Research Article | Vol. 33, No. 6 / 24 Mar 2025 / Optics Express 13848
Optics EXPRESS

Enhanced transformer method coupled with transfer learning for surface defect segmentation of myopia control spectacle lenses

RUOXIN WANG,[1,2] CHI FAI CHEUNG,[1,3,4,*] BO WANG,[1,3] ZHANCHEN ZHU,[1] AND DENNIS YAN-YIN TSE[4,5]

[1]State Key Laboratory of Ultra-Precision Machining Technology, Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China
[2]Center for Industrial Artificial Intelligence, Department of Mechanical Engineering, A. James Clark School of Engineering, University of Maryland, College Park, Maryland, USA
[3]Research Centre for SHARP Vision, The Hong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China
[4]The Hong Kong Polytechnic University-Wenzhou Technology Innovation and Research Institute, Wenzhou, China
[5]School of Optometry, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China
*benny.cheung@polyu.edu.hk

Abstract: Myopia requires visual correction. The complications associated with myopia affect a large population of schoolchildren around the world. Nanostructured myopia control spectacle lenses (NMCSLs) containing nano surface features are commonly used as a non-invasive approach for slowing down the progression of myopia. However, the effective segmentation of surface defects generated in the precision manufacturing of the NMCSL heavily relies on highly efficient and effective defect detection and characterization methods. As a result, this paper presents an enhanced transformer method coupled with the transfer learning (E2Trans) method, which combines the powerful feature extraction abilities of the transformer and the knowledge re-usage abilities of transfer learning to realize high-efficiency and high-accuracy defect segmentation. To further improve the segmentation performance, two auxiliary decoders are added to adjust the training loss. To validate the model's performance, a lens defect dataset is built, and a series of experiments are conducted. The results show that our proposed model can segment five lens defects, including notches, black spots, bubbles, fibers, and scratches with high segmentation accuracy and speed. In addition, a detection system is developed for real-time lens defect detection.

1. Introduction

1.1. Myopia control

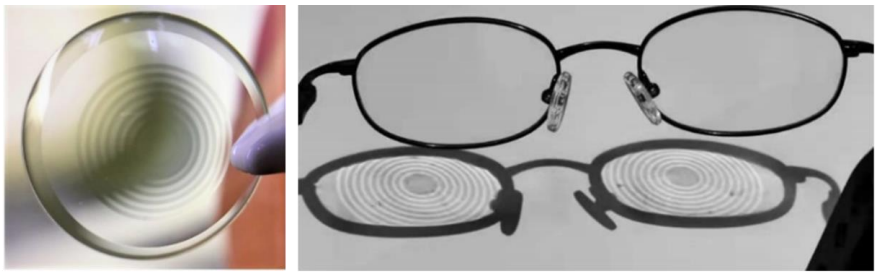
Holden et al. [1] reported that approximately 1.4 billion people had myopia in 2000, and projected that the number of myopic people will increase to 4.8 billion by 2050. Among them, about 0.9 billion people will have high myopia. The increase in myopia among school children has become a major public health issue globally. This could result from the significant amount of time children spend reading and using computers or smartphones while spending inadequate time outdoors [2,3]. The socio-economic consequence of myopia is enormous because myopia permanently alters the structure of the posterior segment of the eye, increasing the risk of developing blinding diseases such as myopic macular degeneration and glaucoma [4]. In particular, myopic macular

#558277 https://doi.org/10.1364/OE.558277
Journal © 2025 Received 30 Jan 2025; revised 9 Mar 2025; accepted 9 Mar 2025; published 17 Mar 2025

Optics EXPRESS

degeneration is the second leading cause of low vision in Hong Kong [5], the fourth in Singapore [6] and the fifth in the US [7].

Vision health is closely related to the quality of people's daily lives, and the formation of myopia is irreversible. During childhood and adolescence, the eyes are in a sensitive period of rapid development. It is particularly effective and critical to actively prevent and intervene in the formation and progression of myopia at this stage. Biologists found that excessive ocular growth can be controlled through a physiological feedback mechanism termed "emmetropization" by imposing a defocused optical image onto the retina [8]. The defocused image, often referred to as "Myopic Defocus", is the underlying principle for conventional myopia interventions such as orthokeratology, which involves the overnight application of a rigid gas permeable contact lens; or the day-wear defocus incorporated soft contact lens (DISC), as shown in Fig. 1(a), developed by Lam [9]. However, contact lenses are often contra-indicated for many school children due to hygiene, safety concerns, and their limited capacity to correct astigmatism. Consequently, spectacle lenses that incorporate a myopia control function would be a user-friendly option appealing to all patients. Accordingly, nanostructured myopia control spectacle lenses (NMCSLs), as shown in Fig. 1(b), have been developed by the authors of The Hong Kong Polytechnic University. They proposed a novel optical power distribution design and ultra-precision mold-processing method to realize the mass production of NMCSLs. However, controlling the quality of NMCSLs in the manufacturing process is difficult, resulting in high scrap rates and low efficiency. Therefore, a fast and accurate defect detection technique is urgently required in NMCSL manufacturing.



(a) Defocus incorporated soft contact lens (b)Myopia control spectacle lenses

Fig. 1. Myopia control lenses

1.2. Defect detection in injection molding

Injection molding is the most popular technique to manufacture lenses, especially lenses with micro or nanostructures. To improve the quality of products and enhance the efficiency of processes, there are many studies focused on detecting and analyzing defects. Obregon et al. [10] proposed a rule-based explanation (RBE) framework to detect sink mark defects in the injection molding process. They predicted defects through ensemble methods and then RBE was used to combine the decision trees and extract the rules from them. Some plot studies have depicted the relationships between features and predictive models. Bianchi et al. [11] developed a thermal control system to avoid premature solidification, in which a particle swarm optimization-based finite element method was used to make the cooling rate of molded components uniform. The results showed that time delay neural networks have better detection accuracy. Chen et al. [12] developed an artificial neural network (ANN) to conduct online defect detection. They trained the ANN using the real-time temperature and pressure data to predict the part diameter. The proposed method showed better performance compared with

Optics EXPRESS

the multilinear linear regression model. Anguraj [13] adopted a support vector machine to classify the qualified and unqualified injection-molded components online based on the data from temperature and pressure sensors. The Bayesian regularization approach was used to make the classification more refined. Khosravani et al. [14] applied a case-based reasoning method to conduct fault detection in the injection molding process. The occurrence weight of faults' causes was used to define similarity measurement and case retrieval. The experiments showed that the proposed method can reduce the machine breakdown time. Although many studies have been conducted, little attention is paid to classifying and recognizing defects on inject-molded workpieces simultaneously, which plays an important role in understanding the inject-molding process and performing further process optimization.

As a result, we aim to improve the quality and efficiency of NMCSL production. We propose a deep learning-based segmentation method, and an enhanced transformer method coupled with transfer learning (E2Trans) to detect surface defects on NMCSLs. In the model, two auxiliary decoders are added to decode the learned feature into different segmentation masks. Then the auxiliary loss is added to the loss function to adjust the feature learning so as to improve segmentation performance. In addition, to accelerate the training speed and reduce the data collection cost, a transfer learning technique is used to transfer the knowledge from a public dataset to NMCSL detection. Then the dataset is collected for model performance validation. The main contributions of this work can be summarized as follows.

a) An enhanced transformer method coupled with transfer learning is proposed.

b) Five types of defects can be segmented by the proposed method.

c) The method yields better segmentation performance compared with other models.

d) A detection system is developed for real-time lens defect detection.

2. Related work

2.1. Deep learning-based segmentation methods

Deep learning-based segmentation methods have been used in many industries, such as medicine [15], agriculture [16], electronics [17], and manufacturing [18,19] since they push the detection problem down to the pixel level. There are many types of deep learning methods proposed in recent years [20], which can be roughly divided into convolutional neural networks (CNN)-based models, recurrent neural network (RNN)-based models, generative adversarial network (GAN)-based models and transformer-based models.

CNN-based models. Liu et al. [21] proposed a deep learning-based method, ParseNet, in which global context is added through the average feature based on a fully convolutional network. The experimental result on SiftFlow and Pascal-Context showed that the segmentation accuracy increased consistently. Lin et al. [22] combined CNNs with conditional random fields (CRFs) to improve the segmentation accuracy, where a CRF was used to capture the contextual information from the neighboring patches of images. The piecewise training technique was adopted to achieve efficient learning. Paszke et al. [23] proposed an efficient neural network (ENet) to enable segmentation to be used in mobile applications. They designed an initial block to obtain 16 feature maps. An ENet bottleneck module was carefully designed to learn the features. The performance on benchmark datasets showed that ENet can yield similar accuracy but 18× faster.

RNN-based models. Shuai et al. [24] developed a directed acyclic graph-RNN to capture the contextual dependencies of image patches, which can aggregate the context of neighboring feature maps. In addition, a class-weighted loss was added in the training process to deal with the class occurrence imbalance issue. Liang et al. [25] proposed a graph long short-term memory network (LSTM), in which an undirected graph was established for each image with the arbitrary-shaped

Optics EXPRESS

superpixels as nodes. In addition, a confidence-driven scheme was introduced to update the states of the LSTM. Xiang et al. [26] proposed data-associated RNNs (DA-RNN) for RGB-D video semantic labeling. A new block diagram, a data-associated recurrent unit was proposed to update the hidden state. The outputs of DA-RNN were combined with a mapping technique, KinectFusion, to reconstruct 3D scenes.

GAN-based models. Luc et al. [27] combined CNN with GAN models to improve segmentation accuracy. They used the CNN to perform per-pixel class predictions, and the GAN was used to distinguish whether the segmentation map comes from the CNN or ground truth. Souly et al. [28] proposed a semi-supervised segmentation method based on the GAN to mitigate the shortage of labeled data. They used the generator to generate fake images. Hence, the fake data, unlabeled data, and labeled data were fed into a discriminator which can obtain confidence maps of each class and classify fake or real data simultaneously. Xue et al. [29] proposed an end-to-end trainable network, SegAN, for medical segmentation. The fully convolutional encoder-decoder network served as the segmentor to produce label maps. And the predicted label maps and the ground truth label maps were fed into a critic network. These two networks were trained alternately to maximize the multi-scale loss.

Transformer-based models. Transformers have become more and more dominant in vision tasks since they can provide simple, unified, and robust solutions [30]. There are many studies proposed recently. Liu et al. [31] developed a new vision transformer, which uses shifted windows to compute representations in a hierarchical way. The proposed transformer, Swin Transformer, is suitable for not only any image scales, but also many vision tasks, such as classification, detection, and semantic segmentation. Xie et al. [32] proposed SegFormer for semantic segmentation, which combines transformers and a lightweight multilayer perceptron (MLP) as an encoder-decoder network. Transformers were hierarchically concatenated to avoid positional coding, and a simple MLP was used to combine the representation from different layers so that it is powerful and efficient. Cheng et al. [33] proposed a masked-attention mask transformer (Mask2Former), which can be applied to any segmentation task, including panoptic segmentation, instance segmentation, and semantic segmentation. The designed mask-attention can limit the cross-attention in a predicted mask region to learn local features. Similarly, Jain et al. [34] tried to deal with three kinds of segmentation tasks in one framework. They designed a novel training strategy to jointly train a network for each segmentation task. In addition, a task token was introduced to enable the model to dynamically perform various tasks' training and prediction.

2.2. Deep learning-based segmentation methods

Defect segmentation is a challenging issue in many industries, such as material, civil infrastructures, textile, agriculture, and manufacturing. Neven et al. [35] introduced a multi-task model based on U-Net to segment surface defects on sheet steel and predict severity. In addition, they improved the model's performance by adding process parameters and a sensor fusion technique, whereby the segmentation accuracy increased by 6.8% in mIoU. Wang et al. [36] segmented defects on sewer pipes through a unified neural network, which combined a convolutional neural network with a conditional random field. The experiments showed that the proposed method achieves better segmentation accuracy and speed. Huang et al. [37] proposed a convolutional neural network to segment fabric defects using a few labeled data, which includes a segmentation network and a decision network. The proposed method can achieve high accuracy with almost 50 defect samples. Marino et al. [38] adopted a weakly supervised learning approach to detect potato skin defects. A deep convolutional neural network was first used to detect six types of defects, and then on a defect activation map, a segmentation method was applied to obtain the morphology of potatoes. In the end, the support vector machine was used to classify damaged or greened potatoes. Wang et al. [39] proposed an unsupervised defect segmentation method to

Optics EXPRESS

segment the defects on a 3D-printed workpiece. They introduced a self-attention mechanism into a CNN-based feature learning network to capture more global features.

Although many deep learning-based segmentation methods have been proposed and applied in defect segmentation in many industries, few studies focus on lens defect segmentation. In addition, transformer-based methods show better performance in many computer vision tasks. Therefore, in this study, we aimed to segment lens defects through a transformer-based segmentation method.

3. Enhanced transformer method coupled with transfer learning for defect segmentation

3.1. Overall framework

The proposed method is shown in Fig. 2. E2Trans is an encoder-decoder network based on Segformer [32], which is a simple and efficient segmentation network. Since defects with different sizes, in the encoder, we adopted four hierarchically stacked transformer blocks to learn the different scale features from overlap patch embedding. For the number of transformer blocks, deeper transformers capture long-range dependencies but increase computational costs, and shallower networks focus on local features but lack global context. To achieve local and global context balance and reduce computation cost, we use four transformer blocks. In each block, efficient multi-head self-attention layers and mixed feed-forward network layers are adopted to learn the features from image patches. The main decoder combines coarse to fine features learned by each transformer block by a designed MLP block; then, an MLP layer is used to predict the segmentation mask. To enhance the segmentation performance, two auxiliary decoders with the same structure as the main decoder are added to compute auxiliary loss. For the weight of auxiliary decoders, higher weights on early layers can encourage low-level feature learning and higher weights on deep layers can ensure high-level semantic consistency. In this paper, we use different weights to balance the low-level and high-level feature learning. This framework includes two stages. In the first stage, the network without auxiliary decoders is trained on a public dataset. Then, in the second stage, the learned basic features are transferred to a lens domain, and the network with auxiliary decoders is further trained to predict lens defects. In the whole training process, images are first converted into overlapping patch embeddings through a convolutional layer with different input sizes, patch sizes, and embedding dimensions.

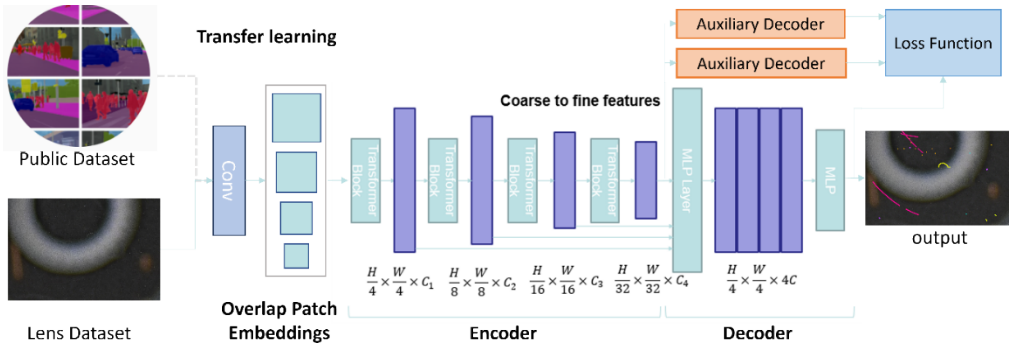


Fig. 2. The overall framework of E2Trans

3.2. Enhanced transformer-based segmentation method

Assume an input image as $x \in R^{H\times W\times 3}$, where H, W are the height and width of the image. Through a convolutional layer with different settings in kernel size (Ke), stride size (S), and padding size (P), the input image is divided into 4×4 patches and multi-level feature maps

Optics EXPRESS

$F_i \in R^{H_i \times W_i \times C_i}$ are obtained, where $H_i = \frac{H}{2^{i+1}}, W_i = \frac{W}{2^{i+1}}$, $i = \{1, 2, 3, 4\}$. The $\{Ke,\ S, P\}$ are set as $\{7, 4, 3\}$ and $\{3, 2, 1\}$ to produce overlapping patches with the same size as the non-overlapping process. In each transformer block, there are T feature extraction blocks followed by an overlapped patch merging layer, in which each block is composed of an efficient self-attention layer and a mixed feed-forward network layer. Since the self-attention layer is a computation bottleneck, the sequence reduction (SeqRe) process [40] is adopted to achieve efficient self-attention (EAttn), which can be defined as Eqs. (1) and (2).

$$x_{EAttn} = EAttn(Q, K, V) = Softmax\left(\frac{Q \cdot SeqRe(K)}{\sqrt{A_h}}\right) \tag{1}$$

$$SeqRe(K) = Linear(C_i \cdot R_i, C_i)\left(Reshape\left(\frac{N_i}{R_i}, C_i \cdot R_i\right)(K)\right) \tag{2}$$

where Q, K, and V are query, key, and value in self-attention with the same size of $N_i \times C_i$ and $N_i = H_i \times W_i$, A_h is the number of attention heads, R_i is the reduction radio, *Reshape* aims to reshape the size of K to $\left(\frac{N_i}{R_i}, C_i \cdot R_i\right)$, and *Linear* aims to output a tensor with C_i dimensions with the $C_i \cdot R_i$-dimensional input. In the mixed feed-forward network (MFFN) layer, a 3×3 convolutional layer is applied to provide position information, which can be expressed as Eq. (3).

$$x_{Mffn} = MFFN(x_{EAttn}) = MLP(GELU(Conv_{3\times3}(MLP(x_{EAttn})))) + x_{EAttn} \tag{3}$$

where GELU is the activation function called Gaussian Error Linear Units, and x_{EAttn} is the output of the efficient self-attention layer. Then, the x_{Mffn} is fed into an overlapped patch merging layer, which is a linear layer to shrink features from $H_i \times W_i \times C_i$ to $H_{i+1} \times W_{i+1} \times C_{i+1}$.

In the main decoder, only MLP layers are applied to obtain the final segmentation masks. An MLP block aims to make all feature maps into feature map $\hat{F}_i \in R^{\frac{H}{4} \times \frac{W}{4} \times C}$ and another MLP layer is adopted to combine the multi-level features from four transformer blocks to perform mask prediction. In the MLP block, it is composed of an MLP layer and an up-sampling layer. The MLP layer firstly maps the feature map F_i with C_i dimensions of each transformer block to a feature map with C dimensions. Then an up-sampling layer is adopted to up-sample the feature maps to $\frac{H}{4} \times \frac{W}{4}$. It can be expressed as Eq. (4).

$$\hat{F}_i = Upsample\left(\frac{H}{4} \times \frac{W}{4}\right)(Linear(C_i, C)(F_i)) \tag{4}$$

In the end, another MLP layer is applied to fuse the features of four transformer blocks, in which the feature maps $\hat{F}_i$ are contacted first and a linear layer is applied to output a feature map with the size of $\frac{H}{4} \times \frac{W}{4} \times C$, as shown in Eq. (5):

$$F = Linear(4C, C)(Concat(\hat{F}_i)) \tag{5}$$

The final prediction layer is a linear layer to generate segmentation mask $M \in R^{\frac{H}{4} \times \frac{W}{4} \times N_{cls}}$. In this paper, we add two auxiliary decoders which have the same structure as the decoder but with different weights to participate in loss function computation so as to improve the segmentation accuracy.

3.3. Loss function

Cross-entropy loss is a common loss function used in segmentation tasks [41,42,43] to estimate the difference between input logits and target. Assume that z is the output of the model, c is the ground truth, and the loss function can be expressed as Eq. (6):

$$l(z, c) = -log\frac{\exp(z_c)}{\sum_{i=0}^{N_{cls}-1}\exp(z_i)} \tag{6}$$

where $z = [z_0, \cdots, z_{N_{cls}-1}]$ is the output of an image and $\exp(*) = e^*$.

Optics EXPRESS

To improve the segmentation performance, we added two auxiliary losses, and the input logits z_1, z_2 are obtained by two auxiliary decoders. Assume that z_{main} is the output of the main decoder; then the whole loss function l_{total} can be defined in Eq. (7):

$$l_{total}(z_{main}, z_1, z_2, c) = l_{main} + w_1 l_{aux1} + w_2 l_{aux2} s.t.\ \ l_{main} = l(z_{main}, c) l_{aux1} = l(z_1, c) l_{aux2} = l(z_2, c) \tag{7}$$

Here, auxiliary decoder loss plays an important role in improving gradient flow, multi-scale feature learning and generalization. When deetermining the gradient during the training process, as shown in Eq. (8),

$$\frac{\partial l_{total}}{\partial F_i} = \frac{\partial l_{total}}{\partial F_i} + w_1 \frac{\partial l_{aux1}}{\partial F_i} + w_2 \frac{\partial l_{aux2}}{\partial F_i} \tag{8}$$

l_{aux1} and l_{aux2} can provide additional gradients to ensure that earlier feature maps are well-trained so as to stabilize optimization and prevent early layer degradation. These additional gradients can provide supervision at multiple feature levels F_i so as to ensure all feature levels are useful. In addition, in the loss function, auxiliary decoder loss also acts as a regularizer, reducing overfitting. This regularizer can prevent deep layers from over-specializing in the final prediction while neglecting intermediate features.

3.4. Evaluation metrics

Three types of evaluation metrics, which are commonly used in segmentation research, were applied to assess the model's performance: intersection over union (*IoU*), pixel accuracy (*PA*), and mean pixel accuracy (*mPA*) [44,45,46]. *IoU* means the overlapping ratio between the predicted segmentation mask and the ground truth, which is defined as Eq. (9):

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

where A and B stand for the ground truth mask and the predicted segmentation mask, respectively. The average *IoU* throughout all classes is known as *mIoU*. *PA* attempts to calculate the ratio between the amount of correctly classified pixels and the total number of pixels so as to realize the pixel-level performance evaluation for all classes. The pixel accuracy of N_{cls} classes ($N_{cls} - 1$ foreground classes and the background) is defined by Eq. (10):

$$PA = \frac{\sum_{i=0}^{N_{cls}-1} p_{ii}}{\sum_{i=0}^{N_{cls}-1} \sum_{j=0}^{N_{cls}-1} p_{ij}} \tag{10}$$

where p_{ij} is the number of pixels of class i predicted as belonging to class j.

Similar to *MIoU*, *PA* also has an extension metric through average PA through all classes, named Mean Pixel Accuracy (mPA) as shown in Eq. (11), in which the ratio of correct pixels is computed in a per-class manner and then averaged over the total number of classes.

$$mPA = \frac{1}{N_{cls}} \sum_{i=0}^{N_{cls}-1} \frac{p_{ii}}{\sum_{j=0}^{N_{cls}-1} p_{ij}} \tag{11}$$

4. Experimental

4.1. Experimental setting and dataset collection

To train the proposed model, a labeled lens defect dataset is necessary. Firstly, many NMCSLs, shown in Fig. 3(b), were manufactured by an injection molding machine, Toyo Si-80, which is shown in Fig. 3(a). The entire injection molding experiment consisted of three steps: filling,

Optics EXPRESS

packing, and cooling, as shown in Fig. 3(e). A pile of plastic pellets was first heated, far above the transition temperature of the polymer. Then the appropriate injection pressure and speed were set to ensure that enough melt can be injected into the mold cavity to complete the filling. Once the cavity was filled, the pressure was maintained to maintain the cavity and compensate for the volume shrinkage of the lens. Finally, a cooler is used to cool the molded assembly to room temperature to ensure that the lens is solidified and can be successfully ejected. When cooling was completed, the mold was opened, and the finished lenses were released. Therefore, the process is difficult to control since too many process parameters and physical processes are involved in it. In this experiment, the parameter settings are shown in Table 1, where the melt temperature, mold temperature, injection pressure, holding pressure, holding time, and injection time were carefully set to fabricate NMCSLs so as to meet the requirements of glasses as shown in Fig. 1(b). Hence, NMCSLs that have defects were selected for the next step. The defect images were captured by a vision system (Fig. 3(c)), which is composed of a camera, a lens, a light source, and a computer. The detailed information is listed in Table 2. Then the five types of defects were labeled by labelme [47] and preprocessed for model training. As shown in Fig. 4, notches, black spots, bubbles, fibers, and scratches on the lens were labeled. In the dataset, compared with other defects, bubbles are the most frequent defects. Since bubbles are too small to be detected, we did not conduct undersampling to let the model learn more. The dataset has a total of 220 images, and the ratio of training and validation dataset is 0.9:0.1, where data enhancement techniques, including random resizing with a ratio range from 0.5 to 2.0, random cropping with a ratio of 0.75, and random flipping with a probability of 0.5 were adopted in each training epoch so that model does not see same images to learn more features. In addition, in each image, there are 20-50 defects, therefore we have around 10000 defects to learn the features. Furthermore, we first train the model on the public dataset with 5000 images to learn basic features from a large dataset. Since there are some limitations of train-test split in the small dataset, such as the risk of overfitting and unrepresentative validation data, 10-fold cross-validation experiments were conducted to show the segmentation performance on each defect class of the proposed model.

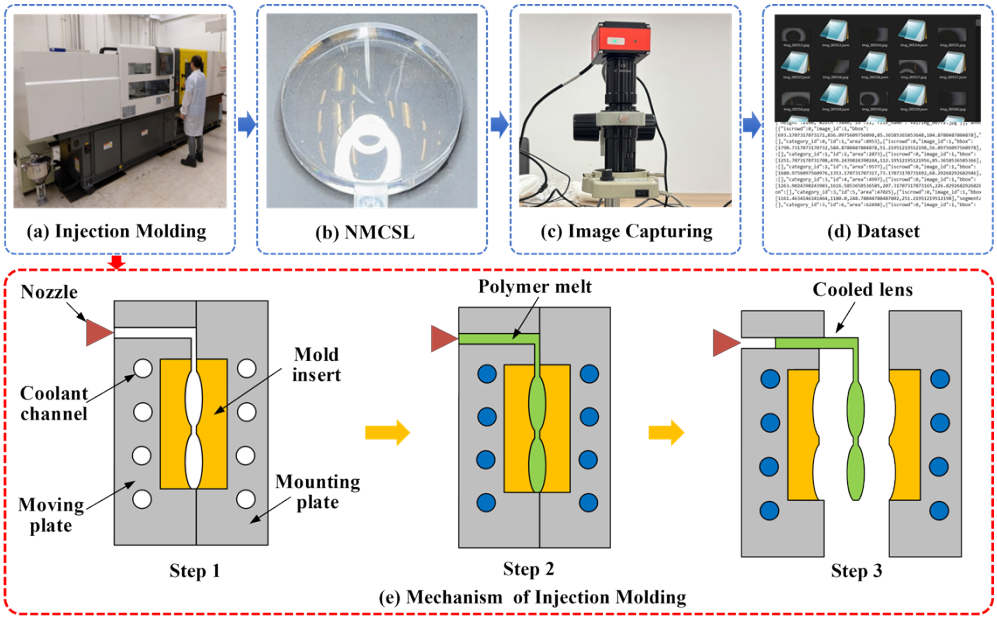


Fig. 3. The experimental mechanism and data collection.

Optics EXPRESS

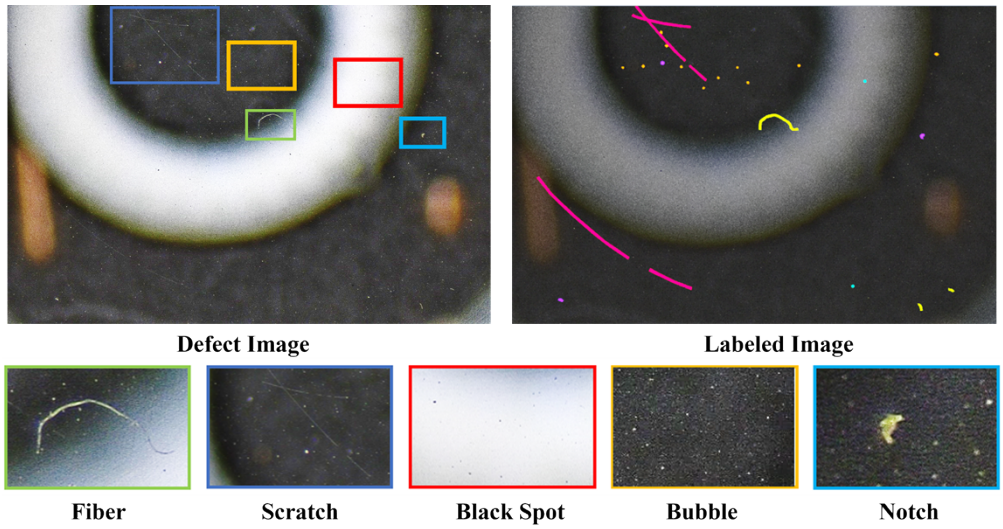


Fig. 4. The five types of lens defects.

Table 1. The parameter setting of the injection-molding process

Parameters	Values
Precision injection molding machine	Toyo Si-80
Melt Temperature	235-250 C°
Mold Temperature	146 C°
Injection Pressure	1600 kgf/cm^2
Holding Pressure	1600-1450 kgf/cm^2
Holding Time Injection Time Cooling Time	70s 21s 75s

Table 2. The parameter settings of the image capturing system

Camera model	RZSP-4KCH-IV
Resolution	3840*2160
Pixel size	2.0μm*2.0μm
Frame rate	60 FPS
Lens model	SPZ0745-ZC1.0
Objective magnification	0.75-5X
Light source	LED ring light

Optics EXPRESS

4.2. Implementation details

The input image size was set as 2048×640. The number of self-attention heads, A_h, in each transformer block was set to $\{1, 2, 5, 8\}$. The numbers of feature extraction blocks, T, in four transformer blocks were set to 3,6,40, and 3, respectively. The reduction ratios $\{R_i | i = 1, 2, 3, 4\}$ and the numbers of feature dimensions $\{C_i | i = 1, 2, 3, 4\}$ of four transformer blocks were set to $\{8, 4, 2, 1\}$ and $\{64, 128, 320, 512\}$, respectively. The weights, w_1, w_2, of two auxiliary decoders were set to 0.4 and 0.4, respectively. The optimizer was AdamW, which is the Adam with weight decay. Here, the rate of weight decay was set to 0.01 and the learning rate was $6 * 10^{-5}$. The learning rate started from $1 * 10^{-6}$ and linearly changed in the first 1,500 iterations. Then the polynomial learning rate decay was adopted from 1,500 iterations until the training process ended. The model was first trained on the Cityscapes dataset [48], and then our lens defect dataset was fed into the proposed model to fine-tune the model. The model was trained using 16,000 iterations. In this study, all experiments were conducted on a computer with an i9-12900 H 5.1 GHz CPU with 16 Cores and NVIDIA GEFORCE RTX A4500 GPU with a memory of 20GB.

4.3. Results and discussion

As shown in Fig. 5, all the types of defects on lenses can be segmented although they have different morphologies and sizes which are marked by different colors. Since some bubbles and black spots are too small, we highlighted them with color boxes. Each class segmentation accuracy is listed in Table 3, which is the average accuracy of 10-fold cross-validation experiments. The results show that fiber defects can be segmented easier than other kinds of defects with an accuracy of 47.48 in *IoU* and 72.41 in *PA*. The reason is that fiber defects have distinct features in both bright field and dark fields. Notch defects have lower segmentation accuracy, reaching 37.89 of *IoU* and 47.37 of *PA*. The segmentation performance on bubble defects is low since it is so small, and the features of them are similar to the background.

Table 3. The performance of E2Trans in each class

Class	IoU	PA
Background	99.89	99.96
Notches	37.89	47.37
Black Spots	27.25	34.96
Bubbles	7.44	8.38
Fibers	47.48	72.41
Scratches	28.89	40.95

Transfer learning effect. To explore the transfer learning influences on the proposed model, we compared the model performance under three conditions, including no transfer learning, transfer learning from the ADE dataset [49], and transfer learning from the Cityscapes dataset. All the experiments were conducted with the model with four transformer blocks, and the auxiliary decoder weights were set to 0.4. According to the results listed in Table 4, the segmentation performances of transfer learning from the Cityscapes dataset on *mPA* and *mIoU* were both better than the other kinds of conditions.

Table 4. Transfer learning performance comparison

Transfer learning Dataset	PA	mPA	mIoU
None	99.93	52.48	43.51
ADE Dataset	99.91	50.07	40.07
Cityscapes Dataset	99.92	53.90	44.02

Optics EXPRESS

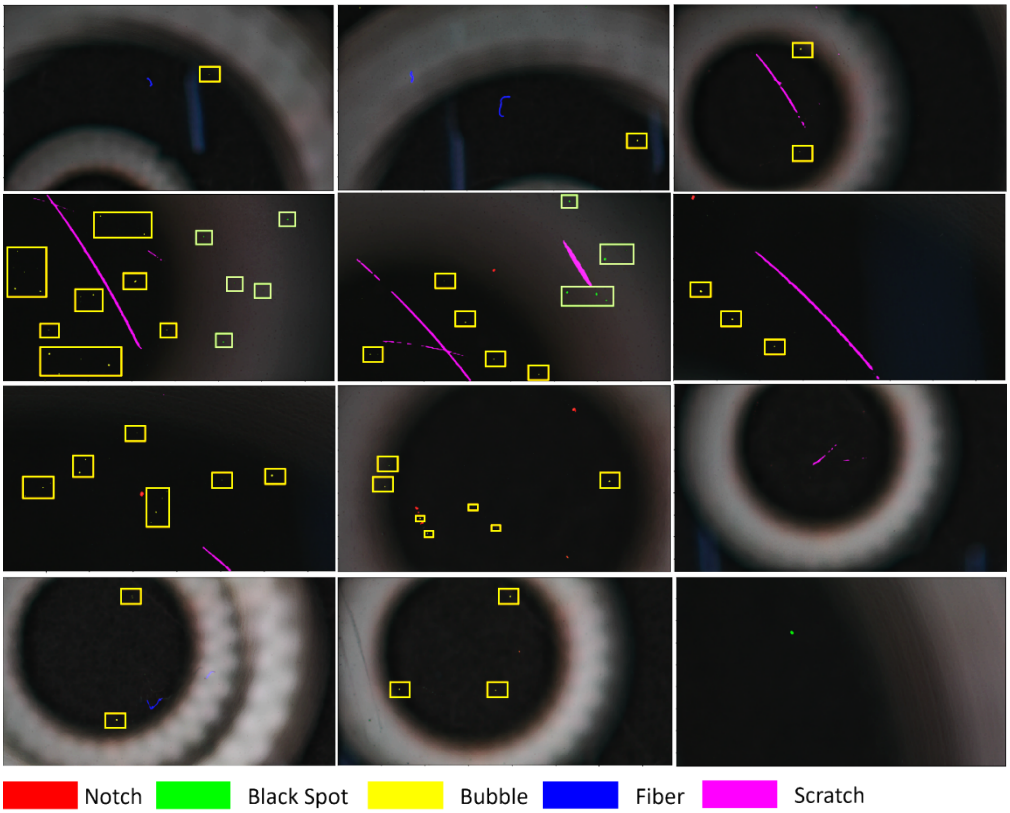


Fig. 5. The defect segmentation results of E2Trans.

Auxiliary decoder effect. Comparison experiments with different auxiliary decoder (AD) settings were also conducted to explore the segmentation accuracy of our model. Firstly, we compared the performance of the model without ADs (w/ TL_4 w/o AD), the model with ADs when the weights were set to 0.4 (w/ TL_4 w/ AD_0.4), and the model with ADs when the weights were set to 0.6 (w/ TL_4 w/ AD_0.6). All these three models were pretrained on the Cityscapes dataset and with four transformer blocks. From the results listed in Table 5, we can see that the best performance was achieved when the weights of ADs were set to 0.4. To further explore the weights of Ads' influence on segmentation performance, we conducted two experiments on the model without transfer learning in which the number of transformer blocks was 6. The weights of ADs in these experiments were set to 0.4 and 0.6, separately. As shown in Table 5, the performance of w/o TL_6 w/ AD_0.6 model had better performance. Therefore, the weights of ADs had a significant effect on the segmentation performance, and the value settings will depend on the model structures.

Transformer block effect. Since the transformer blocks play an important part in the segmentation performance, we conducted a series of experiments to explore the segmentation accuracy of different numbers of transformer blocks. As shown in Table 6, we compared the performance of the model with four, five, and six transformer blocks. All the experiments were conducted on the model without transfer learning, and the weights of ADs were set to 0.4. The results showed that the model with six transformer blocks had the best segmentation performance. The worst performance was obtained when the number of transformer blocks was set to five. Therefore, the number of transformer blocks needs to be carefully set.

Optics EXPRESS

Table 5. The auxiliary decoder performance comparison

Model	PA	mPA	mIoU
w/ TL_4 w/o AD	99.92	50.68	42.36
w/ TL_4 w/ AD_0.4	99.92	53.90	44.02
w/ TL_4 w/ AD_0.6	99.92	50.56	42.11
w/o TL_6 w/ AD_0.4	99.92	49.48	42.06
w/o TL_6 w/ AD_0.6	99.92	50.40	41.10

Table 6. The transformer block performance comparison

#Transformer Blocks	PA	mPA	mIoU
6	99.92	49.48	42.06
5	99.91	51.08	40.42
4	99.93	52.48	43.51

Other types of lenses. To show the generalizability of our model, we evaluated segmentation performance on glass molded lenses with a molding temperature of 590 C°, molding speed of 0.4 mm/s, and molding time of 10 s. As shown in Fig. 6, the results show that our model can be extended to different processes for producing lenses.

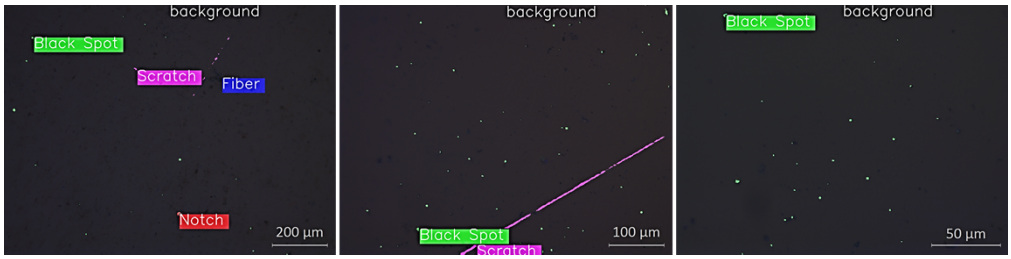


Fig. 6. The defect segmentation results on glass molded lenses.

In the end, we compared the performance of E2Trans with state-of-the models, such as Segformer series models, Mask2Former [33], PSPNet [50], DeepLabv3+ [51], SETR [52], and STDC2 [53] in regard to segmentation accuracy and inference time. Here, the backbone of DeepLabv3 + is ResNet-101; the decoder of SETR is the naive decoder. The results listed in Table 7 show that our model achieved good segmentation accuracy while maintaining a high inference speed, reaching 53.90 mPA and 0.214 seconds. The segmentation accuracy of Segfomer series models becomes better and better as the network depth increases. The Mask2Former achieved better performance, but the segmentation inference time was long. In contrast, SETR had the shortest inference time but suffered from low segmentation performance. As a result, our model is more suitable for lens defect defection since mass production needs a highly efficient detection method.

4.4. Lens defect detection system

To show the performance of the detection system, we conducted a lens defect segmentation experiment on a developed detection system, as shown in Fig. 7. The vision system consists of a 4 K high-definition camera (RZSP-4KCH-IV), a high-magnification zoom lens with a zoom range of 0.7-5 and the working distance of 175, a ring LED light source, and a server with GPU. The computer was connected to the camera through a GigE interface and obtained real-time video

Optics EXPRESS

Table 7. The performance comparison with state-of-the-art models

Model	PA	mPA	mIoU	Inference Time (s)
Segformer-B0	99.92	35.29	31.63	0.0836
Segformer-B1	99.92	35.45	32.60	0.0799
Segformer-B2	99.92	38.66	33.52	0.0776
Segformer-B3	99.92	39.53	33.88	0.0977
Segformer-B4	99.85	41.89	28.82	0.1543
Segformer-B5	99.92	52.24	35.10	0.1622
Mask2Former	99.93	66.79	53.95	0.5577
PSPNet	99.91	36.63	30.46	0.1797
DeepLabv3+	99.90	41.57	33.70	0.2549
SETR	99.89	35.15	29.20	0.0146
STDC2	99.91	39.66	33.75	0.2610
E2Trans	99.92	53.90	44.02	0.2014

stream information through an IP address. Here, category 6 cable was adopted. The algorithm was then deployed on the server for platform calling. The monitor is used to show the detection results. On the platform, as shown in Fig. 8, the hardware device was connected by clicking the 'connect camera' button to achieve real-time photo acquisition, while the 'capture' button is used to take and save photos, and the 'detection' button is used to call the algorithm and output the results. The 'Logs' shows the details of operations. The 'Result' provides the final decision. The 'Defect Details' gives each defect's name and location. The 'Reasons' provide the possible reasons for defects. This study shows that our algorithm can segment defects correctly, and the defection system is running smoothly.

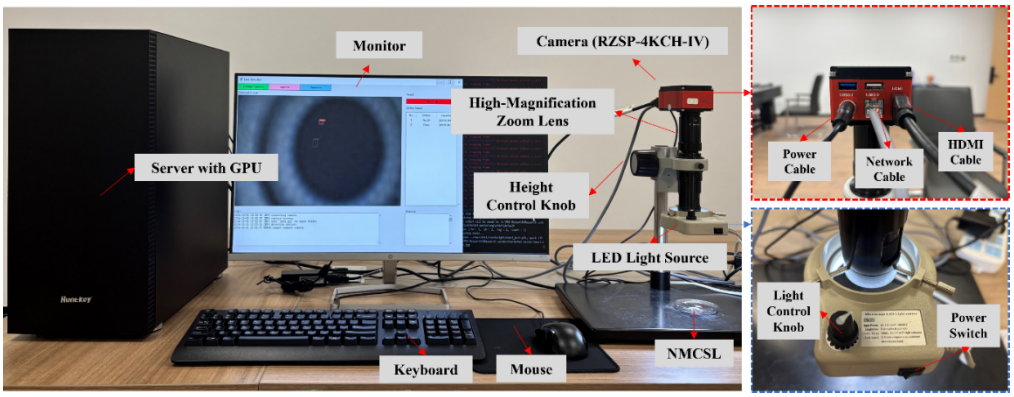


Fig. 7. The configuration of the lens detection system.

Optics EXPRESS

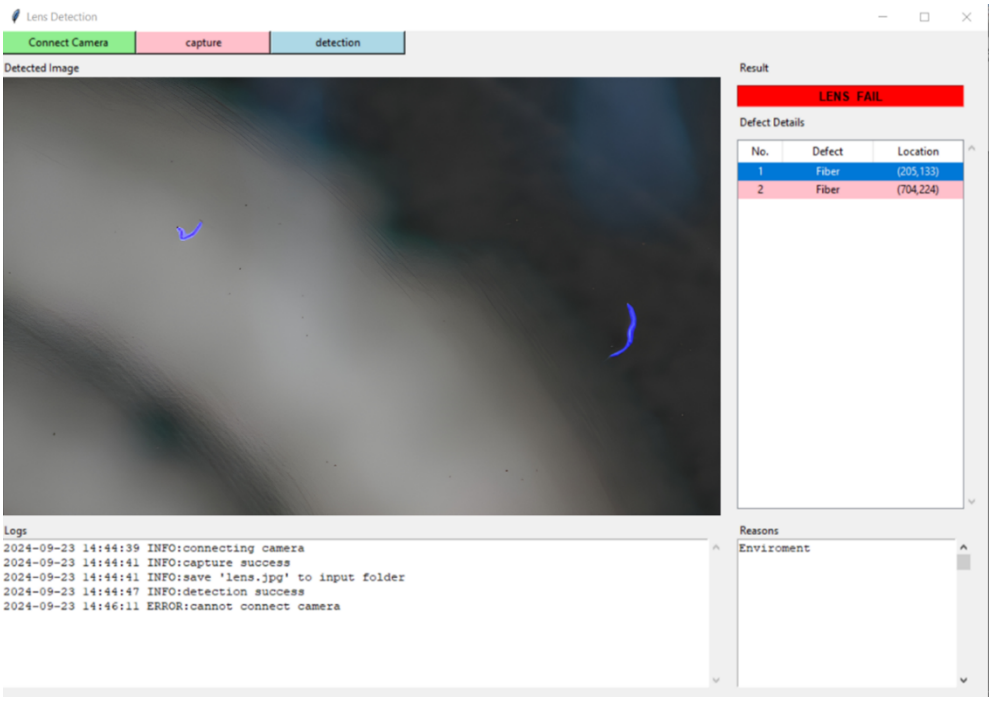


Fig. 8. The detection results of the lens detection system.

5. Conclusions

In this paper, a defect segmentation method, named E2Trans is presented to defect five myopia control spectacle lens defects, including notches, black spots, bubbles, fibers, and scratches, in which two auxiliary decoders are added to improve the segmentation performance. It utilizes the powerful feature extraction ability of transformers and the knowledge transfer ability of transfer learning to achieve highly efficient defect segmentation with good segmentation accuracy. To train the model, the lens defect dataset was collected by manufacturing many myopia control spectacle lenses via injecting molding. The images were captured by an electric microscope system and labeled by hand. A series of experiments were conducted to show our model's superiority in regard to segmentation accuracy and speed with 53.90 *mPA*, 44.02 *mIoU*, and 0.2014 s inference time. A case study was conducted to show that the developed system is capable of detecting defects smoothly.

Funding. Research Grants Council of the Government of the HKSAR (C5031-22G); Research Centre of Smart Vision of Hong Kong Polytechnic University (BBCU); Contract Research Project between The Hong Kong Polytechnic University and Vision Science Technology Co. Ltd. (ZDCP).

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. B. A. Holden, T. R. Fricke, D. A. Wilson, *et al.*, "Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050," Ophthalmology **123**(5), 1036–1042 (2016).
2. A. Grzybowski, P. Kanclerz, K. Tsubota, *et al.*, "A review on the epidemiology of myopia in school children worldwide," BMC Ophthalmol. **20**(1), 27 (2020).
3. M. He, F. Xiang, Y. Zeng, *et al.*, "Effect of time spent outdoors at school on the development of myopia among children in China: a randomized clinical trial," JAMA **314**(11), 1142–1148 (2015).

Optics EXPRESS

4. A. E. G. Haarman, C. A. Enthoven, J. W. L. Tideman, *et al.*, "The Complications of myopia: a review and meta-analysis," Invest. Ophthalmol. Vis. Sci. **61**(4), 49 (2020).
5. M. Yap, J. Cho, and G. Woo, "A survey of low vision patients in Hong Kong," Clin. Exp. Optom. **73**(1), 19–22 (1990).
6. B. Seet, T. Y. Wong, D. T. Tan, *et al.*, "Myopia in Singapore: taking a public health approach," Br. J. Ophthalmol. **85**(5), 521–526 (2001).
7. B.J. Curtin, "The myopias. basic science and clinical management," Philadelphia: Harper & Row (1985).
8. D. Y. Tse, C. S. Lam, J. A. Guggenheim, *et al.*, "Simultaneous defocus integration during refractive development," Invest. Ophthalmol. Vis. Sci. **48**(12), 5352 (2007).
9. C. S. Lam, W. C. Tang, D. Y. Tse, *et al.*, "Defocus Incorporated Soft Contact (DISC) lens slows myopia progression in Hong Kong Chinese schoolchildren: a 2-year randomised clinical trial," Br. J. Ophthalmol. **98**(1), 40–45 (2014).
10. J. Obregon, J. Hong, and J. Y. Jung, "Rule-based explanations based on ensemble machine learning for detecting sink mark defects in the injection moulding process," J. Manuf. Syst. **60**, 392–405 (2021).
11. M. F. Bianchi, A. A. Gameros, D. A. Axinte, *et al.*, "Regional temperature control in ceramic injection moulding: An approach based on cooling rate optimization," J. Manuf. Process. **68**, 1767–1783 (2021).
12. J. C. Chen, G. Guo, and W. N. Wang, "Artificial neural network-based online defect detection system with in-mold temperature and pressure sensors for high precision injection molding," Int. J. Adv. Manuf. Tech. **110**(7-8), 2023–2033 (2020).
13. D. K. Anguraj, "A Bayesian regularization approach to predict the quality of injection-moulded components by statistical SVM for online monitoring system," J. Ubiqui. Comput. Commun. Tech. **3**(4), 277–288 (2022).
14. A. V. Hulagadri, "Quality classification of defective parts from injection moulding," arXiv (2020).
15. D. Karimi, Q. Zeng, P. Mathur, *et al.*, "Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images," Med. Image Anal. **57**, 186–196 (2019).
16. L. G. Divyanth, A. Ahmad, and D. Saraswat, "A two-stage deep-learning based segmentation model for crop disease quantification based on corn field imagery," Smart Agricultural Technology **3**, 100108 (2023).
17. H. Su and J. Lee, "Machine learning approaches for diagnostics and prognostics of industrial systems using open source data from PHM data challenges: a review," Int. J. Progn. Health M. **15**(2), 3993 (2024).
18. T. Fu, P. Li, and S. Liu, "An imbalanced small sample slab defect recognition method based on image generation," J. Manuf. Process. **118**, 376–388 (2024).
19. C. Xu, G. Wei, Y. Guan, *et al.*, "High-performance deep learning segmentation for non-destructive testing of X-ray tomography," J.Manuf. Process. **128**, 98–110 (2024).
20. S. Minaee, Y. Boykov, F. Porikli, *et al.*, "Image segmentation using deep learning: A survey," IEEE Trans. Pattern Anal. Mach. Intell. **44**(7), 3523–3542 (2021).
21. L. C. Chen, G. Papandreou, I. Kokkinos, *et al.*, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," arXiv (2014).
22. G. Lin, C. Shen, Van Den Hengel, *et al.*, "Efficient piecewise training of deep structured models for semantic segmentation," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3194–3203 (2016).
23. A. Paszke, A. Chaurasia, S. Kim, *et al.*, "Enet: A deep neural network architecture for real-time semantic segmentation," arXiv (2016).
24. B. Shuai, Z. Zuo, B. Wang, *et al.*, "Scene segmentation with dag-recurrent neural networks," IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1480–1493 (2018).
25. X. Liang, X. Shen, J. Feng, *et al.*, "Semantic object parsing with graph lstm.,". In *Computer Vision–ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14 125–143 (2016).
26. Y. Xiang and D. Fox, "DA-RNN: Semantic mapping with data associated recurrent neural networks," arXiv (2017).
27. P. Luc, C. Couprie, S. Chintala, *et al.*, "Semantic segmentation using adversarial networks," arXiv (2016).
28. N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," In *Proceedings of the IEEE international conference on computer vision*, 5688–5696 (2017).
29. Y. Xue, T. Xu, H. Zhang, *et al.*, "Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation," Neuroinformatics **16**(3-4), 383–392 (2018).
30. X. Li, H. Ding, W. Zhang, *et al.*, "Transformer-based visual segmentation: A survey," arXiv (2023).
31. Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," In *Proceedings of the IEEE/CVF international conference on computer vision*,10012–10022 (2021).
32. E. Xie, W. Wang, Z. Yu, *et al.*, "SegFormer: Simple and efficient design for semantic segmentation with transformers," Adv. Neural Infor. Process. Syst. **34**, 12077–12090 (2021).
33. B. Cheng, I. Misra, A. G. Schwing, *et al.*, "Masked-attention mask transformer for universal image segmentation," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299 (2022).
34. J. Jain, J. Li, M. T. Chiu, *et al.*, "Oneformer: One transformer to rule universal image segmentation," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2989–2998 (2023).
35. R. Neven and T. Goedemé, "A multi-branch U-Net for steel surface defect type and severity segmentation," Metals-Basel **11**(6), 870 (2021).
36. M. Wang and J. C. Cheng, "A unified convolutional neural network integrated with conditional random field for pipe defect segmentation," Comput.-Aided Civ. Inf. **35**(2), 162–177 (2020).

Optics EXPRESS

37. Y. Huang, J. Jing, and Z. Wang, "Fabric defect segmentation method based on deep learning," IEEE Trans. Instrum. Meas. **70**, 1–15 (2021).
38. S. Marino, P. Beauseroy, and A. Smolarz, "Weakly-supervised learning approach for potato defects segmentation," Eng. Appl. Artif. Intel. **85**, 337–346 (2019).
39. R. Wang, C. F. Cheung, and C. Wang, "Unsupervised defect segmentation in selective laser melting," IEEE Trans. Instrum. Meas. **72**, 1–10 (2023).
40. W. Wang, E. Xie, X. Li, *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578(2021).
41. K. Hu, Z. Zhang, X. Niu, *et al.*, "Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function," Neurocomputing **309**, 179–191 (2018).
42. H. Xu, M. Yang, L. Deng, *et al.*, "Neutral cross-entropy loss based unsupervised domain adaptation for semantic segmentation," IEEE Trans. Image Process. **30**, 4516–4525 (2021).
43. M. Yeung, E. Sala, C. B. Schönlieb, *et al.*, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," Comput. Med. Imag. Grap. **95**, 102026 (2022).
44. D. Li, Z. Duan, X. Hu, *et al.*, "Pixel-level recognition of pavement distresses based on U-Net," Adv. Mater. Sci. Eng. **2021**(3), 1–11 (2021).
45. S. Li, X. Zhao, and G. Zhou, "Automatic pixel-level multiple damage detection of concrete structure using fully convolutional network," Comput.-Aided Civ. Inf. **34**(7), 616–634 (2019).
46. H. Liu, L. Bian, and J. Zhang, "Image-free single-pixel segmentation," Opt. Laser Technol. **157**, 108600 (2023).
47. Tzutalin, "LabelIing: Git code," Github, (2022) [accessed 4 October 2023]. Retrieved from https://github.com/tzutalin/labelImg
48. M. Cordts, M. Omran, S. Ramos, *et al.*, "The cityscapes dataset for semantic urban scene understanding," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223 (2016).
49. S. Shao, Z. Li, T. Zhang, *et al.*, "Objects365: A large-scale, high-quality dataset for object detection," In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439 (2019).
50. H. Zhao, J. Shi, X. Qi, *et al.*, "Pyramid scene parsing network," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890 (2017).
51. L. C. Chen, Y. Zhu, G. Papandreou, *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," In *Proceedings of the European conference on computer vision (ECCV)*, 801–818 (2018).
52. S. Zheng, J. Lu, H. Zhao, *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890 (2021).
53. M. Fan, S. Lai, J. Huang, *et al.*, "Rethinking bisenet for real-time semantic segmentation," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9716–9725 (2021)