applied sciences

MDPI

Article

Pavement Distress Detection Using Three-Dimension Ground Penetrating Radar and Deep Learning

Jiangang Yang [1], Kaiguo Ruan [2], Jie Gao [1,*], Shenggang Yang [2] and Lichao Zhang [2]

[1] School of Civil Engineering and Architecture, East China Jiaotong University, Nanchang 330013, China; 2851@ecjtu.edu.cn

[2] School of Transportation and Logistics, East China Jiaotong University, Nanchang 330013, China; ruankaiguo@ecjtu.edu.cn (K.R.); yangshenggang@ecjtu.edu.cn (S.Y.); zlc5566@ecjtu.edu.cn (L.Z.)

* Correspondence: gaojie@ecjtu.edu.cn

Abstract: Three-dimensional ground penetrating radar (3D GPR) is a non-destructive examination technology for pavement distress detection, for which its horizontal plane images provide a unique perspective for the task. However, a 3D GPR collects thousands of horizontal plane images per kilometer of the investigated pavement. The existing detection methods using GPR images are time-consuming and risky for subjective judgment. To solve the problem, this study used deep learning methods and 3D GPR horizontal plane images to detect pavement structural distress, including cracks, repairs, voids, poor interlayer bonding, and mixture segregation. In this study, two deep learning methods, called CP-YOLOX and SViT, were used to achieve the aim. A dataset for anomalous waveform localization (3688 images) was first created by pre-processing 3D-GPR horizontal plane images. A CP-YOLOX model was then trained to localize anomalous waveforms. Five SViT models with different numbers of encoders were adopted to perform the classification of anomalous waveforms using the localization results from the CP-YOLOX model. The numerical experiment results showed that 3D GPR horizontal plane images have the potential to be an assistant for pavement structural distress detection. The CP-YOLOX model achieved 87.71% precision, 80.64% mAP, and 33.57 sheets/s detection speed in locating anomalous waveforms. The optimal SViT achieved 63.63%, 68.12%, and 75.57% classification accuracies for the 5-category, 4-category, and 3-category datasets, respectively. The proposed models outperformed other deep learning methods on distress detection using 3D GPR horizontal plane images. In the future, more radar images should be collected to improve the accuracy of SViT.

Keywords: 3D GPR; horizontal radar images; deep learning; distress recognition and localization

check for updates

Citation: Yang, J.; Ruan, K.; Gao, J.; Yang, S.; Zhang, L. Pavement Distress Detection Using Three-Dimension Ground Penetrating Radar and Deep Learning. *Appl. Sci.* **2022**, *12*, 5738. https://doi.org/10.3390/app12115738

Academic Editors: Jinyi Lee, Azouaou Berkache, Zheng Tong and Minhhuy Le

Received: 8 May 2022
Accepted: 3 June 2022
Published: 5 June 2022

1. Introduction

Three-dimensional ground penetrating radar (3D GPR) is an emerging non-destructive inspection technology that is efficient, accurate, and multi-dimensional [1,2]. It has been a major tool for pavement distress detection and pavement condition evaluation [3–5]. Pavement distress detection is defined as the process of classifying and locating instances of pavement distresses in images or videos. Compared to a 2D GPR, 3D GPR uses stepping frequency and antenna array technologies to collect the full structure data of a pavement section. Informative 3D data can be used to detect internal pavement distress, such as cracks, repairs, voids, poor interlayer bonding, and mixture segregation [6]. However, a 3D GPR can obtain thousands of radar images per kilometer in different dimensions. GPR images have been processed using traditional machine learning algorithms in several studies [7]. Rebecca M.W. et al. [8] combined support vector machines (SVMs) and hidden Markov models (HMMs) for Crevasse detection in ice sheets. Zhou et al. [9] combined SVM with H-Alpha Decomposition for subsurface target classification of GPR. The existing processing methods are fallible and unreliable for pavement distress detection using 3D GPR data.

In the last decade, machine learning has produced breakthroughs due to the rapid development of deep learning. Deep learning provided excellent performance in the fields of image recognition, speech recognition, and information security [10–12]. In particular, convolutional neural networks (CNNs) have made excellent achievements in object detection thanks to their powerful feature extraction architecture [13,14]. Additionally, the vision transformer (ViT) model, which has emerged in the last two years, also has achieved remarkable success in object detection [15,16]. Such cases provide a new idea for GPR image processing [17]. For example, Liu et al. [18] detected and located reinforced steel bars in concrete using GPR images and a Single Shot MultiBox Detector (SSD) model with good accuracy and detection speed. Li et al. [19] and Liu et al. [20] achieved the automatic detection of concealed cracks and voids by using the You Only Look Once (YOLO) model and 3D GPR, respectively. Sha et al. [21] proposed three CNN models to classify, localize, and measure structural cracks and potholes in asphalt pavements using GPR images. Hou et al. [22] proposed a data enhancement method based on a convolutional self-encoder structure, which significantly improved the accuracy of crack classification. Yan et al. [23] proposed a pavement distress detection model based on faster region convolutional neural network (Faster R-ConvNet), which reduced the ratio of missing and false detection. In addition, Sha et al. [24] used cascaded CNN models to overcome the low-accuracy problem of traditional CNNs in identifying low-resolution GPR images. Tong et al. and Gao et al. [25–27] used GPR images to identify, locate, measure, and reassemble 3D models of internal cracks by building CNN models. In addition, they developed a Faster R-ConvNet model to accurately identify internal pavement distress (reflection cracks, water-damage pits, and uneven settlements). Moreover, GPR signals were directly input into a network-in-network architecture to achieve the training and testing of the model, and the results showed that the method was effective in detecting cracks, water-damage pits, and uneven settlements. Long [28] used a 3D-to-2D data transformer for reverse-time offset imaging and adopted a single-shot detector to investigate subsurface structures. Wang et al. [29] implemented GPR data enhancement using cycle generative adversarial networks and localized radar hyperbolic waveforms by a Faster R-ConvNet. Kim et al. [30] combined GPR images from different channels into ones for pavement distress detection, and the results showed that the method effectively reduced the error rate. Omwenga M. M. et al. [31] proposed deep reinforcement learning (DRL)-based autonomous cognitive GPR (AC-GPR) to achieve the automatic detection of subsurface targets with superior accuracy and speed.

The waveform characteristics of pavement distress vary from one channel to another in a 3D GPR. The majority of the reported studies focus on the longitudinal radar images since most distresses are obvious in these images. However, these methods ignore a fact that 3D GPR can obtain pavement internal information in three dimensions and the horizontal radar images also can provide a unique view of pavement distress detection. However, there are very few research studies that study this problem.

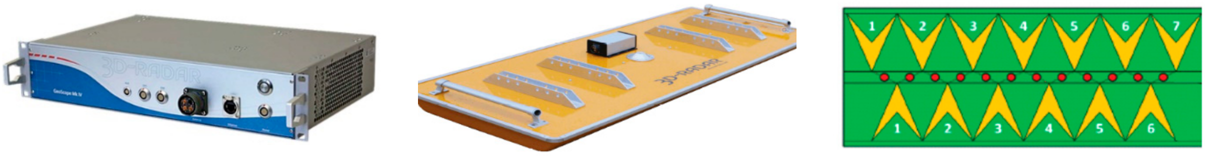
Motivated by the above reason, this study aims to detect pavement distress based on the waveform characteristics of different distress in horizontal radar images. In horizontal radar images, anomalous waveforms are located by a CP-YOLOX model, which is a modified version of You Only Look Once X (YOLOX) [32]. On this basis, to investigate the capacity of horizontal GPR images to represent different pavement distresses, the localization results from the CP-YOLOX model were intercepted to build a classification data set according to three class-membership strategies. Five SViT models, as simplified versions of Vision Transformer [16], classify anomalous waveforms into one of the possible distress categories. The objective of this study is to identify distress types from horizontal radar images using deep learning methods and to provide assistance for distress detection in the future by combining longitudinal radar images.

The rest of this paper is organized as follows. The methods of our study are presented in Section 2, which include the collection of 3D GPR images, the processing method of anomalous waveform localization using the CP-YOLOX model, and the method for distress classification using SViT models. Section 3 presents the numerical experiment results. The conclusions are summarized in Section 4.

2. Proposed Approaches

2.1. Acquisition and Pre-Processing GPR Images

GeoScope 3D Radar was used to collect 3D GPR data, which mainly consisted of a GeoScope^{TM} MK IV mainframe, a DXG1212 shallow ground-coupled antenna array, an RT3D acquisition software, and a 3dr-Examiner data processing software, as shown in Figure 1. The antenna matching mode was a conventional mode with 12 channels; trigger spacing was 5 cm; time window was 25 ns; dwell time was 3 µs. A 3D GPR using a stepping frequency of 100~3000 MHz and antenna array technology captured the full range of wavelengths of pavement interior information during data acquisition, achieving a balance between depth and resolution in a single acquisition.



(**a**) (**b**) (**c**)

Figure 1. GeoScope 3D Radar equipment: (**a**) radar mainframe, (**b**) multichannel antenna array, (**c**) Layout of the DXG1212 antenna array, where the numbers 1–7 indicate the serial number.

The investigated road sections were on Zhangshu-Ji'an Highway, Jiangxi, China. Structure types I–IV in Figure 2 show four main types of pavement structures on the highway. The lengths of the road sections with structure types I, II, III, and IV are 21.9 km, 8.3 km, 175.4 km, and 4.0 km, respectively. The four structures were the most common in China. All 1574 original GPR images were collected and each of them represents a road section with a length of 60 m and a width of 0.8 m. The 416 × 416 resolution image is a common input size for the YOLO model, and the detector can detect anomalous waveforms well at this resolution. Therefore, these images were cropped and resized to finally obtain 3688 horizontal radar images with a resolution of 416 × 416, as shown in Figure 3. The actual width and length of each cropped radar image were 0.8 m and 20 m, respectively.

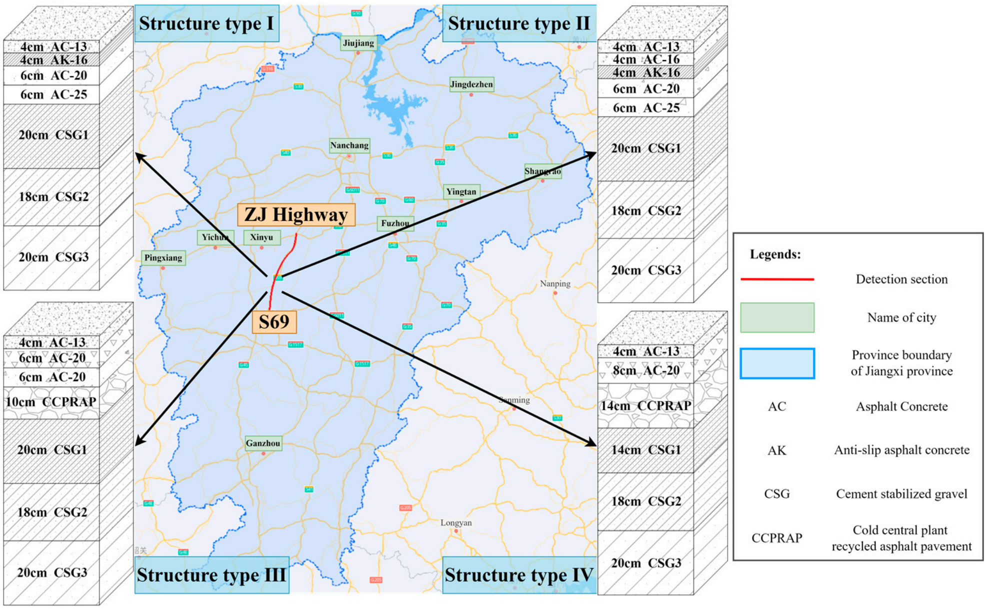


Figure 2. Zhangshu-Ji'an Highway, which had four different pavement structures.

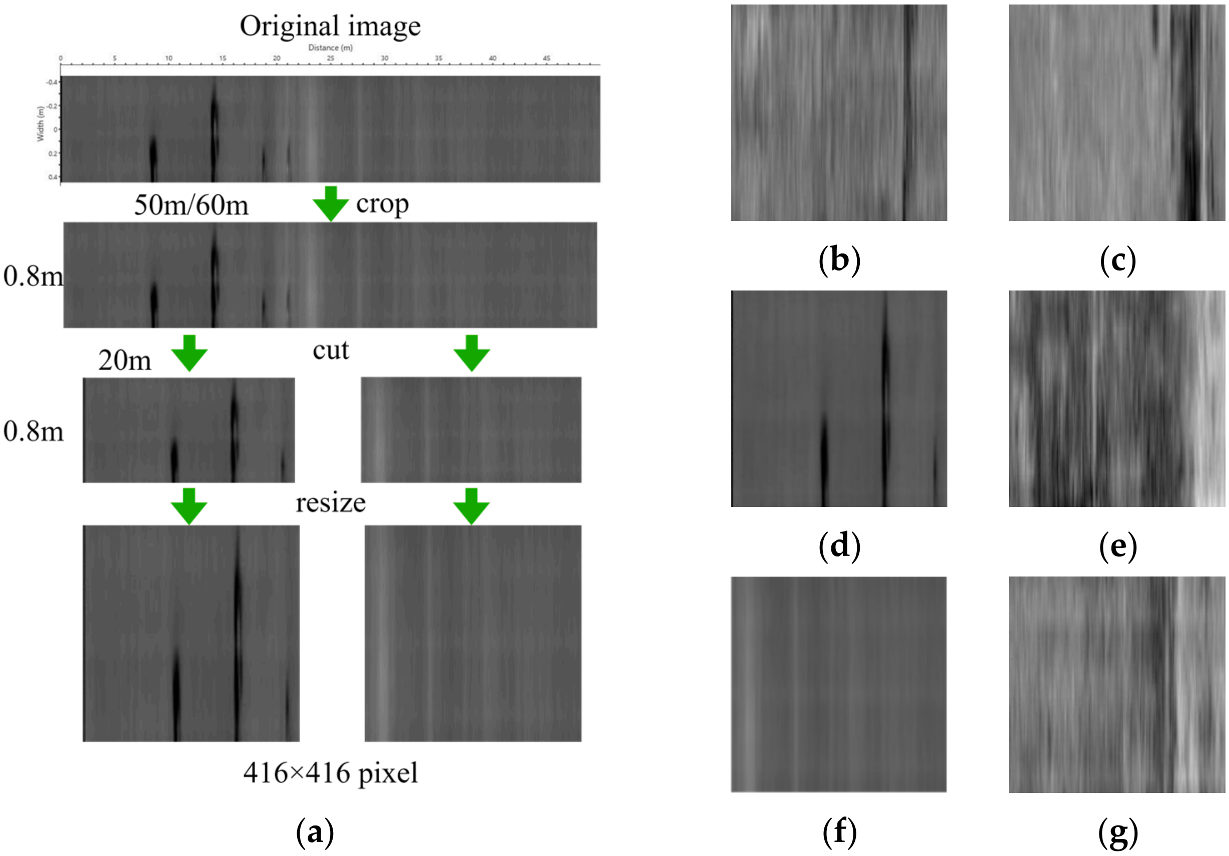


Figure 3. Pre-processing of radar images: (**a**) processing of horizontal radar images; (**b**–**g**) cropped horizontal radar images.

A crack is a typical type of distress in the pavement and appears in horizontal radar images as stripes with a strong reflective appearance. Pavement repairs were performed to prevent the pavement from deteriorating. The waveform at the repair location would show a highlight feature compared to its nearby location due to the difference in repair materials and raw materials. Voids refers to the phenomenon of bottom cavity of the pavement structure layer caused by settlement and distortion between the old and new asphalt pavement. It is often depicted as a blocky highlight on horizontal radar images. Due to environmental constraints during construction, there was poor interlayer bonding at the interlayer. It appears as a highlight anomaly on horizontal radar images. Mixture segregation is the result of an uneven paving process due to poorly mixed materials or uncontrolled temperatures during production, mixing and paving. This mixture segregation is often depicted as a messy highlight feature on horizontal radar images due to the large number of voids. Typical pavement structural distresses (cracks, repairs, voids, poor interlayer bonding, and mixture segregation) were shown in Figure 4. Different distresses were represented clearly in the horizontal radar images. Therefore, horizontal radar images provide a unique perspective for pavement distress detection and can be used as an aid for pavement inspection. However, the horizontal radar images still had two problems. First, the waveform characteristics were complex because the same pavement distress may have shown different characteristics at different depths. For example, voids may be shown as black or white highlights at different depth locations. In addition, due to the complex and various internal structures of pavements, except for typical pavement distresses, there were also a large number of noise waveforms that derive from many real-world factors, such as antenna vibration and background noise. These anomalous waveforms significantly affect distress detection performances. Therefore, a robust and accurate method was needed to process the horizontal radar images.

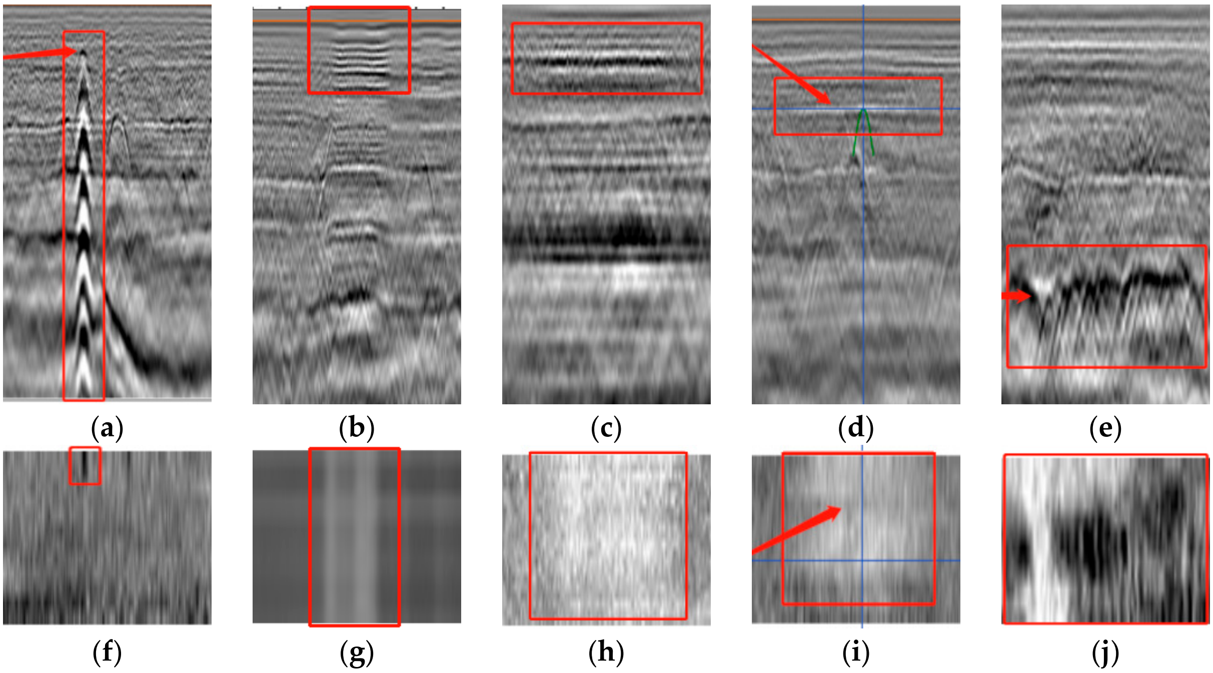


Figure 4. Radar waveform features of different pavement distress: longitudinal radar images of (**a**) crack, (**b**) repair, (**c**) voids, (**d**) poor interlayer bonding, and (**e**) mixture segregation; horizontal radar images of (**f**) crack, (**g**) repair, (**h**) voids, (**i**) poor interlayer bonding, and (**j**) mixture segregation.

2.2. Proposed Localization Model

2.2.1. Distress Localization Dataset

It is necessary to generate a dataset of 3D GPR horizontal images before building a deep learning model for anomalous waveform localization. The collected 3688 horizontal radar images were divided into training, validation, and test sets, corresponding to 2470, 618, and 600 images, respectively. This study then made block-level labels for each image. Labellmg software in the Python environment was used to label the anomalous waveform areas in an image, such as the example shown in Figure 5a. A bounding box indicated the location of an anomalous waveform area. Table 1 presents the total number of bounding boxes in the training, validation, and test sets.

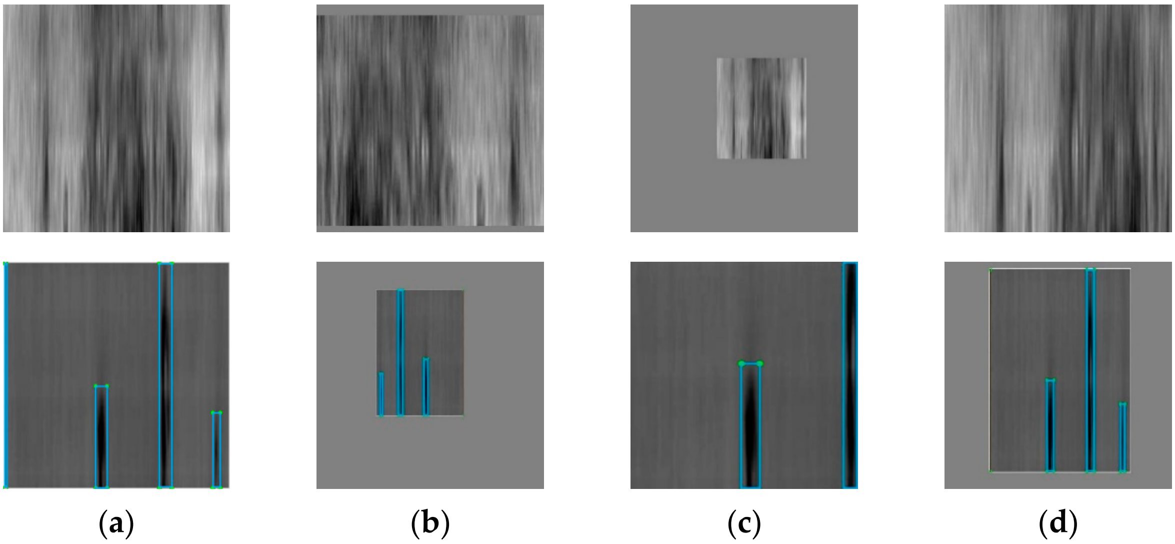


Figure 5. Data enhancement: (**a**) original images and (**b**–**d**) images of random data enhancement.

Table 1. Number of training, validation, and test samples.

Type	Training and Validation Samples		Test Samples	
	Number of Objects	Number of Images	Number of Objects	Number of Images
Anomalous waveform	7453	3088	1592	600

In order to alleviate the overfitting and improve the accuracy of the location model, random data augmentation was performed on the dataset. In this study, random data enhancement had the following three main points:

(1) Randomly crop an image;
(2) Randomly resize an image in terms of its length and width;
(3) Randomly distort the color gamut of an image.

The three data enhancement methods were performed simultaneously, such as the examples shown in Figure 5b–d. The randomly enhanced image was filled with zeros in the remaining positions.

2.2.2. Structure of CP-YOLOX

In order to localize anomalous waveforms, a CSP PAN YOLOX (CP-YOLOX) model has been proposed, which is modified from the original YOLOX [32]. The proposed model can be divided into three parts: CSPDarkNet-SPP as Backbone, Path Aggregation Network (PAN) as Neck, and Decoupled Head as Prediction [33–35]. The architecture of the model is shown in Figure 6.

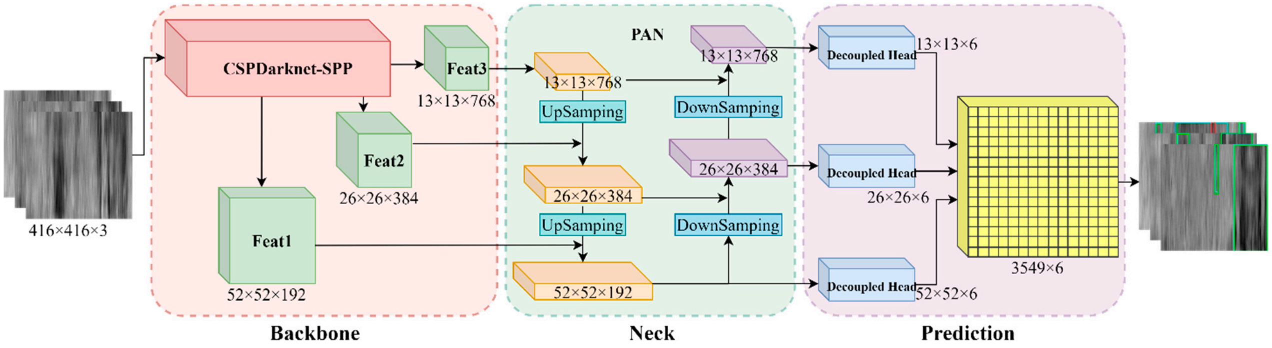


Figure 6. Architecture of a CP-YOLOX network.

(1) Backbone

The backbone extracts high-dimensional features from inputs using several convolutional and pooling layers. In this study, the CSPDarkNet-SPP network was used as the backbone of CP-YOLOX. The architecture and parameters of the network were shown in Figure 7 and Table 2, respectively. The CSPDarkNet-SPP network incorporates Focus, Cross Stage Partial (CSP), and Spatial Pyramid Pooling (SPP) to reduce the floating operations, increase perceptual field, and enhance feature extraction efficiency.

In the CSPDarkNet-SPP network, an input image passes through the Focus structure, which concentrates the information in the length and width dimensions into the channel dimension. This operation obtained two-fold downsampling features without information loss. The features were then compressed and extracted using a Convolutional Batch SiLU (CBS) structure and four CSP layers. With the last CSP layer, an SPP operation was added to increase the perceptual field of the network by utilizing maximum pooling layers with different sizes, which improved feature extraction efficiency.

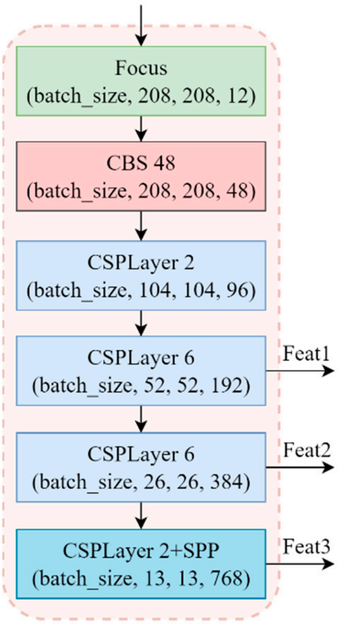

Focus
(batch_size, 208, 208, 12)
CBS 48
(batch_size, 208, 208, 48)
CSPLayer 2
(batch_size, 104, 104, 96)
CSPLayer 6
(batch_size, 52, 52, 192)
Feat1
CSPLayer 6
(batch_size, 26, 26, 384)
Feat2
CSPLayer 2+SPP
(batch_size, 13, 13, 768)
Feat3

Figure 7. Flow chart of structure of CSPDarkNet-SPP.

Table 2. Main parameters of CSPDarkNet-SPP.

Layers		Kernel Size and Number		Input Size	Output Size	Stride
Focus		3 × 4 = 12		416 × 416	208 × 208	-
CBS3		3 × 3, 48		208 × 208	208 × 208	1
CSPLayer 2	CBS3	3 × 3, 96		208 × 208	104 × 104	2
	CBS1	1 × 1, 48	1 × 1, 48	104 × 104	104 × 104	1
	Res unit	-	$\left\{\begin{matrix}1 \times 1, & 48\\ 3 \times 3, & 48\end{matrix}\right\} \times 2$	104 × 104	104 × 104	1
	CBS1	1 × 1, 96		104 × 104	104 × 104	1
CSPLayer 6	CBS3	3 × 3, 192		104 × 104	52 × 52	2
	CBS1	1 × 1, 96	1 × 1, 96	52 × 52	52 × 52	1
	Res unit	-	$\left\{\begin{matrix}1 \times 1, & 96\\ 3 \times 3, & 96\end{matrix}\right\} \times 6$	52 × 52	52 × 52	1
	CBS1	1 × 1, 192		52 × 52	52 × 52	1
CSPLayer 6	CBS3	3 × 3, 384		52 × 52	26 × 26	2
	CBS1	1 × 1, 192	1 × 1, 192	26 × 26	26 × 26	1
	Res unit	-	$\left\{\begin{matrix}1 \times 1, & 192\\ 3 \times 3, & 192\end{matrix}\right\} \times 6$	26 × 26	26 × 26	1
	CBS1	1 × 1, 384		26 × 26	26 × 26	1
CSPLayer 2 + SPP	CBS3	3 × 3, 768		26 × 26	13 × 13	2
	SPP	1 × 1, 384 1 × 1, 5 × 5, 9 × 9, 13 × 13, 384 1 × 1, 768		13 × 13	13 × 13	1
	CBS1	1 × 1, 384	1 × 1, 384	13 × 13	13 × 13	1
	Res unit	-	$\left\{\begin{matrix}1 \times 1, & 384\\ 3 \times 3, & 384\end{matrix}\right\} \times 2$	13 × 13	13 × 13	1
	CBS1	1 × 1, 768		13 × 13	13 × 13	1

A CSP layer divides its input into two parts by performing a 1 × 1 convolution operation in which one part passes through multiple residual modules to extract features, and then the features are concatenated with another part. The residual module consists of two CBS and one residual edge. The first CBS was used for channel adjustment to reduce the computation cost, while the second one was used for feature extraction. The CSP layer reduces the calculation of the network as well as the redundancy of the network information while ensuring efficient feature extraction. Based on the proposed model, the number of CSP layers represents the number of residual module cycles, and the cycle numbers of the four residual modules were 2, 6, 6, and 2.

The CBS consists of a convolutional layer, a batch normalization (BN) layer, and a SiLU activation function [36]. There were two sizes of convolutional kernels in the convolutional layer. The 1 × 1 convolutional kernels were used for channel adjustment to increase the nonlinear fitting ability, while the 3 × 3 convolutional kernels were used for feature extraction. The BN layer normalized the outputs of a convolutional layer to accelerate network convergence and alleviate overfitting. The SiLU activation function was a comprehensive version of the Sigmoid and ReLU functions and is defined as follows.

$$\mathrm{f(x)= x{\cdot}sigmoid(x)}, \tag{1}$$

$$\mathrm{sigmoid(x)} = \frac{1}{1+\mathrm{e}^{-\mathrm{x}}}. \tag{2}$$

Its smooth and non-monotonic characteristics work well on deep neural networks. The function is shown in Figure 8.

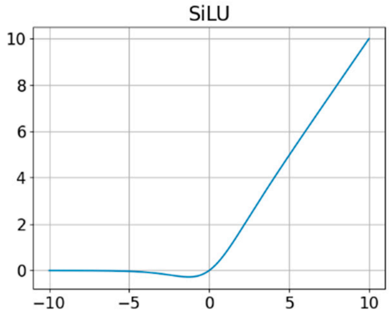


Figure 8. Activation function of SiLU.

(2) Neck

Neck was an architecture between backbone and prediction, which fused the multi-dimension features from the backbone network and imported them into the prediction architecture, as shown in Figure 9. Three different features (Feat1, Feat2, and Feat3) from the CSPDarkNet-SPP backbone were imported into the neck architecture. The low-dimensional feature Feat1 contained strong local information, while the high-dimensional feature Feat3 included distress-semantics information. In this study, a Path Aggregation Network (PAN) was used as the neck architecture to fused these features, as shown in Figure 9, for which its parameters are shown in Table 3. In PAN, multi-dimension features were fused by sequentially upsampling and using convolutional layers.

(3) Prediction

A decoupled head was proposed as the prediction architecture, using neck architecture features as inputs. The decoupled head decoupled the classification and regression tasks, as shown in Figure 10. The regression task aimed to predict the bounding box of each anomalous waveform area, while the classification task classified each anomalous waveform area into one of the distress classes. The decoupled head network used a CBS layer to generate two feature vectors based on the input features. The first vector was then passed through two CBS layers to produce a classification prediction vector, and the second vector was also passed through two CBS layers to produce a regression prediction vector. After concatenating the two vectors, the dimension of the concatenated vector was adjusted by using a 1 × 1 convolution layer. The final output of the model was a feature vector with the size of 1 × 6, where 6 can be divided into 1 + 1 + 4, corresponding to anomalous waveform or not, confidence, and location of the prediction boxes.

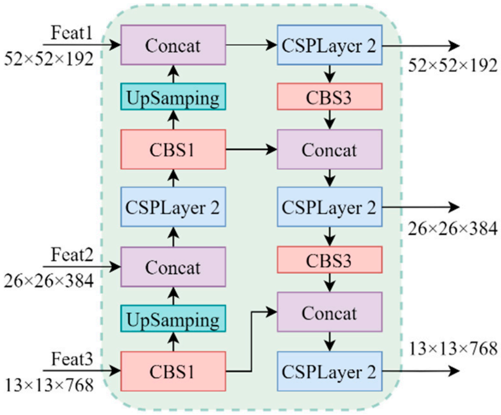


Figure 9. Flow chart of structure of PAN.

Table 3. Structure parameters of PAN.

Layers	Kernel Size	Channel	Input Size	Output Size	Stride
CBS1	1 × 1	384	13 × 13	13 × 13	1
UpSamping	-	384	13 × 13	26 × 26	-
Concat	-	768	26 × 26	26 × 26	-
CSPLayer 2	(1 × 1, 3 × 3) × 2	384	26 × 26	26 × 26	1
CBS1	1 × 1	192	26 × 26	26 × 26	1
UpSamping	-	192	26 × 26	52 × 52	-
Concat	-	384	52 × 52	52 × 52	-
CSPLayer 2	(1 × 1, 3 × 3) × 2	192	52 × 52	52 × 52	1
CBS3	3 × 3	192	52 × 52	26 × 26	2
Concat	-	384	26 × 26	26 × 26	-
CSPLayer 2	(1 × 1, 3 × 3) ×2	384	26 × 26	26 × 26	1
CBS3	3 × 3	384	26 × 26	13 × 13	2
Concat	-	768	13 × 13	13 × 13	-
CSPLayer 2	(1 × 1, 3 × 3) ×2	768	13 × 13	13 × 13	1

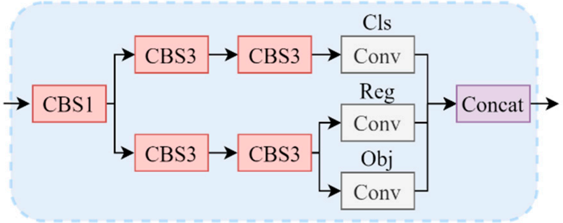


Figure 10. Flow chart of decoupled head.

The proposed model adopted the anchor-free strategy, which did not require predefining anchor sizes. In addition, this study also adopted the simplify optimal transport assignment (SimOTA) [37] strategy to dynamically matched positive samples for different distresses. Anchor-free and SimOTA were used in conjunction to reduce the complexity of the detection head and to increase the robustness of the model.

2.3. Proposed Classification Model

2.3.1. Distress Classification Dataset

Even though the proposed location model can determine anomalous waveform areas, it cannot easily determine the categories of these areas. For example, there was no easy method to determine if the anomalous waveform area was caused by cracks or uneven settlement. The main reason for this was that few studies had demonstrated the ability of horizontal GPR images to represent different types of pavement distress. However, the class-membership strategies of anomalous waveform images affected distress classification. An over-fine classification strategy had the risk of misclassification, despite sometimes providing a precise decision; an over-coarse strategy cannot provide an informative decision. In this study, three class-membership strategies, as shown in Table 4, were proposed to demonstrate the capacity of horizontal GPR images to represent different pavement distresses. Among them, HD1 (Horizontal Distress 1) indicated that all anomalous waveforms were considered cracks; HD16 indicated that the anomalous waveforms could represent cracks or noises; HD45 indicated that the anomalous waveforms may be poor interlayer bonding or mixture segregation; HD6 indicated that the anomalous waveforms were background noises.

Table 4. Class-membership strategies of anomalous waveform in horizontal radar image.

Classification Method	Category					Category Number
1	HD1	HD16	HD45	HD6	HD126	5
2	HD1	HD16	HD45	HD6		4
3	HD1	HD45		HD6		3

As shown in Table 1, 9045 anomalous waveforms were manually labeled, which were then intercepted from the horizontal radar images to create a classification dataset. Figure 11 illustrates how the image interception method works. The anomalous waveform areas were cropped from the radar images based on the position coordinates. The intercepted anomalous waveform images had different shapes, while a classification model was required to had the inputs with a fixed size. Thus, the cropped anomalous waveform images were resized by adding zeros to the missing parts. Finally, these filled images were assigned to one of the distress categories using different class-membership strategies. Therefore, three datasets with the same images but different labels were generated, as shown in Table 5.

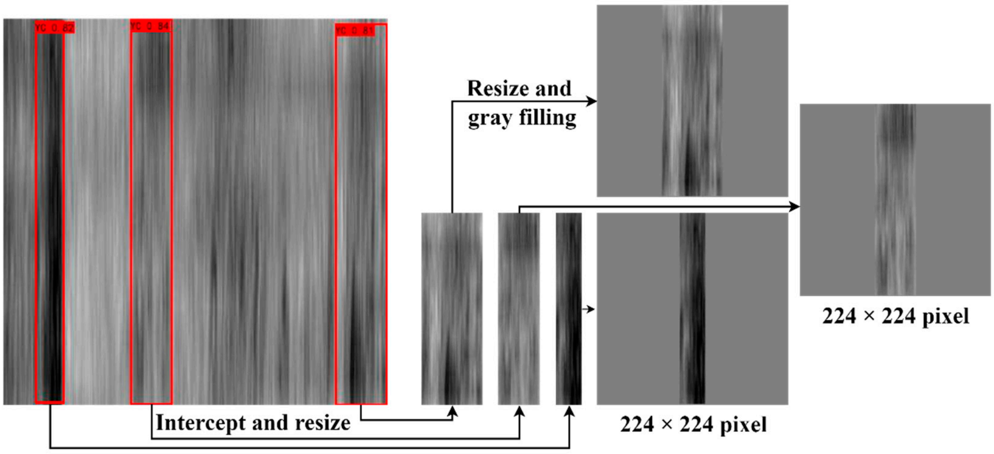
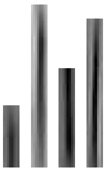
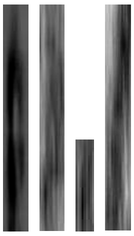
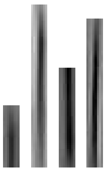
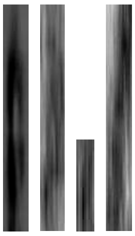
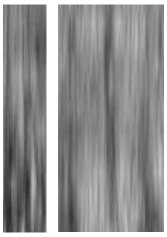
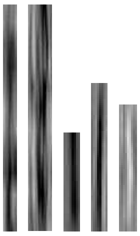
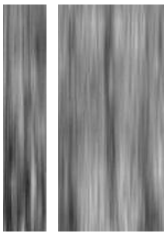


Figure 11. Process of intercepting anomalous waveform areas from a horizontal radar image. The red boxes are manual labeled.

In all three datasets, the number of samples in different categories was balanced. In the 5-category dataset, all 395 images were selected for HD126, and 400 images were randomly selected for each of the other categories to create the dataset, and the total number of datasets was 1995. In the 4-category dataset, the number of randomly selected images were 1072, 1100, 1100, and 1100, respectively, for a total of 4372 images. In the 3-class dataset, the numbers of randomly selected images were 1600, 1366, and 1600, and the total number of images was 4566. For the three classification methods, the ratio of the training and validation sets was 8:2, and the test set consisted of all the anomalous waveforms in 600 horizontal radar images with a total of 1592 images.

Table 5. Characteristics of three classification datasets.

Category		**HD1**	**HD16**	**HD45**	**HD6**	**HD126**
5 Categories						
Training	Samples	1072	2172	1366	2448	395
	Data set	400	400	400	400	395
Testing	Data set	178	467	120	748	79
Category		**HD1**	**HD16**	**HD45**	**HD6**	
4 Categories						
Training	Samples	1072	2172	1366	2843	
	Data set	1072	1100	1100	1100	
Testing	Data set	178	467	120	827	
Category		**HD1**		**HD45**	**HD6**	
3 Categories						
Training	Samples	3244		1366	2843	
	Data set	1600		1366	1600	
Testing	Data set	645		120	827	

2.3.2. Structure of SViT

In this study, Simplify Vision Transformers (SViTs) were used to perform the classification task, which was a simplified version of Vision Transformer (ViT). An SViT can be separated into three modules, namely Embedding, Transformer Encoder, and MLP Head, as shown in Figure 12.

(1) Embedding

As shown in Figure 12, the embedding module converted the input image into the data format required by the transformer encoder module. The input image was split into

196 blocks of 16 × 16 pixels in the embedding module. The blocks were then flattened into one-dimensional vectors by a flattening layer to obtain 196 vector sequences with 256 dimensions. Finally, the position information is embedded in each vector sequence.

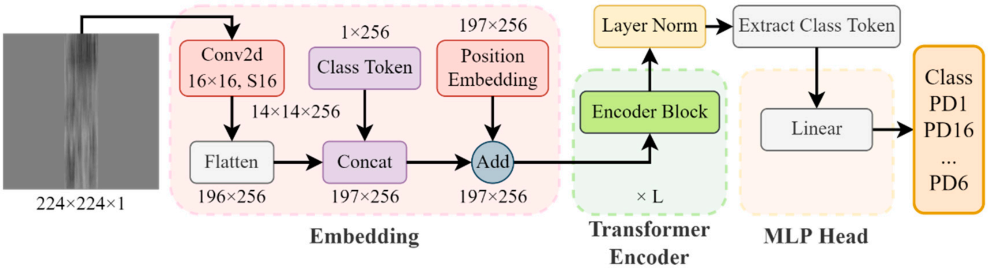


Figure 12. Overall flow chart of SVIT.

(2) Transformer Encoder

The transformer encoder module, as the core of the SViT model, extracted features from its input vector sequences. The architecture of the transformer encoder is shown in Figure 13. The transformer encoder consisted of two residual structures. The first residual structure was multi-head attention, while its second counterpart was multi-layer perceptron (MLP). Multi-head attention divided the input data into multiple parts to perform attention computation. In this study, the number of heads in multi-head attention was four, i.e., the vector sequence of 197 × 256 was divided into four subsequences of 197 × 64 and imported into the attention module. In the attention module, the vector sequence was further split into Query (Q), Key (K), and Value (V), and the results of attention and multi-head attention were computed as follows.

$$\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V})=\mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d}_{\mathrm{k}}}}\right)\cdot\mathrm{V}, \quad (3)$$

$$\mathrm{MultiHead}(\mathrm{Q},\mathrm{K},\mathrm{V})=\mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_{\mathrm{h}})\cdot\mathrm{W}^{\mathrm{O}}, \quad (4)$$

$$\mathrm{head}_{\mathrm{i}}=\mathrm{Attention}\left(\mathrm{QW}_{\mathrm{i}}^{\mathrm{Q}},\mathrm{KW}_{\mathrm{i}}^{\mathrm{K}},\mathrm{VW}_{\mathrm{i}}^{\mathrm{V}}\right). \quad (5)$$

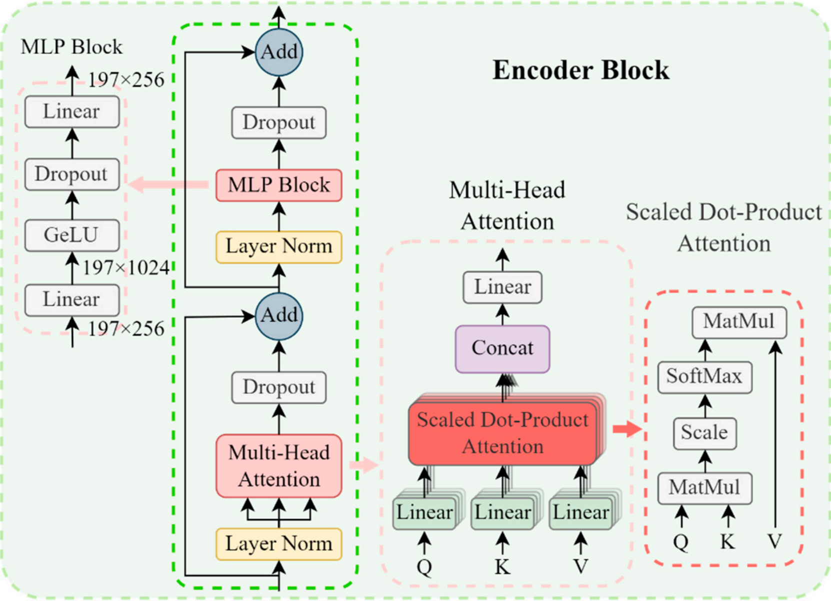


Figure 13. Schematic diagram of structure of transformer encoder.

The number of neurons in the first fully connected layer of the MLP module was 1024, which was four times the size of the inputs, and the number of neurons in the second fully connected layer was 256, which was the same size as the inputs. The coefficient of the dropout was 0.1, and the GELU activation function [38] was adopted as follows.

$$\mathrm{GELU}(x) = xP(X \leq x) = x\Phi(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}\left(x/\sqrt{2}\right)\right], \quad (6)$$

$$\mathrm{GELU}(x) \approx 0.5x\left(1 + \tanh\left[\sqrt{2/\pi}\left(x + 0.0447x^3\right)\right]\right). \quad (7)$$

The function of GELU was shown in Figure 14.

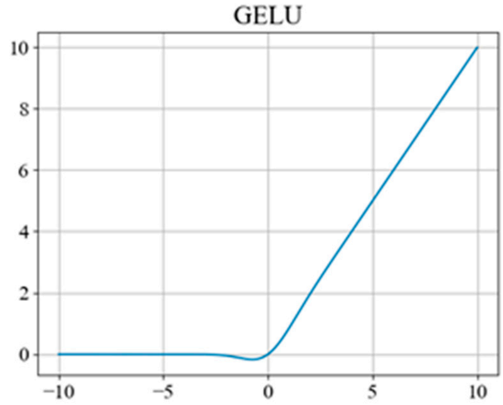


Figure 14. Activation function of GELU.

The transformer encoder module was repeated several times to extract multi-level features. In this study, five transformer encoders with 3, 6, 9, 12, and 15 cycles were used, and the corresponding models were called SViT-3, SViT-6, SViT-9, SViT-12, and SViT-15, as shown in Table 6.

Table 6. Detailed parameters of the 5 SViT models.

Model	Patch Pixel	Layer	Hidden Size	MLP Size	Heads	Params	FLOPs
SViT-3	16 × 16	3	256	1024	4	2.49 M	1.09 G
SViT-6	16 × 16	6	256	1024	4	4.86 M	2.15 G
SViT-9	16 × 16	9	256	1024	4	7.23 M	3.21 G
SViT-12	16 × 16	12	256	1024	4	9.6 M	4.27 G
SViT-15	16 × 16	15	256	1024	4	11.96 M	5.33 G

(3) MLP Head

Based on the features from the transformer encoder, the MLP Head classified the input image into one possible class. After completing feature extraction by the transformer encoder module, the output shape was 197 × 256, which contained the class token embedded in the embedding module. The class token learned the features of anomalous waveforms in the transformer encoder and extracted the class token individually into the MLP head, which enabled the classification of anomalous waveforms.

2.4. Learning Strategy

(1) Overall

The weights and bias of the location and classification models were updated by the backward propagation algorithm [14] as follows:

$$\mathrm{W_{ij}(q+1)= W_{ij}(q)} - \eta\frac{\partial \mathrm{L}}{\partial \mathrm{W_{ij}(q)}}, \tag{8}$$

$$\mathrm{b_j(q+1)= b_j(q)} - \eta\frac{\partial \mathrm{L}}{\partial \mathrm{b_j(q)}}, \tag{9}$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{W_{ij}(q)}}$ and $\frac{\partial \mathrm{L}}{\partial \mathrm{b_j(q)}}$ were the gradients of the loss function with respect to weight $\mathrm{W_{ij}(q)}$ and bias $\mathrm{b_j(q)}$, and η was the learning rate.

The convergence of the model can be accelerated by using different learning rates during different training stages. In this study, an exponential descent function was used to represent the learning rate:

$$lr = lr_b \times \gamma^{epoch}, \tag{10}$$

where lr_b was the base learning rate; γ was the coefficient of learning rate decay.

The location and classification models were trained in TensorFlow 2.5 framework. The training device was a cloud server with AMD EPYC 7302 CPU, 64 G RAM, and NVIDIA GeForce RTX 3090 GPU with 24 GB memory. The testing device was a laptop with Intel i7-9750H CPU, 16 G RAM, and NVIDIA GeForce GTX 1650 GPU with 4 GB memory.

(2) Training of CP-YOLOX

In the learning strategy, three different loss functions were combined in the CP-YOLOX model to compute the gaps between predicted and target information, including the decision of anomalous waveform or not, confidence, and location of the prediction boxes. The loss of category and confidence was defined by the cross-entropy loss function, and the loss of box position was computed by the CIoU loss function [39] as follows:

$$\mathrm{L_{CIoU}}= 1 - \mathrm{IoU}+\frac{\rho^2\left(b,\ b^{gt}\right)}{c^2}+\alpha v, \tag{11}$$

$$\alpha =\frac{v}{(1-\mathrm{IoU})+v}, \tag{12}$$

$$v =\frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w^p}{h^p}\right)^2, \tag{13}$$

where IoU was the intersection over union between the prediction and truth boxes; b and b^{gt} represent the center points of the prediction and truth boxes, respectively; ρ was the Euclidean distance between the two center points; c represents the diagonal distance that can contain the minimum closed area of both the prediction and truth boxes; v represents the similarity of the aspect ratio; w^{gt}, h^{gt}, w^p, and h^p were the length and width of the truth and prediction boxes; α was the weight parameter.

The proposed CP-YOLOX model was trained in three stages, and each stage had 100 epochs. The values of lr_b in the three stages were 1×10^{-3}, 1×10^{-4}, and 5×10^{-5}, while γ was 0.97, as shown in Figure 15a. The Adam optimizer [40] was used for training, and the parameters of the optimizer were 0.9 and 0.999, as shown in Table 7. The training of the model was terminated when the loss value of the validation set could no longer be reduced after 20 epochs.

Table 7. The detailed parameters of model training.

Model	Learning Rate Decay Coefficient	Batch Size	Epoch	Base Learning Rate	Optimizer
CP-YOLOX	γ = 0.97	16	100 100 100	1×10^{-3} 1×10^{-4} 5×10^{-5}	Adam
SViT	γ = 0.97	32	20	1×10^{-4}	Adam

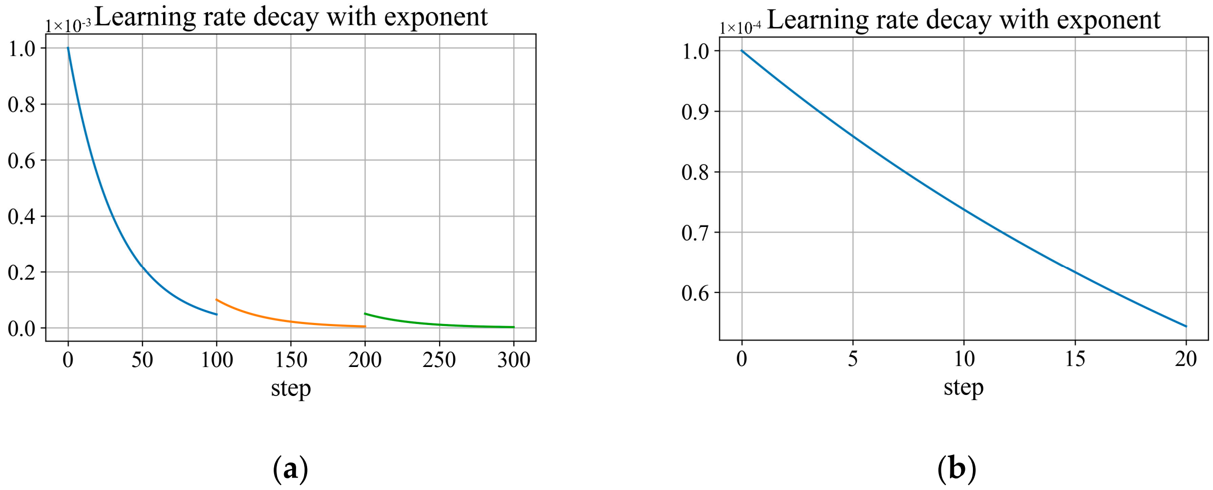


Figure 15. Learning rate decline curve: (**a**) three stages of training for the CP-YOLOX; (**b**) training for the SViT.

(3) Training of SViT

SViT models use cross-entropy loss functions to compute the gap between prediction and target classes. The models also adopted Equation (10) to represent the learning rate. The models were trained for 20 epochs with lr_b of 1×10^{-4} and γ of 0.97, and the same Adam optimizer was used for training, as shown in Figure 15b and Table 7. The training of the model was terminated when the loss value of the validation set could no longer be reduced after 10 epochs.

3. Results and Discussion

3.1. Analysis of Localization Results

The CP-YOLOX model achieved the lowest loss value of 3.060 in the validation set at the 258th epoch. After that, the training continued for 20 epochs and the loss value stopped decreasing. The loss curve of the model was shown in Figure 16.

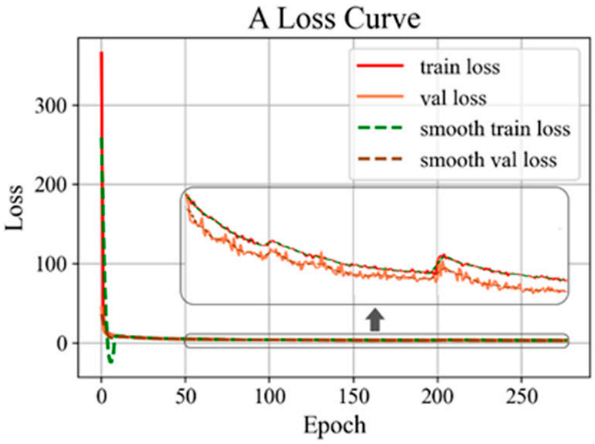
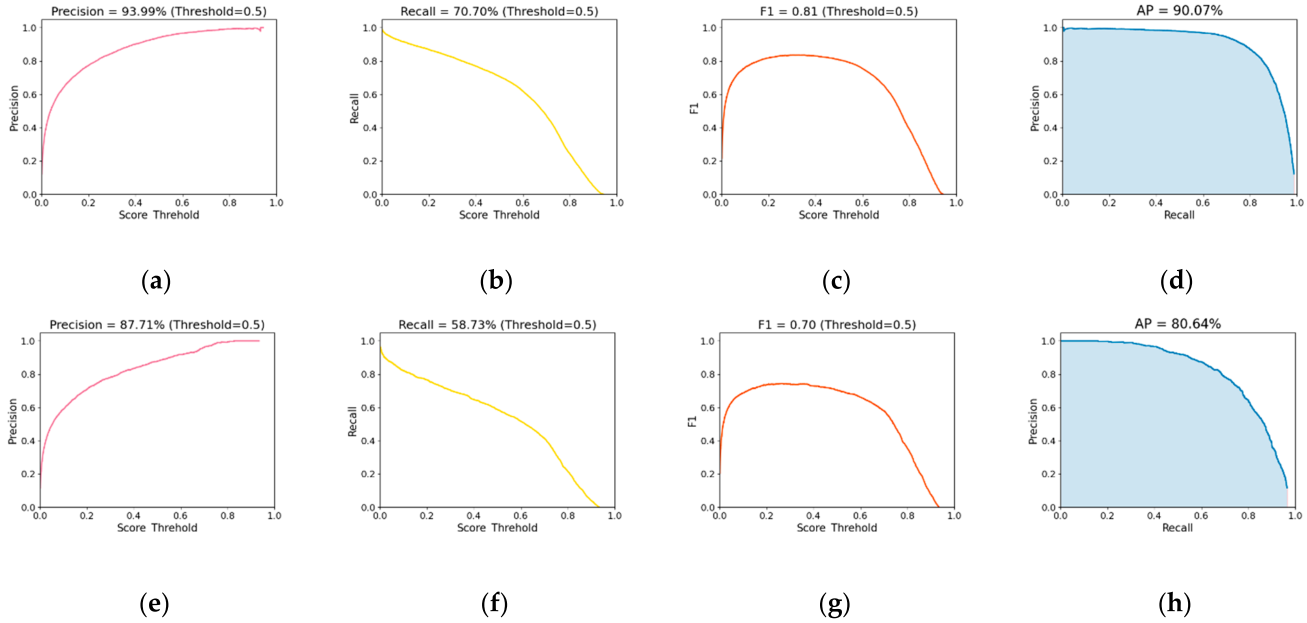


Figure 16. Decline curves of loss values in training and validation sets.

Five metrics were used to evaluate the well-trained localization model, Precision, Recall, F1 score, and mean average precision (mAP) [14]. The results were shown in Figure 17, which were obtained at a confidence threshold of 0.5. In the test set, CP-YOLOX model achieved 87.71%, 58.73%, and 80.64% for precision, recall, and mAP, respectively. The model required 29.8 ms to processed one image, and the frame per second reached 33.57 images/s. The evaluation results were shown in Table 8.

Table 8. Results of the test set.

Model	Size (Pixels)	Precision (%)	Recall (%)	F1 Score	mAP (%)	FPS (Images/s)	Inference Time (ms)
CP-YOLOX	416 × 416	87.71	58.73	0.70	80.64	33.57	29.8

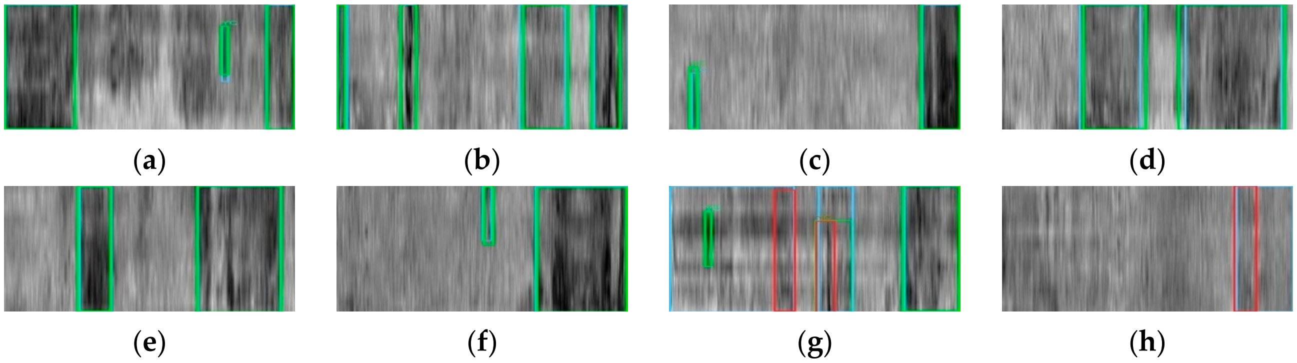


(**a**) (**b**) (**c**) (**d**)

(**e**) (**f**) (**g**) (**h**)

Figure 17. Curves of evaluation metrics for the training and test sets: (**a**–**d**) the Precision, Recall, F1 score, and AP values of the training set, respectively; (**e**–**h**) the Precision, Recall, F1 score, and AP values of the testing set, respectively.

The well-trained model was used to localized the anomalous waveforms in the horizontal radar images with a confidence threshold of 0.5. Some results were shown in Figure 18. Prediction and truth boxes were plotted into the radar images, with green and red boxes being prediction boxes and blue boxes being truth boxes. The prediction boxes display green with IoU ≥ 0.5; otherwise, it shows red. Prediction result and confidence levels were plotted into radar image, as shown in Figure 19. As a general rule, the proposed model was able to identify the anomalous waveform areas in horizontal radar images.



(**a**) (**b**) (**c**) (**d**)

(**e**) (**f**) (**g**) (**h**)

Figure 18. The prediction results of CP-YOLOX on the test set: (**a**–**h**) Predictions based on eight randomly selected images.

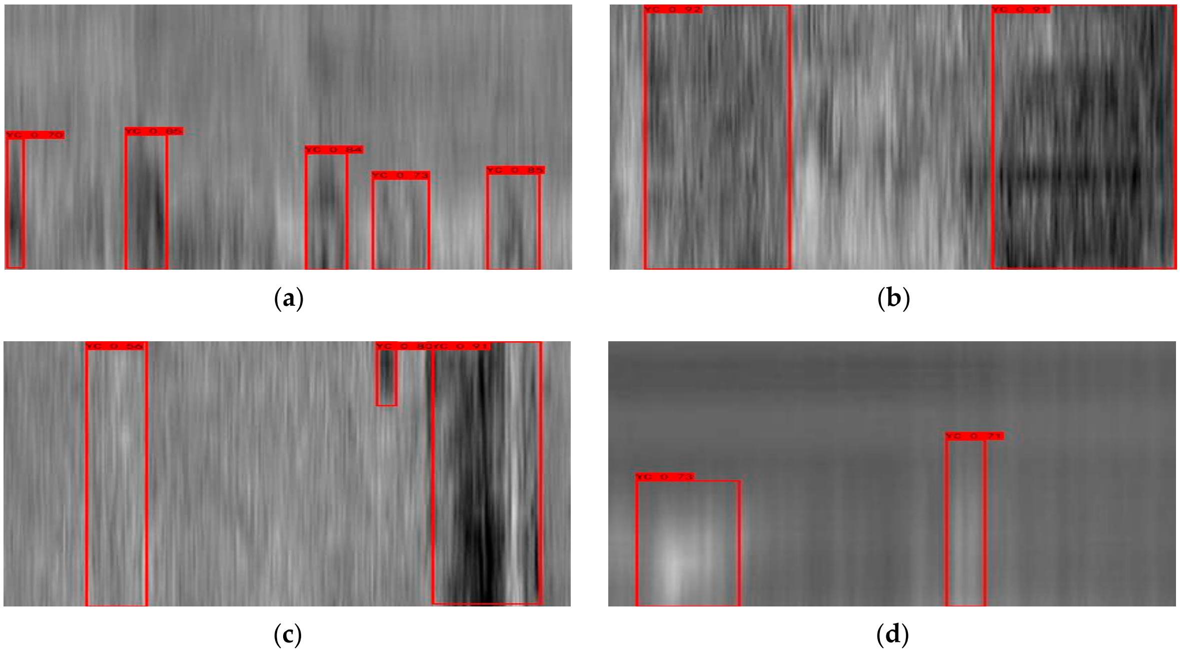
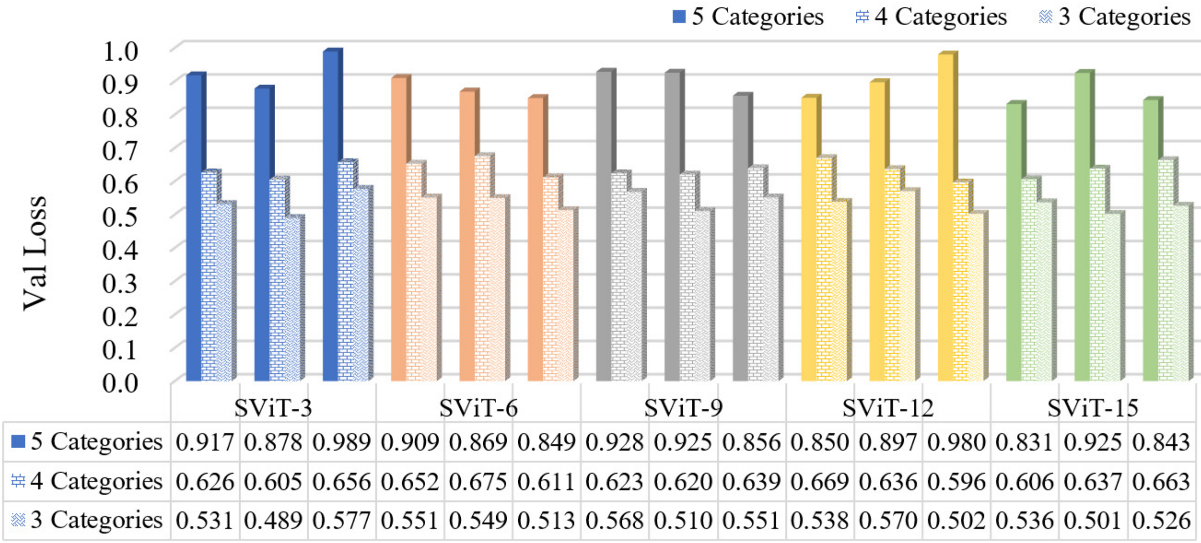


Figure 19. The prediction results of CP-YOLOX on the original image: (**a**–**d**) Predictions based on four randomly selected images.

3.2. Analysis of Classification Results

3.2.1. Results of Training and Testing

In order to reduce the error, each of the five SVIT models was trained three times to find optimal one using the validation set. The five models were trained three times for each of the three data sets in Table 5. A total of 45 training results were obtained, as shown in Figure 20.



	SViT-3			SViT-6			SViT-9			SViT-12			SViT-15		
■ 5 Categories	0.917	0.878	0.989	0.909	0.869	0.849	0.928	0.925	0.856	0.850	0.897	0.980	0.831	0.925	0.843
▨ 4 Categories	0.626	0.605	0.656	0.652	0.675	0.611	0.623	0.620	0.639	0.669	0.636	0.596	0.606	0.637	0.663
▨ 3 Categories	0.531	0.489	0.577	0.551	0.549	0.513	0.568	0.510	0.551	0.538	0.570	0.502	0.536	0.501	0.526

Figure 20. Loss values of validation sets on different models under three data sets.

The weights with the smallest loss in the validation set among the three results were selected, and the corresponding loss value decline curves were shown in Table 9.

Table 9. Loss value decline curve for 15 best weights.

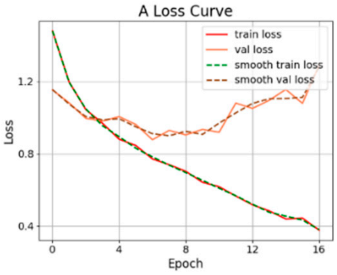
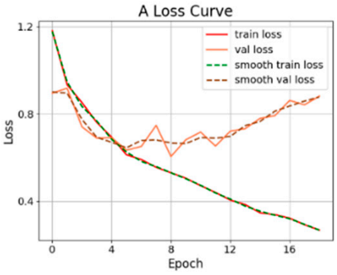
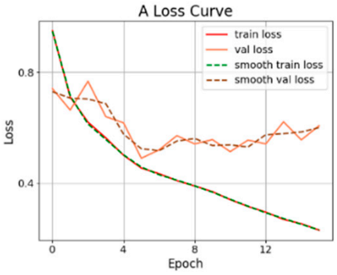
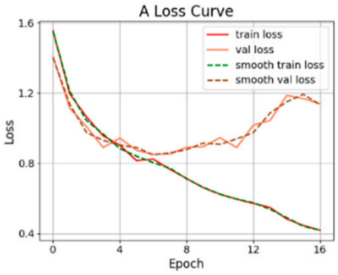
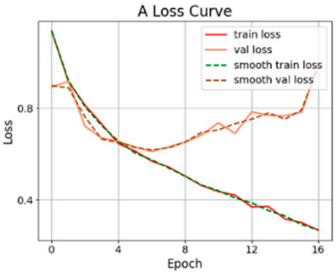
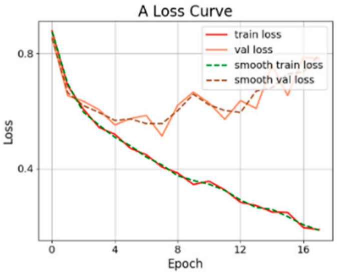
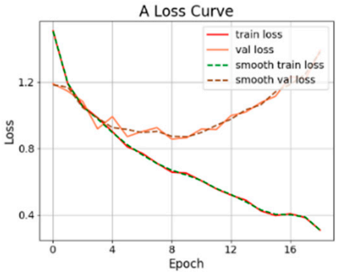
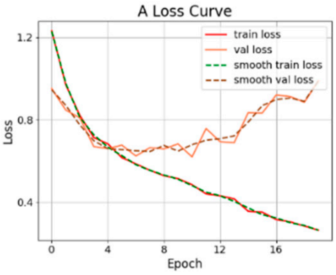
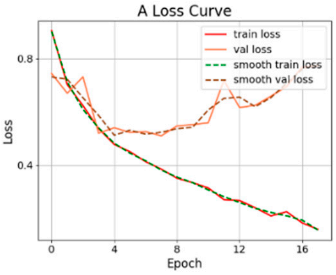
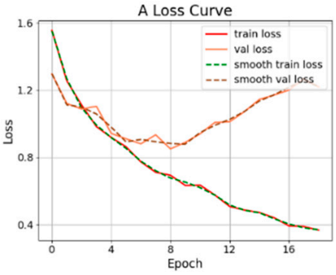
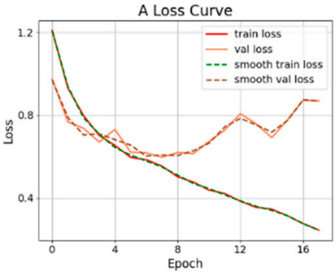
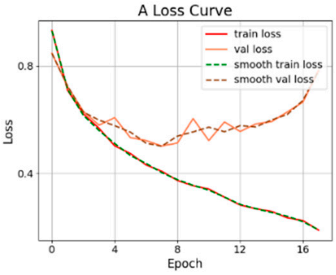
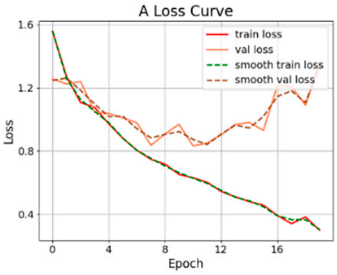
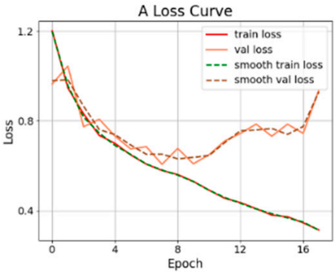
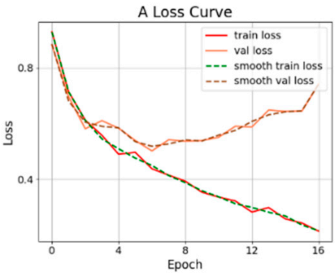
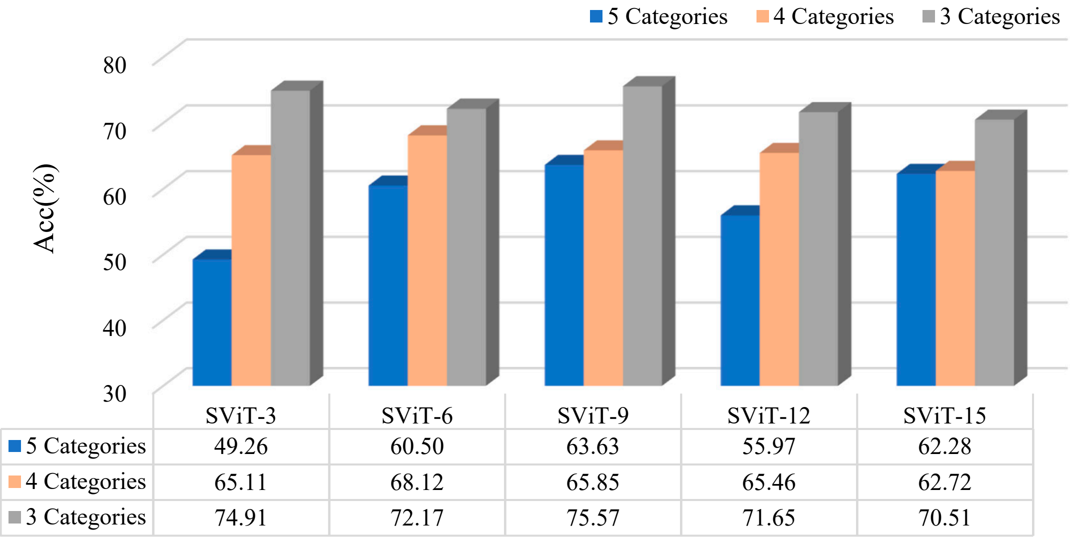
Model	5 Categories Dataset	4 Categories Dataset	3 Categories Dataset
SViT-3			
SViT-6			
SViT-9			
SViT-12			
SViT-15			

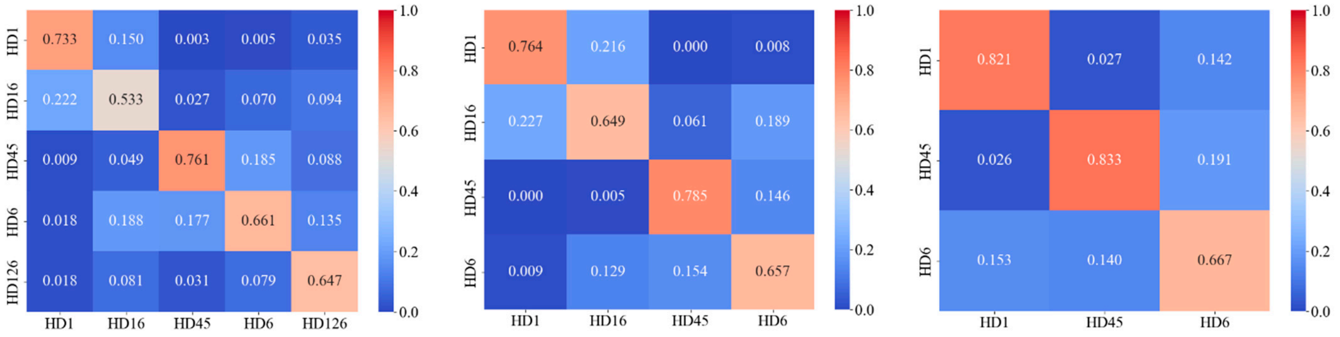
Figure 21 presents the accuracies of 15 optimal models in the test sets. On the SViT-9 model, the 5-categories and 3-categories test sets had the highest accuracy with 63.63% and 75.57%, respectively. On the SViT-6 model, the 4-categories test sets had the highest accuracy with 68.12%. The SViT model predicted the highest accuracy of 75.57% for the 3-categories test sets, corresponding to the categories of crack, poor interlayer bonding, and mixture segregation.



	SViT-3	SViT-6	SViT-9	SViT-12	SViT-15
■ 5 Categories	49.26	60.50	63.63	55.97	62.28
■ 4 Categories	65.11	68.12	65.85	65.46	62.72
■ 3 Categories	74.91	72.17	75.57	71.65	70.51

Figure 21. Accuracy of different models on 3 test sets.

Based on the results in Figure 21, the SViT-6 model was tested in the 4-categories test set, while the SViT-9 models were tested in the 5- and 3-categories test sets. The results are shown in Figure 22. In the three class-membership strategies, the models were the most accurate in classifying pavement distresses without background noises, with HD1 accuracies of 73.3%, 76.4%, and 82.1% and HD45 accuracies of 76.1%, 78.5%, 83.3%, respectively. The models had poor accuracies in predicting distress containing background noises, in which all accuracies were lower than 70%. In horizontal radar images, SViT model was capable of detecting cracks, poor interlayer bonding, and mixture segregation distress. The model was tested well on a 3-category dataset, but the best distress classification strategy needs to be determined along with longitudinal radar images. The future study should focus on how to suppress the noise waveform and on more detailed analysis of the disturbance waveform.



	HD1	HD16	HD45	HD6	HD126
HD1	0.733	0.150	0.003	0.005	0.035
HD16	0.222	0.533	0.027	0.070	0.094
HD45	0.009	0.049	0.761	0.185	0.088
HD6	0.018	0.188	0.177	0.661	0.135
HD126	0.018	0.081	0.031	0.079	0.647

(**a**)

	HD1	HD16	HD45	HD6
HD1	0.764	0.216	0.000	0.008
HD16	0.227	0.649	0.061	0.189
HD45	0.000	0.005	0.785	0.146
HD6	0.009	0.129	0.154	0.657

(**b**)

	HD1	HD45	HD6
HD1	0.821	0.027	0.142
HD45	0.026	0.833	0.191
HD6	0.153	0.140	0.667

(**c**)

Figure 22. Accuracy of different disease in 3 test sets: (**a**) 5-categories; (**b**) 4-categories; (**c**) 3-categories.

3.2.2. Prediction Results of Different Models

SViT was a simplified model based on ViT, which was different from the traditional CNN models described in Section 2.3. Different models were compared with the proposed model. Comparison models included ViT and CNN-based MobileNet and ResNet50 [41,42]. The floating-point operations (FLOPs) and accuracies of different models were shown in Figure 23 and Table 10. SViT outperformed ViT and MobileNet models on accuracy, parameter, and FLOPs, even though its accuracy was slightly lower than one of ResNet50 model. As a result, the proposed model had fewer parameters and FLOPs, ensuring a high level of accuracy, which enables it to perform distress classification quickly. This demonstrated the effectiveness of the proposed model in identifying pavement distress.

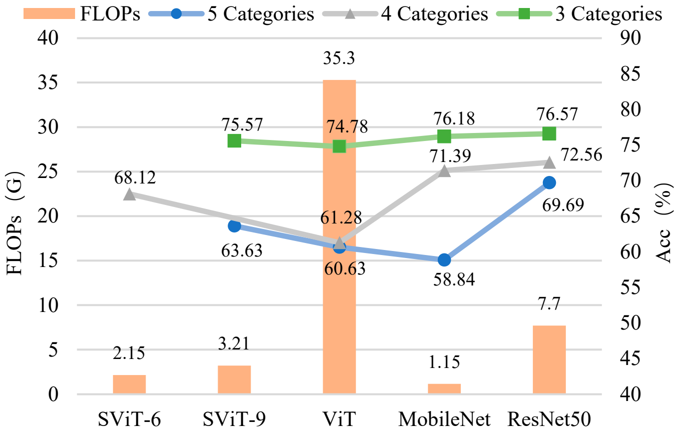


Figure 23. Accuracy and FLOPs of each model under different data sets.

Table 10. Accuracy results of 3 datasets with different models.

Model	Acc (%)			Params	FLOPs
	5 Categories	4 Categories	3 Categories		
SViT-6	—	68.12	—	4.86 M	2.15 G
SViT-9	63.63	—	75.57	7.23 M	3.21 G
ViT	60.63	61.28	74.78	85.8 M	35.3 G
MobileNet	58.84	71.39	76.18	3.23 M	1.15 G
ResNet50	69.69	72.56	76.57	23.6 M	7.7 G

3.2.3. The Influence of Number of Samples on the Model

A ViT model surpassed traditional CNN models in the field of image recognition once given sufficient samples in the learning set [16]. During training, the fitting ability of SViT was excellent, and only a few epochs were required to complete the training. Figure 24 presents the accuracies of different models with different datasets. With more samples, the accuracy of the SViT model was close to that of ResNet50. The accuracy gap between the two models shrank from 6.06% to 1% as the number of single-category samples increased. SViT still had the potential to outperform the ResNet50 model if more samples were available. Therefore, more 3D GPR images should be collected to improve the performance of the proposed model.

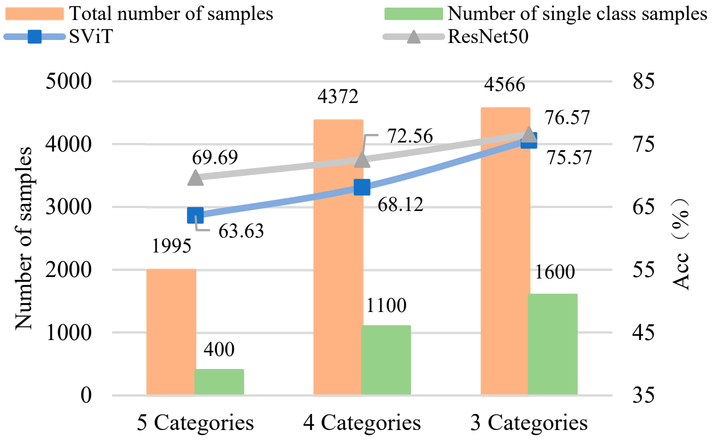


Figure 24. Comparison study of sample number and model accuracy.

4. Conclusions

Pavement distress was detected using deep learning and horizontal GPR images in this study. The anomalous waveform areas in horizontal radar images were located by a CP-YOLOX model. Five SViT models classified anomalous waveforms into one of the possible distress categories. A GPR image dataset collected from China demonstrated the effectiveness of the proposed models. The following conclusions can be drawn.

(1) The proposed CP-YOLOX model could localize anomalous waveforms caused by pavement distresses. With a confidence threshold of 0.5, the CP-YOLOX model localized anomalous waveforms with mAP of 80.64%, Precision of 87.71%, and Recall of 58.73%. The model processed radar images with a speed of 33.57 images/s.

(2) The proposed SViT model was capable of detecting cracks, poor interlayer bonding, and mixture segregation distress in horizontal radar images. For the category without background noise, the model had a high prediction accuracy. Future studies should focus on how to suppress the noise waveform and on more detailed analysis of the disturbance waveform.

(3) The proposed SViT model had fewer parameters and FLOPs, ensuring a high level of accuracy, which enables it to perform distress classification quickly. With the increase in GPR images, the gap between SViT and ResNet50 shrunk from 6.06% to 1%, indicating that more data samples had the potential to improve the performance of SViT. This demonstrated the superiority of the proposed model on the pavement distress classification.

(4) In the three classification datasets, the 3-categories dataset had the highest accuracy, followed by the 4-categories dataset, and the 5-categories dataset had the lowest accuracy. However, the model trained based on the 5-categories dataset provided the most detailed basis for distress classification. Subsequently, we need to combine the horizontal detection results with the longitudinal radar images to determine the best classification method using 3D GPR.

Author Contributions: Conceptualization, J.Y.; methodology, K.R.; software, K.R.; validation, K.R. and J.G.; formal analysis, K.R.; investigation, S.Y. and L.Z.; resources, J.Y.; data curation, S.Y. and L.Z.; writing—original draft preparation, K.R.; writing—review and editing, K.R. and J.G.; visualization, K.R.; supervision, J.G.; project administration, J.Y.; funding acquisition, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Jiangxi Province of China (grant numbers 20202BABL214046); Science and technology project of Jiangxi Provincial Department of Education (GJJ210623); Key R & D projects of Xinjiang Uygur Autonomous Region (Grant No. 2021B01005); Fundamental Research Funds for the Central Universities, CHD (Grant No. 300102210501).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data, models, and code generated or used during the study appear in the submitted article.

Conflicts of Interest: No conflicts of interest exist in the submission of this manuscript, and the manuscript is approved by all authors for publication. We would like to declare on behalf of my co-authors that the work described was original research that has not been published previously and is not under consideration for publication elsewhere, in whole or in part. All authors listed have approved the manuscript that is enclosed.

Abbreviations

Abbreviations	**Full Name**
3D GPR	Three-dimensional ground penetrating radar
DL	Deep Learning
CNN	Convolutional Neural Network
ViT	Vision Transformer
SViT	Simplify Vision Transformers
YOLOX	You Only Look Once X
CP-YOLOX	CSP PAN YOLOX
CSP	Cross Stage Partial
SPP	Spatial Pyramid Pooling
CBS	Convolutional Batch normalization SiLU
PAN	Path Aggregation Network
SimOTA	Simplify Optimal Transport Assignment
HD	Horizontal Distress
MLP	Multi-Layer Perceptron
BP	Backward Propagation
mAP	mean average precision
NMS	Non-Max Suppression
FLOPs	floating-point operations

References

1. Benedetto, A.; Tosti, F.; Ciampoli, L.B.; D'Amico, F. An overview of ground-penetrating radar signal processing techniques for road inspections. *Signal Process.* **2017**, *132*, 201–209. [CrossRef]
2. Zajícová, K.; Chuman, T. Application of ground penetrating radar methods in soil studies: A review. *Geoderma* **2019**, *343*, 116–129. [CrossRef]
3. Tong, Z. *Research on Pavement Distress Inspection Based on Deep Learning and Ground Penetrating Radar*; Chang'an University: Xi'an, China, 2018; pp. 1–12.
4. Cai, J.; Song, C.; Gong, X.; Zhang, J.; Pei, J.; Chen, Z. Gradation of limestone-aggregate-based porous asphalt concrete under dynamic crushing test: Composition, fragmentation and stability. *Constr. Build. Mater.* **2022**, *323*, 126532. [CrossRef]
5. Klewe, T.; Strangfeld, C.; Kruschwitz, S. Review of moisture measurements in civil engineering with ground penetrating radar —Applied methods and signal features. *Constr. Build. Mater.* **2021**, *278*, 122250. [CrossRef]
6. Luo, C.X. *Research on the Application of Road Nondestructive Testing Technology Based on Three-Dimensional Ground Penetrating Radar*; South China University of Technology: Guangzhou, China, 2018; pp. 55–69.
7. Travassos, X.L.; Avila, S.L.; Ida, N. Artificial Neural Networks and Machine Learning techniques applied to Ground Penetrating Radar: A review. *Appl. Comput. Inform.* **2018**, *17*, 296–308. [CrossRef]
8. Williams, R.M.; Ray, L.; Lever, J.H.; Burzynski, A.M. Crevasse Detection in Ice Sheets Using Ground Penetrating Radar and Machine Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4836–4848. [CrossRef]
9. Zhou, H.; Feng, X.; Zhang, Y.; Nilot, E.; Zhang, M.; Dong, Z.; Qi, J. Combination of Support Vector Machine and H-Alpha Decomposition for Subsurface Target Classification of GPR. In Proceedings of the 17th International Conference on Ground Penetrating Radar (GPR), Rapperswil, Switzerland, 18–21 June 2018; pp. 635–638.
10. Kwon, H.; Kim, Y. BlindNet backdoor: Attack on deep neural network using blind watermark. *Multimedia Tools Appl.* **2022**, *81*, 6217–6234. [CrossRef]
11. Kwon, H. MedicalGuard: U-Net Model Robust against Adversarially Perturbed Images. *Secur. Commun. Netw.* **2021**, *2021*, 1–8. [CrossRef]
12. Kwon, H.; Yoon, H.; Choi, D. Data Correction For Enhancing Classification Accuracy By Unknown Deep Neural Network Classifiers. *KSII Trans. Internet Inform. Syst.* **2021**, *15*, 3243–3257. [CrossRef]
13. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
14. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
17. Tong, Z.; Gao, J.; Yuan, D. Advances of deep learning applications in ground-penetrating radar: A survey. *Constr. Build. Mater.* **2020**, *258*, 120371. [CrossRef]
18. Liu, H.; Lin, C.; Cui, J.; Fan, L.; Xie, X.; Spencer, B.F. Detection and localization of rebar in concrete by deep learning using ground penetrating radar. *Autom. Constr.* **2020**, *118*, 103279. [CrossRef]

19. Li, S.; Gu, X.; Xu, X.; Xu, D.; Zhang, T.; Liu, Z.; Dong, Q. Detection of concealed cracks from ground penetrating radar images based on deep learning algorithm. *Constr. Build. Mater.* **2020**, *273*, 121949. [CrossRef]
20. Liu, Z.; Wu, W.; Gu, X.; Li, S.; Wang, L.; Zhang, T. Application of Combining YOLO Models and 3D GPR Images in Road Detection and Maintenance. *Remote Sens.* **2021**, *13*, 1081. [CrossRef]
21. Sha, A.M.; Tong, Z.; Gao, J. Recognition and Measurement of Pavement Disasters Based on Convolutional Neural Networks. *China J. Highway Trans.* **2018**, *31*, 1–10.
22. Hou, Y.; Chen, Y.H.; Gu, X.Y.; Mao, Q.; Cao, D.D. Automatic Identification of Pavement Objects and Cracks Using the Convolutional Auto-encoder. *China J. Highway Transp.* **2020**, *33*, 288–303.
23. Yan, B.F.; Xu, Y.G.; Luan, J.; Lin, D.; Deng, L. Pavement Distress Detection Based on Faster R-CNN and Morphological Operations. *China J. Highway Transp.* **2021**, *34*, 181–193.
24. Sha, A.M.; Cai, R.N.; Gao, J.; Tong, Z.; Li, S. Subgrade distresses recognition based on convolutional neural network. *J. Chang'an Univ. (Nat. Sci. Ed.)* **2019**, *39*, 1–9.
25. Tong, Z.; Gao, J.; Zhang, H. Recognition, location, measurement, and 3D reconstruction of concealed cracks using convolutional neural networks. *Constr. Build. Mater.* **2017**, *146*, 775–787. [CrossRef]
26. Gao, J.; Yuan, D.; Tong, Z.; Yang, J.; Yu, D. Autonomous pavement distress detection using ground penetrating radar and region-based deep learning. *Measurement* **2020**, *164*, 108077. [CrossRef]
27. Tong, Z.; Yuan, D.; Gao, J.; Wei, Y.; Dou, H. Pavement-distress detection using ground-penetrating radar and network in networks. *Constr. Build. Mater.* **2019**, *233*, 117352. [CrossRef]
28. Long, Z.J. *Reverse-Time Migration Applied to Ground Penetrating Rader and Intelligent Recognition of Subsurface Targets*; Xiamen University: Xiamen, China, 2018; pp. 1–4.
29. Wang, H.; Ou, Y.S.; Liao, K.F.; Jin, L.N. GPR B-SCAN Image Hyperbola Detection Method Based on Deep Learning. *Acta Electr. Sinica* **2021**, *49*, 953–963.
30. Kim, N.; Kim, S.; An, Y.-K.; Lee, J.-J. Triplanar Imaging of 3-D GPR Data for Deep-Learning-Based Underground Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 4446–4456. [CrossRef]
31. Omwenga, M.M.; Wu, D.; Liang, Y.; Yang, L.; Huston, D.; Xia, T. Cognitive GPR for Subsurface Object Detection Based on Deep Reinforcement Learning. *IEEE Internet Things J.* **2021**, *8*, 11594–11606. [CrossRef]
32. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
33. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
34. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
36. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
37. Ge, Z.; Liu, S.; Li, Z.; Yoshie, O.; Sun, J. Ota: Optimal transport assignment for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 303–312.
38. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
39. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
40. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
41. Li, Y.; Huang, H.; Xie, Q.; Yao, L.; Chen, Q. Research on a surface defect detection algorithm based on MobileNet-SSD. *Appl. Sci.* **2018**, *8*, 1678. [CrossRef]
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.