more accurate and reliable.

$area_u$ is the sum of the areas of all objects, and it is an important indicator for assessing the size of an object. However, in this step, the overlap of different objects corresponding to the bounding boxes is not considered.

$area_n$ is calculated as the sum of the areas of all bounding boxes, which is an important indicator for assessing the number and distribution of objects. These bounding boxes represent the shape of the objects and can help us to better understand the position and size of the objects.

σ is a Sigmoid function, which is a commonly used nonlinear function widely used in fields such as machine learning and neural networks. It helps to classify and analyze the

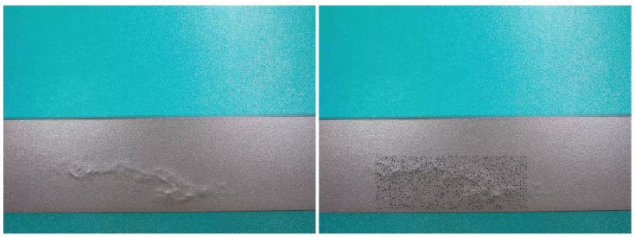


FIGURE 4. Original images and artificial noise.

TABLE 1. Units for magnetic properties.

Defect	Num
Defect0	128
Defect1	365
Defect2	538
Defect3	390
Defect4	173
Defect5	86
Defect6	82
Defect7	407
Defect8	261
Defect9	346
Multi Defects	229

results by transforming the input data so that the results are between 0 and 1.

Finally, the upper limit of the perturbation is set to 0.2. This means that the range of the perturbation is limited, avoiding excessive fluctuations in the calculated results. This also helps to ensure the reliability and repeatability of the calculated results.

Overall, these concepts and functions are very important and are crucial for improving the accuracy and efficiency of the calculations.

As illustrated in Figure 4, it should be noted that we tried to add artificial noise around the object detection box to affect the accuracy of the detection. However, the experimental results showed very serious overfitting. It is concluded that the model could easily recognize this noise around the detection box as a feature indicating the presence of the detection box, causing unnecessary dependence.

IV. EXPERIMENT

A. DENOISE

In order to fully study the characteristics and detection effectiveness of defects in aluminum profiles, a dataset consisting of samples from the actual production process was obtained by collecting pictures of aluminum profiles, which were obtained at a certain period from a certain production line of an enterprise in Nanhai, Foshan, Guangdong Province, China. There are 4356 samples in the dataset, and 3005 of them contain defects for analysis. Samples without defects are not considered.

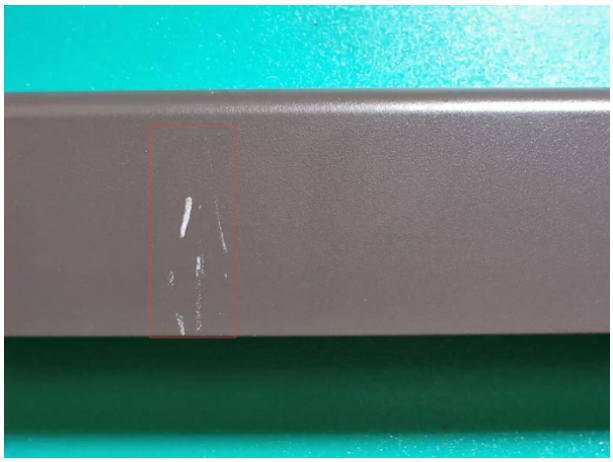


FIGURE 5. Sample of surface defect1.

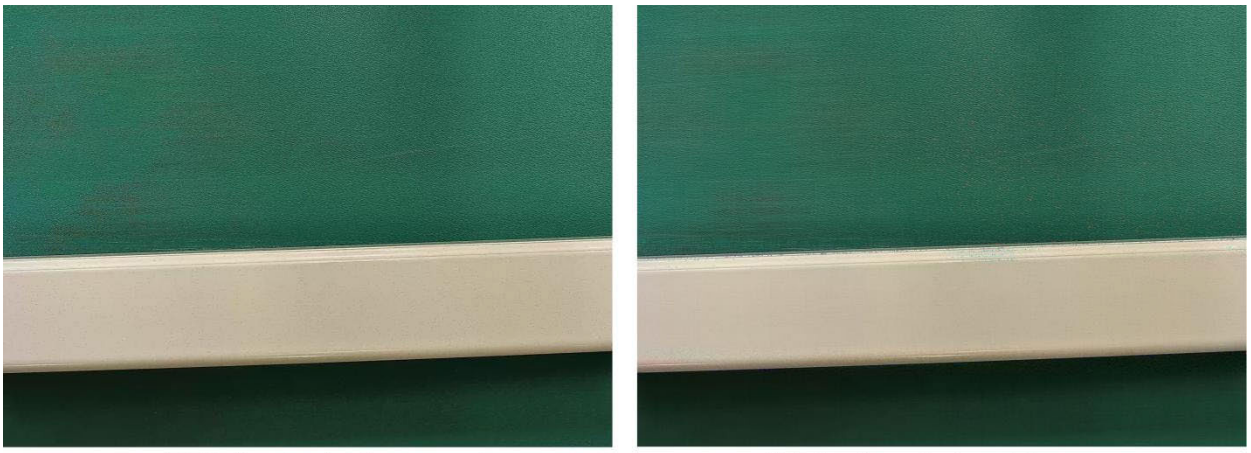
FIGURE 6. Sample of surface defect4.

According to the categories of defects that may occur in the actual production process, a total of 10 different types of defects are included. The dataset is divided into the training set and the data set, with numbers 2505 and 500. Table 1 shows the statistics of the ten defects.

After visualizing a number of samples (as shown in Figures 5 and 6), we find that it is difficult to detect surface defects in aluminum profiles due to small differences between defects and background. Meanwhile, there are lots of variations of defects, which means a large number of defect types and their similarities make the detection tasks more challenging. This is also why our research is meaningful and economically beneficial. The results of this research could reduce production costs and accelerate automation innovation in the industry.

B. MULTIPART FIGURES

Figure 7 shows the effect of data augmentation, comparing before and after augmentation. Through Singular Value Decomposition and salt-and-pepper noise, many jump points are recognized in the picture. By doing so, the noise is artificially added to the picture, which drives the model to improve robustness and enhance its anti-interference ability.



a) Original Input Sample: before Augmentation

b) The Same Image: after Augmentation

FIGURE 7. Augmentation Effect: the Same Image before and after Augmentation.

TABLE 2. Numbers of defects in aluminum profile surface defects dataset.

Model	STF	DeNoise	Fixed Proportion	Area Strategy	AP	AP^{50}	AP^{75}
Faster RCNN(Baseline)	×	×	×	×	0.775406	0.769865	0.776022
YOLOv5(Baseline)	×	×	×	×	0.831565	0.833692	0.839798
DAFNe	×	×	×	×	0.82946	0.834221	0.845845
RTMDet-R-l	×	×	×	×	0.837481	0.845962	0.854316
DERT	×	×	×	×	0.841891	0.846668	0.849194
DERT + STF	√	×	×	×	0.853289	0.855176	0.855519
DERT + STF + DeNoise	√	√	×	×	0.874183	0.876221	0.875936
MeDERT(DERT + STF+ DeNoise + Area Strategy)	√	√	×	√	0.887823	0.885094	0.887615

TABLE 3. Numbers of different type of defects in magnetic tile dataset.

Model	STF	DeNoise	Fixed Proportion	Area Strategy	mIOU	DICE	mPA
Faster RCNN(Baseline)	×	×	×	×	0.9485	0.1012	0.9326
YOLOv5(Baseline)	×	×	×	×	0.9673	0.077	0.9629
SegFormer	×	×	×	×	0.9765	0.0801	0.9724
Mask2Former	×	×	×	×	0.9688	0.0772	0.9775
SETR	×	×	×	×	0.9661	0.0895	0.9674
SETR + STF	√	×	×	×	0.9793	0.0761	0.9789
SETR + STF + DeNoise	√	√	×	×	0.9821	0.0613	0.9761
MeDERT(SETR + STF+ DeNoise + Area Strategy)	√	√	×	√	0.9894	0.0576	0.9853

C. TRAINING CONFIG

1) HARDWARE CONFIGS

In the surface defect detection experiment, we use Python and PyTorch with version 1.8.0. To accelerate the training, we use GPU RTX2080Ti.

2) TRAINING CONFIGS

Here we choose Adam [41] as the optimizer in our experiments. The cross-entropy loss function was used for the optimization objective. We set the learning rate to 0.0005 and $\beta 1$ and $\beta 2$ to 0.99 and 0.999, respectively. to make the

TABLE 4. Numbers of different type of defects in NEU-CLS dataset.

Model	STF	DeNoise	Fixed Proportion	Area Strategy	AP	AP^{50}	AP^{75}
Faster RCNN(Baseline)	×	×	×	×	0.787376	0.784331	0.775427
YOLOv5(Baseline)	×	×	×	×	0.842882	0.843963	0.865238
DAFNe	×	×	×	×	0.856302	0.859916	0.844364
RTMDet-R-l	×	×	×	×	0.849888	0.826564	0.835787
DERT	×	×	×	×	0.850016	0.844495	0.84026
DERT + STF	√	×	×	×	0.854088	0.835317	0.84581
DERT + STF + DeNoise	√	√	×	×	0.860279	0.869036	0.867398
MeDERT(DERT + STF+ DeNoise + Area Strategy)	√	√	×	√	0.861902	0.885304	0.879113

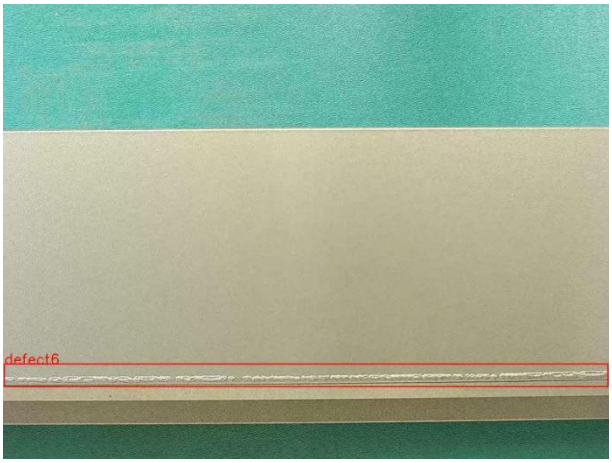


FIGURE 8. Defect detection result: sample1.

FIGURE 9. Defect detection result: sample2.

FIGURE 10. Defect detection result: sample3.

calculation efficient, we set the eps parameter to 1e-08 to prevent the denominator from being 0. The attention head branches are 7 and 20,000 Epochs are used to train our model with a learning rate of 1e-5 using the pre-training method of the COCO model.

D. METRICS

In our experiment, we choose to mean Average Precision(mAP) to measure the effectiveness of detection. First, we arrange all detection results predicted by our proposed model in decreasing order, according to the confidence score. When our prediction and gt get an IoU (Intersection over Union) greater than a certain threshold we predefine, we call them a "match." We use statistics to get the PR (Precision-Recall) curve and further the area under the curve, which is the AP score. Finally, by fusing different AP values, we get mAP.

We use AP, AP^{50}, and AP^{75}, which correspond to different thresholds, to evaluate the effect of the model from different perspectives. They can provide a better comparison between models. When calculating AP^{50} and AP^{75}, the threshold is used to determine whether the output is positive or negative.

FIGURE 11. Defect detection result: sample4.

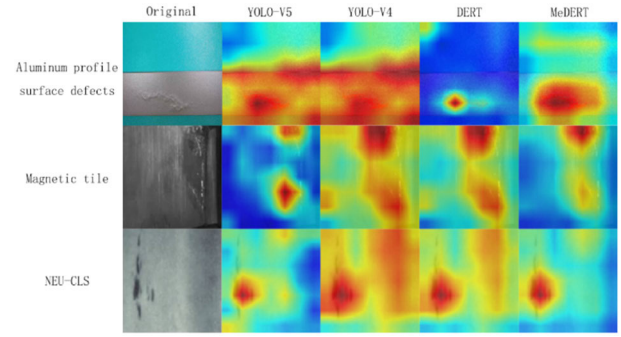


FIGURE 12. Attention map comparison.

Taking the threshold of 50 as an example, when the IoU of the output and the real box is greater than 0.5, it is considered positive. Similarly, 75 corresponds to 0.75. For calculating AP, in the interval [0.5, 0.95], take a value every 0.05 as the threshold and finally average the 10 APs, which is the final AP.

E. RESULT

We chose Faster RCNN and YOLOv4 as the baseline, which are well-known detectors with superior detection performance. Our model is based on modifications of the YOLOv4 backbone, and we examine its impact on the detection task.

We first change the multi-scale feature fusion strategy and then replace the original fusion block with our proposed Span-sensitive Texture Fusion (STF) module. As shown in Table 2-4, our model achieves significant improvement in all AP metrics and beats all other detection models, which fully illustrates the efficiency of our new block. This new structure can effectively recover the details of the original input during upsampling and better capture the jump points to improve object detection. We further add the DeNoise Augmentation strategy that improves all AP metrics by 2%. This fully illustrates that the data augmentation approach in our model effectively alleviates the sensitivity to noise resulting from the STF module and its excessive attention to jump points. It enhances the robustness of the model and allows it to cope with different types of inputs better. At last, the addition of an area strategy further improves the performance and achieves the best results.

All experimental results are shown in Table 2-4, with the maximum values in each column marked in bold. Our final model achieves the best metrics in all aspects. Even compared with the latest models, such as DERT or YOLOv5, it demonstrates an obvious advantage.

Figures 8, 9, 10, and 11 show examples of the results. It can be seen that our proposed model achieves better results for different hues, different categories, and different numbers of defects. It provides accurate detection and demonstrates robustness even in cases where there are different defects.

V. CONCLUSION

In this paper, we propose MeDERT, a new defect detection model for aluminum profiles that aims to improve the performance of the DERT model. To achieve this, we incorporate several modifications to the DERT model and introduce a new structure called the Span-sensitive Texture Fusion (STF) module. This new module enhances the defect detection performance by focusing more attention on jump points and better recovering confidence in detail loss.

In addition to the STF module, we also introduce Singular Value Decomposition and salt-and-pepper noise to enhance the robustness of the model through data augmentation. These techniques are designed to simulate real-world noise and balance the high sensitivity to jump points introduced by the STF module. To further improve the robustness of the model, we design an area strategy-based noise addition strategy that adds noise to the input data.

Our experimental results show that MeDERT outperforms recent models such as YOLOv5 and DERT and becomes the state-of-the-art (SOTA) model in the field of aluminum profile defect detection. The results demonstrate the advantages of MeDERT over other models, indicating that our proposed modifications have effectively improved the defect detection performance. The results of this study contribute to the advancement of the field and provide valuable insights into the development of future models for aluminum profile defect detection.

ACKNOWLEDGMENT

Conceptualization and methodology, Chenglong Wang and Heng Xie; software, validation, and original draft preparation, Heng Xie; review and editing, Chenglong Wang. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] M. Senthikumar, V. Palanisamy, and J. Jaya, "Metal surface defect detection using iterative thresholding technique," in *Proc. 2nd Int. Conf. Current Trends Eng. Technol. (ICCTET)*, Jul. 2014, pp. 561–564.

[2] 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, 2017.

[3] 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.

[4] Z. Xue-wu, D. Yan-qiong, L. Yan-yun, S. Ai-ye, and L. Rui-yu, "A vision inspection system for the surface defects of strongly reflected metal based on multi-class SVM," *Exp. Syst. Appl.*, vol. 38, no. 5, pp. 5930–5939, May 2011.
[5] F. Riaz, K. Kamal, T. Zafar, and R. Qayyum, "An inspection approach for casting defects detection using image segmentation," in *Proc. Int. Conf. Mech., Syst. Control Eng. (ICMSC)*, May 2017, pp. 101–105.
[6] R. Shanmugamani, M. Sadique, and B. Ramamoorthy, "Detection and classification of surface defects of gun barrels using computer vision and machine learning," *Measurement*, vol. 60, pp. 222–230, Jan. 2015.
[7] Y.-J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyüköztürk, "Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 9, pp. 731–747, Nov. 2018.
[8] Z. Xiao, Y. Leng, L. Geng, and J. Xi, "Defect detection and classification of galvanized stamping parts based on fully convolution neural network," *Proc. SPIE*, vol. 10615, Apr. 2018, Art. no. 106150K.
[9] 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.
[10] *Aluminum Profile Surface Defects Dataset*. Accessed: Nov. 6, 2018. [Online]. Available: https://tianchi.aliyun.com/competition/entrance/231682/information
[11] L. Jun, Y. U. Jiajia, and Y. Zhang, "Surface defect detection of magnetic tile based on contourlet transform," *Light Ind. Mach.*, 2013.
[12] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1493–1504, Apr. 2020.
[13] Y. Wang, H. Liu, and L. Zhang, "A new approach for detecting anomalies in industrial equipment based on machine learning," *Robot. Comput.-Integr. Manuf.*, vol. 79, Jun. 2022, Art. no. 102470.
[14] Y. Wu, J. Wang, Y. Zhang, X. Li, and Y. Li, "Fabrication and characterization of hydroxyapatite/chitosan composite scaffolds for bone tissue engineering," *Materials*, vol. 15, no. 6, p. 2205, 2022.
[15] X. Li, W. Zhou, M. Liu, and Q. Wu, "Intelligent fault diagnosis of power transformer using transfer learning and one-class support vector machine," *IEEE Access*, vol. 9, pp. 117819–117830, 2021.
[16] Y. Zhang, X. Guo, J. Zhang, and Z. Wang, "A new hybrid method for time series forecasting based on extreme learning machine and genetic algorithm," *Math. Problems Eng.*, vol. 2021, Feb. 2021, Art. no. 5592878.
[17] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2014, p. 1556.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, p. 10, doi: 10.1109/CVPR.2016.90.
[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[21] L. C. Chen, G. Papandreou, and I. Kokkinos, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Jun. 2017, doi: 10.1109/tpami.2017.2699184.
[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 21–37.
[23] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.
[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[25] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[26] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.
[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.
[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.
[29] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
[30] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[31] G. Jocher. (2021). *YOLOv5*. [Online]. Available: https://github.com/ultralytics/yolov5
[32] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
[33] X. Zhang, W. Liang, Q. Wang, J. Li, and X. Sun, "Deep learning for pulmonary nodule detection: A survey," *Computerized Med. Imag. Graph.*, vol. 84, Oct. 2020, Art. no. 101765.
[34] N. Ghatwary, M. M. Gaber, and A. Saeed, "Deep learning in medical image analysis: A review," *Comput. Methods Programs Biomed.*, vol. 200, Jan. 2021, Art. no. 105880.
[35] A. Bn et al., "COVID-19: Automatic detection from X-ray images by utilizing deep learning methods," *Expert Syst. Appl.*, vol. 176, 2021, doi: 10.1016/j.eswa.2021.114883.
[36] C. Hao, W. Chen, W. Zhu, and L. Yu, "Multi-task deep learning model for joint optic disc and fovea detection in color fundus images," *Computerized Med. Imag. Graph.*, vol. 103, Nov. 2022, Art. no. 101988, doi: 10.1016/j.compmedimag.2021.101988.
[37] X. Li, Y. Huang, L. Liu, X. Zeng, and G. Yang, "Few-shot learning with differentiable classification and attention for ultrasound image classification," *Med. Image Anal.*, vol. 75, Dec. 2022, Art. no. 102234, doi: 10.1016/j.media.2021.102234.
[38] J. A. Quinn, R. Nakasi, P. K. B. Mugagga, P. Byanyima, W. Lubega, and A. Andama, "Deep convolutional neural networks for microscopy-based point of care diagnostics," in *Proc. Mach. Learn. Healthcare Conf. (MLHC)*, Los Angeles, CA, USA, 2016, pp. 271–281.
[39] N. Lessmann, I. Isgum, A. A. Setio, B. D. D. Vos, F. Ciompi, P. A. D. Jong, M. Oudkerk, P. T. M. Willem, M. A. Viergever, and B. V. Ginneken, "Deep convolutional neural networks for automatic coronary calcium scoring in a screening study with low-dose chest CT," in *Proc. SPIE*, 2016, Art. no. 978511, doi: 10.1117/12.2216978.
[40] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Proc. Comput. Sci.*, vol. 90, pp. 200–205, Jan. 2016.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

CHENGLONG WANG received the Ph.D. degree from the School of Engineering, Huazhong Agricultural University, in 2014. He joined Huizhou University as a Teacher. His research interests include machine vision and artificial intelligence.

HENG XIE received the Ph.D. degree in material processing engineering from the South China University of Technology, Guangzhou, China, in 2015. Her research interests include intelligent manufacturing and electrical materials.