Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.019785

ARTICLE

Multi-Feature Fusion-Guided Multiscale Bidirectional Attention Networks for Logistics Pallet Segmentation

Weiwei Cai[1,2], **Yaping Song**[1], **Huan Duan**[1], **Zhenwei Xia**[1] **and Zhanguo Wei**[1,*]

[1]School of Logistics and Transportation, Central South University of Forestry and Technology, Changsha, 410004, China

[2]Graduate School, Northern Arizona University, Flagstaff, AZ 86011, USA

[*]Corresponding Author: Zhanguo Wei. Email: jackwzg007@csuft.edu.cn

Received: 15 October 2021 Accepted: 01 December 2021

ABSTRACT

In the smart logistics industry, unmanned forklifts that intelligently identify logistics pallets can improve work efficiency in warehousing and transportation and are better than traditional manual forklifts driven by humans. Therefore, they play a critical role in smart warehousing, and semantics segmentation is an effective method to realize the intelligent identification of logistics pallets. However, most current recognition algorithms are ineffective due to the diverse types of pallets, their complex shapes, frequent blockades in production environments, and changing lighting conditions. This paper proposes a novel multi-feature fusion-guided multiscale bidirectional attention (MFMBA) neural network for logistics pallet segmentation. To better predict the foreground category (the pallet) and the background category (the cargo) of a pallet image, our approach extracts three types of features (grayscale, texture, and Hue, Saturation, Value features) and fuses them. The multiscale architecture deals with the problem that the size and shape of the pallet may appear different in the image in the actual, complex environment, which usually makes feature extraction difficult. Our study proposes a multiscale architecture that can extract additional semantic features. Also, since a traditional attention mechanism only assigns attention rights from a single direction, we designed a bidirectional attention mechanism that assigns cross-attention weights to each feature from two directions, horizontally and vertically, significantly improving segmentation. Finally, comparative experimental results show that the precision of the proposed algorithm is 0.53%–8.77% better than that of other methods we compared.

KEYWORDS

Logistics pallet segmentation; image segmentation; multi-feature fusion; multiscale network; bidirectional attention mechanism; HSV; neural networks; deep learning

1 Introduction

The recent rapid development of e-commerce has promoted the prosperity of the logistics industry, accompanied by a demand for logistics that has steadily increased [1,2]. The logistics industry is one of the industries with the fastest growths in personnel. Traditional logistics methods [3,4] can no longer meet the fast-paced needs of current society. Smart logistics has emerged to adapt to these changing needs [5–7], and with the rapid development of artificial intelligence [8–10], smart logistics research has expanded toward automation. The traditional logistics model requires considerable human and material resources, which can solve employment problems to a certain extent. However, current smart logistics needs to reduce high labour costs through automation while solving the shortage of labour [11] as it shifts to other industries. Automated equipment can improve warehousing, material handling, packaging, and distribution efficiency while reducing the error rate. Automated forklifts play a key role in smart logistics, and the accuracy of automated forklifts needed to identify logistics pallets determines their work efficiency and error rates.

Traditional forklift use in storage-oriented activities requires that goods be handled manually, requiring workers to ensure the accuracy of handling at all times. However, the enormous daily flow of goods and long-term repetitive operations exhaust workers, leading to workers forking the goods and even causing safety hazards. Goods are managed in storage stacked on pallets. Accurate identification of logistics pallets can enable automated forklifts to transport materials quickly and efficiently, saving time and significantly reducing logistics costs [12]. Traditional image processing technology cannot provide the performance required for high-precision segmentation [13] and recognition of logistics pallets; so, semantic segmentation is being applied to the image segmentation of logistics pallets to meet these performance requirements.

Liu et al. [14] applied the YOLACT deep learning approach used in artificial intelligence to investigate the detection and segmentation of pallets in the carriage and achieved competitive segmentation performance. Jia et al. [15] combined the Otsu algorithm and the marker watershed algorithm to achieve image segmentation of pallet contours, which provided reference values for designing a warehouse robot for wooden pallet visual inspection by reducing the influences of the surrounding environment and the pallet pattern. Zhao et al. [16] designed a novel GPU-based mean shift algorithm that quickly achieved unsupervised segmentation and tracking of instances. Cui et al. [17] proposed a colour feature-based visual segmentation method that obtains pallet colour feature samples from images in the work environment and then applies morphological filtering, Sobel edge detection, and Hough transform algorithms to recognize the pallets. For pallet detection, Chen et al. [18] proposed converting the colour image from RGB space to Hue, Saturation, Value (HSV) and YUV spaces and then using the camera space model to determine the location of the pallet relative to the forklift, thus establishing the relationship between the image space and the real-world space. However, these colour-based approaches are vulnerable to interference from non-simple backdrops. The Haar-based Adaboost approach, according to Syu et al. [19], uses the AS-for-pallets algorithm to detect pallets. In addition, Seelinger et al. [20] presented mobile camera space manipulation (MCSM), a visual guiding control system to help forklift drivers.

In summary, vision-based detection methods [21–23] can effectively detect pallets against an image background. However, there is still a lack of relevant research on the precise segmentation of pallets, and accurate pallet segmentation [24] depends on whether automatic forklifts can fully automate loading and unloading. Therefore, we developed a multi-feature fusion-guided multiscale bidirectional attention (MFMBA) neural network for logistics pallet segmentation. First, multi-feature extraction and fusion make up for the shortcomings of vision-based methods that are easily misled by the background. Second, in an actual complex environment, the sizes and shapes of pallets in the same image may be different, which makes feature acquisition difficult, but the multiscale architecture can extract more semantic features, thereby enhancing the feature mining capabilities of the segmentation model. In addition, the bidirectional attention mechanism [25] assigns bidirectional attention weights to each feature, which further improves the segmentation performance of the model.

These are the study's principal innovations:

(1) This paper proposes an MFMBA network for logistics pallet segmentation. Our study has achieved competitive segmentation performance on datasets in real-world production environments.
(2) To better predict the foreground category (the pallet) and the background category (the goods) in an image, we extract the grayscale, texture, and HSV features from the pallet image and then fuse them using a feature concatenation strategy.
(3) Our novel bidirectional attention mechanism assigns weights to each feature from two directions (horizontally and vertically), which is better than traditional attention mechanisms that only assign attention weights from a single direction.

The remainder of the paper is laid out as follows: Sections 2 and 3 describe related work and explain the theoretical basis for the proposed algorithm. The comparison and ablation experiments are described in Section 4, and we present our conclusions in Section 5.

2 Related Work

2.1 Image Segmentation

The process of assigning a label to each pixel in an image so that pixels with the same label have similar characteristics is known as image segmentation [26,27]. Image segmentation can be defined using the concept of set: assuming that the entire digital image is represented by set R, image segmentation can be understood as dividing R into regions R_1, R_2,..., R_n and all subregions meeting the following conditions:

$$\begin{cases} \bigcup\limits_{i=1}^{n} R_i = R, \quad i = 1, 2, \cdots, n \\ R_i \cap R_j \neq \varnothing, \quad i \neq j, j = 1, 2, \cdots, n \\ Q(R_i) = True, \quad i = 1, 2, \cdots, n \\ Q(R_i \cup R_j) = False \end{cases} \tag{1}$$

where $Q(R_i)$ is an attribute of the pixels of the set R, $\varnothing$ indicates the empty set, $\cap$ is the intersection of sets, and $\cup$ indicates the union of sets. If the union of R_i and R_j forms a connected set, the two areas are defined as adjacent. It can be seen from Eq. (1) that after segmentation, each pixel in the image has a category attribute, and the pixels in any sub-region obtained after segmentation are all connected to four or eight other pixels. In addition, the pixels have one and only one category attribute, that is, sub-regions do not intersect, and two adjacent regions have different attributes.

During image processing and analysis, only a small portion of the image is usually examined. As a result, to study image data, you must first identify and extract the portion of interest from the entire image. The target is then analysed on this basis. Image segmentation is an essential step in the intelligent identification of logistics pallets.

2.2 Attention Mechanism

The attention mechanism [28,29] originated from the study of human attention. Due to the limitations on our information-processing capabilities, humans selectively focus on part of the information they receive. This is also the ability that we need the model to have when receiving and learning a large amount of information. In mathematical terms, attention is learning a set of weight coefficients through the model independently and dynamically assigning this series of weights to each area of the information received by the model. The attention mechanism is widely used in neural networks, especially in image segmentation tasks. The principle of the attention mechanism is shown in Fig. 1. If the input variable is set to $X = [x_1, x_2, \cdots, x_n]$, the equation for calculating the attention distribution is as follows:

$$\alpha_i = \mathrm{softmax}(h(x_i, p)) \text{ and} \tag{2}$$

$$\alpha_1 + \alpha_2 + \alpha_3 \cdots + \alpha_n = 1, \tag{3}$$

where α_i is the weight of attention distribution corresponding to the i-th input variable x_i, which is also a probability distribution and satisfies Eq. (2). $h(x_i, p)$ is called the attention score of the i-th input variable, which is determined by x_i and a pre-set vector p. Common attention scoring methods include bilinear scoring and dot product scoring; their calculation equations are as follows:

$$s(x_i, p) = x_i^T W p \text{ and} \tag{4}$$

$$s(x_i, p) = x_i^T p. \tag{5}$$

After obtaining the attention distribution, multiply the input variable x_i and the corresponding attention distribution a_i, and then sum them as follows:

$$attention(X, p) = \sum_{i=1}^{n} a_i x_i. \tag{6}$$

3 Methodology

The overall architecture of the proposed MFMBA algorithm is depicted schematically in Fig. 1. This paper extracts the HSV feature, grayscale feature (GF), and texture feature (TF) from logistics pallet images, applies a feature-stitching strategy for feature fusion, and inputs the fusion features to the proposed multiscale bidirectional attention network to extract deep features. The sigmoid function is then used to achieve semantic segmentation of the logistics pallets. The MFMBA algorithm is discussed in detail in the following sections.

3.1 Multi-Feature Extraction

To improve segmentation accuracy, we first extract the TF, GF, and HSV feature from the pallet image to better distinguish the foreground category (the pallet) from the background category (the cargo).

Texture features: Texture is an important distinguishing feature on the surface of an object. When the image is transformed into different brightnesses and colours, the pixels follow a specified rule and undergo near-periodical changes. Texture characteristics can effectively deal with logistics pallet images in various light environments. The calculation equation for TF extraction is as follows:

$$LBP(x_c, y_c) = \sum_{P=0}^{P-1} 2^P s(i_p - i_c), s(x) = \begin{Bmatrix} 1 & x \geq 0 \\ 0 & else \end{Bmatrix}, \tag{7}$$

where (x_c, y_c) is the central pixel, i_c represents the brightness of the point, i_p is the brightness of the adjacent pixels, and s represents the Sigmoid function.

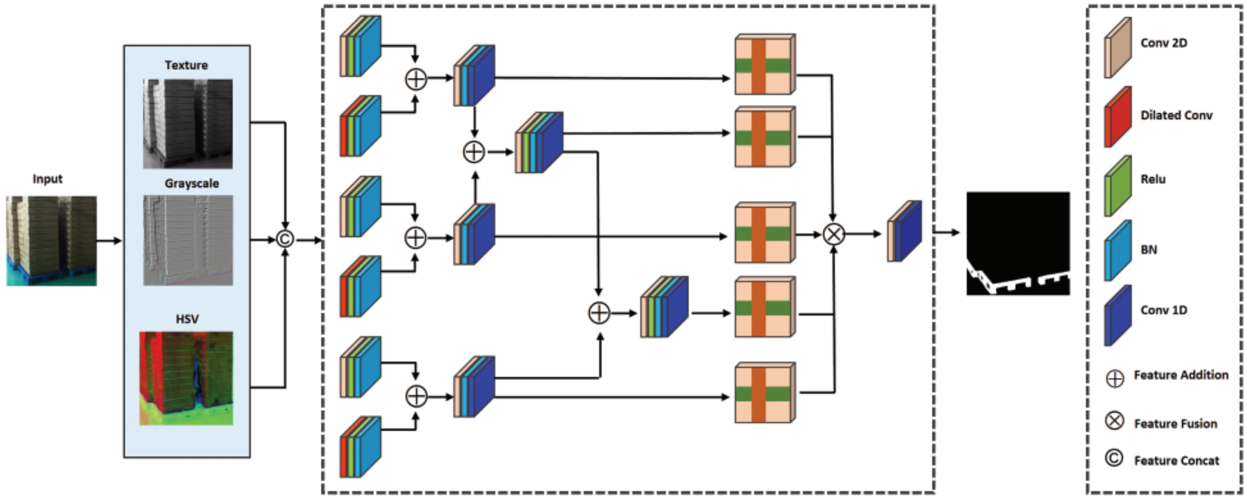


Figure 1: The architecture of the proposed MFMBA algorithm

The basic principle of the local binary pattern (LBP) is that a particular pixel is centred; then, its value is compared with other pixel values in its 3 × 3 window. Every compared pixel value greater than that of the center point equals 1; otherwise, it is 0. Thus, a 3 × 3 window provides eight binary numbers and converts the binary to decimal to obtain the LBP code, which represents the texture. The LBP is shown schematically in Fig. 2.

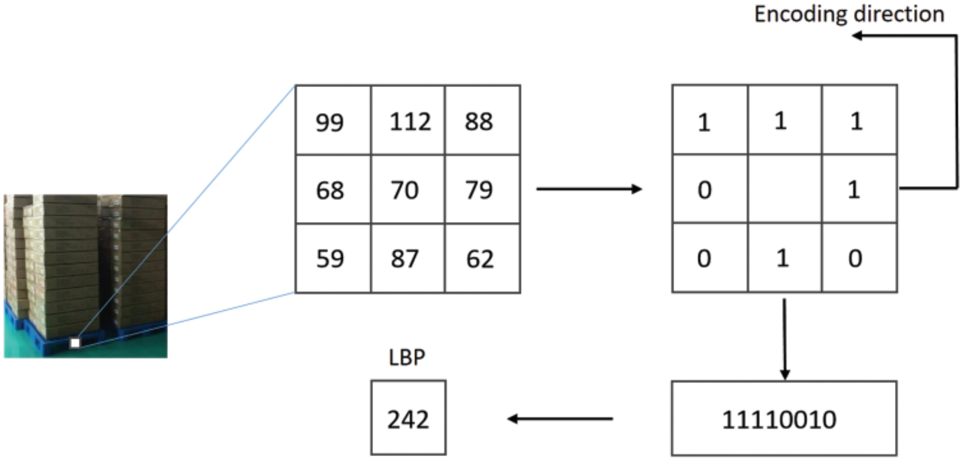


Figure 2: Schematic diagram of an LBP

Grayscale features: Grayscale uses black tones to represent objects; black is used as the reference colour, and blacks of different saturations are used to display the image. Each grayscale image has a brightness value from 0% (white) to 100% (black). Because it has less redundant information, grayscale improves image segmentation. The calculation equation is as follows:

$$Gray = 0.229R + 0.587G + 0.114B, \tag{8}$$

where R, G, and B represent the three-channel colours of the logistics pallet image.

HSV features: The HSV colour space, also known as the hexcone model, was created by A. R. Smith in 1978 based on the intuitive characteristics of colours. Hue (H), saturation (S), and lightness (L) are the colour parameters in this model (V). We must first convert the red, green, and blue coordinates of a colour to real numbers between 0 and 1 before using RGB to represent them. The following are the calculation formulas:

$$\begin{cases} red = R/255 \\ green = G/255 \\ blue = B/255 \end{cases}. \tag{9}$$

Next, we calculate the values of H, S, and V as follows:

$$H = \begin{cases} \dfrac{60(G-B)}{V-\min(R,G,B)}, & V = R \\ 120 + \dfrac{60(B-R)}{V-\min(R,G,B)}, & V = G \\ 240 + \dfrac{60(R-G)}{V-\min(R,G,B)}, & V = B \end{cases}, \tag{10}$$

$$S = \begin{cases} \dfrac{V-\min(R,G,B)}{V} & V \neq 0 \\ 0 & otherwise \end{cases}, \tag{11}$$

$$V = \max(R,G,B). \tag{12}$$

RGB features are output as HSV features using the equation above. The new output vector block will be used as a feature sequence in our MFMBA model. Furthermore, the calculation result may contain $H < 0$. H requires additional calculation processing at this time. The following shows:

$$H = \begin{cases} H + 360 & H < 0 \\ H & otherwise \end{cases}, \tag{13}$$

where $H \in [0, 360]$, $S \in [0, 1]$, and $V \in [0, 1]$.

3.2 Multiscale Hybrid Convolution

Using a multiscale convolution kernel in the proposed algorithm has two distinct advantages. The most significant benefit of multiscale convolution kernels is that differently sized kernels can extract features from logistics pallet images of various scales, allowing the filter to extract and learn richer characterisation information. Also, the convolutional neural network trains the model by learning the filter's parameters (weight and offset), that is, it continuously learns the filter's parameters to find the optimal value closest to the label. This article employs a multiscale convolution kernel to allow a given convolution layer to have multiple filters, thereby diversifying the weight and deviation learning, thus extracting and learning the semantic features of the logistics pallet image fully and effectively.

Multiscale inference methods [30–32] are commonly used in computer vision models for the best results. Fine details are better predicted at larger sizes, larger objects are better predicted at smaller sizes, and the network's receiving field understands the scene better at smaller sizes. This paper proposes a multiscale hybrid convolution model [33] that is different from the traditional multiscale structure shown in Fig. 3. To extract features in the three sizes of 11×11, 7×7, and 3×3, we use traditional convolution and hole convolution. The following is the calculation formula:

$$\begin{aligned} &Y_c{}^1 = \varphi\left\{\sum_{i=0}^{n1} w_{ij}^1 * x_i^1 + b_J^1\right\} \\ &Y_d{}^1 = \varphi\left\{\sum_{\mathrm{l}=0}^{n1}\sum_{\mathrm{m}=0}^{\mathrm{n1}} w_{l,m}^1 * x_{j+l,k+m}^1 + b_J^1\right\}, \end{aligned} \tag{14}$$

$$\begin{aligned} &Y_c{}^2 = \varphi\left\{\sum_{i=0}^{n2} w_{ij}^2 * x_i^2 + b_J^2\right\} \\ &Y_d{}^2 = \varphi\left\{\sum_{\mathrm{l}=0}^{n2}\sum_{\mathrm{m}=0}^{\mathrm{n2}} w_{ij}^2 * x_{j+1,k+m}^2 + b_J^2\right\}, \end{aligned} \tag{15}$$

$$\begin{aligned} &Y_c{}^3 = \varphi\left\{\sum_{i=0}^{n3} w_{ij}^3 * x_i^3 + b_J^3\right\} \\ &Y_d{}^3 = \varphi\left\{\sum_{\mathrm{l}=0}^{n3}\sum_{\mathrm{m}=0}^{\mathrm{n3}} w_{ij}^3 * x_{j+1,k+m}^3 + b_J^3\right\}, \end{aligned} \tag{16}$$

where h_j is the pixel feature vector's hidden state information, k is the feature point, $j * k$ is the size of the feature map, and $l * m$ is the size of the hollow convolution's local receptive field.

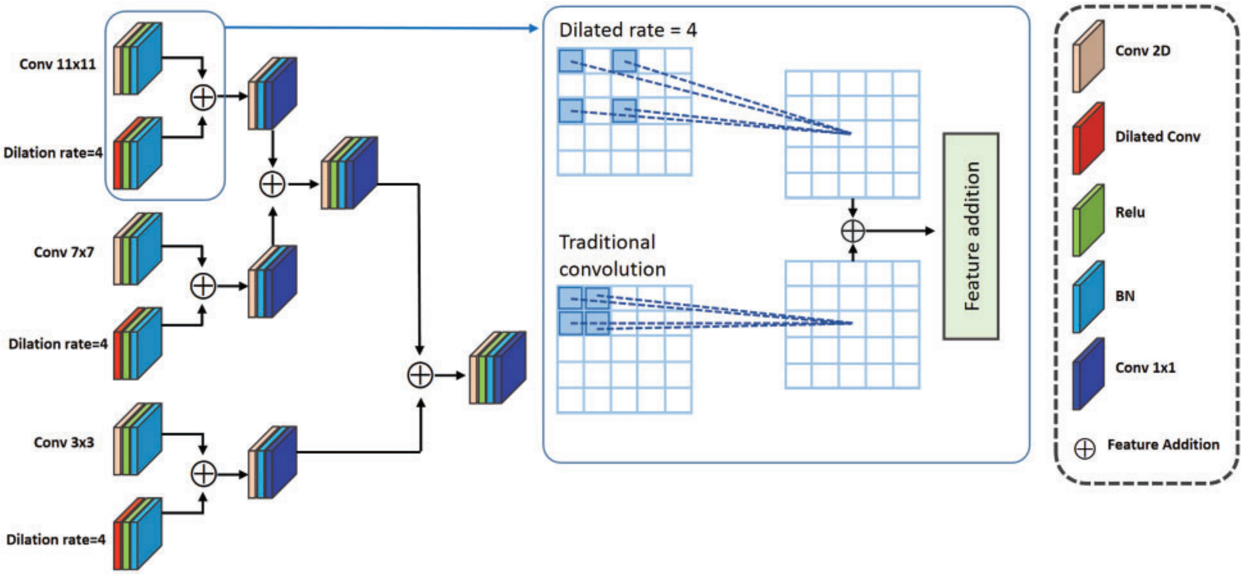


Figure 3: Schematic diagram of multiscale hybrid dilated convolution

3.3 Bidirectional Attention Mechanism

The model is divided into three parts and has a novel bidirectional attention mechanism, which is the first section of the model. To effectively detect the local semantic information of each pixel in the pallet image, we map all the characteristics onto a two-dimensional space and apply a bidirectional weight to each feature using bidirectional attention. The second section of the model includes the two types of weight features to broaden the weight coefficient. The third section combines the two types of weight features to produce the greatest value, which is then utilized to complement the weight coefficient result obtained in the second step. The bidirectional attention mechanism is shown schematically in Fig. 4.

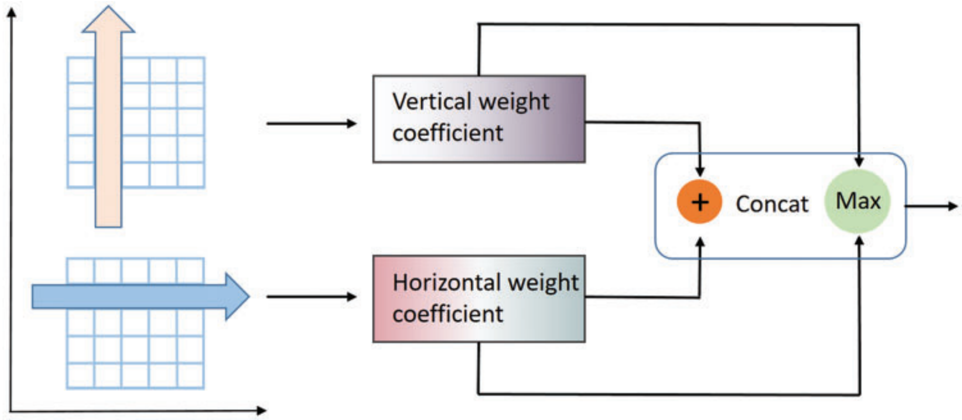


Figure 4: Schematic diagram of the bidirectional attention mechanism model

Let the feature map extracted from the previous convolutional layer be $m_{i,j}^{h} \in R^{H\times W}$, where H and W are the feature map's height and width, respectively; then, input $m_{i,j}^{k}$ into the horizontal attention module to obtain the attention weight. The steps in the calculation are as follows:

$$Att_h = \frac{\exp(W_h m_{i,j} + b_h)}{\sum_{i,j}\exp(W_h m_{i,j} + b_h)}, \tag{17}$$

where W_h and b_h are the weight parameters of the dense layer, and Att_h represents the attention coefficient in the horizontal direction.

For the vertical attention mechanism, we transpose the matrix of the feature map to obtain the feature map in the vertical direction. The calculation equation is as follows:

$$m_{j,i}^{v} = (m_{i,j}^{v})^{T}, \tag{18}$$

where $m_{j,i}^{v}$ represents the feature map flipped vertically. Similarly, input it to the vertical attention module to obtain the vertical attention weight. The calculation equation of the weight coefficient is as follows:

$$Att_v = \frac{\exp(W_v m_{j,i} + b_v)}{\sum_{j,i}\exp(W_v m_{j,i} + b_v)}. \tag{19}$$

Therefore, the calculation equation of the output of the bidirectional attention mechanism model is as follows:

$$Add = (Att_h + Att_v), \tag{20}$$

$$Max = (Att_h, Att_v), \text{and} \tag{21}$$

$$BA = concatenate[Att_h, Att_v, Add, Max], \tag{22}$$

where BA represents the output of the bidirectional attention mechanism model.

3.4 Feature Fusion

In this section, feature fusion is performed on the output of the multiscale semantic feature and the bidirectional attention mechanism and is then segmented by the sigmoid function. The calculation equation is as follows:

$$\begin{cases} M_1 = add[Y_c^1, {Y_d}^1] \\ M_2 = add[Y_c^2, {Y_d}^2] \\ M_3 = add[Y_c^3, {Y_d}^3] \\ M_4 = \text{concatenate}[M_1, M_2] \\ M_5 = \text{concatenate}[M_3, M_4] \end{cases}, \tag{23}$$

$$F = add[BA(M_1), BA(M_2), BA(M_3), BA(M_4), BA(M_5)], \tag{24}$$

where $M_1, M_2, \ldots, M_5$ represents the fusion output of the hybrid dilated convolution of each scale, *add* represents the summation operation on the feature tensor, and *concatenate* represents the concatenation operation on the feature tensor.

The final output of segmentation using the sigmoid function is:

$$O = sigmoid(F). \tag{25}$$

4 Experiments and Results

4.1 Dataset

A pallet is a medium that transforms static goods into dynamic goods—it is a loading platform. Since the focus of this article is the intelligent identification and segmentation of logistics pallets in industrial production environments, we collected images of pallets in complex environments from the Internet. The collected images are of different sizes and pixel sizes. We uniformly cropped the size of the pallet image to 256 × 256. To obtain the pallet image segmentation dataset, we used ENVI software to annotate each image manually. An example of the pallet image after cropping and annotation is shown in Fig. 5.

4.2 Experiment Environments

All of the experiments in this article were conducted on a computer with a single NVIDIA GTX1080 GPU to fairly verify and compare the performance of the proposed algorithms (8 GB). The keras2.1.5 deep learning library was used to construct the model. We used Python 3.6.5 as our programming language, and we processed 1280 samples each time in batches. The setting of each of the above hyperparameters was tested extensively in this study. These parameters are the best in this experiment. Table 1 summarizes the final hyperparameters. Furthermore, we used Adam [34] as the optimizer for the proposed algorithm, which converges quickly. Table 1 lists the most important parameters: The learning rate is 0.01; α indicates that the first-order moment estimation's exponential decay rate is 0.99; β indicates that the second-order moment estimation's exponential decay rate is 0.999; Epsilon is set to 1e-8; and Decay indicates that the learning rate decay is 3e-8.

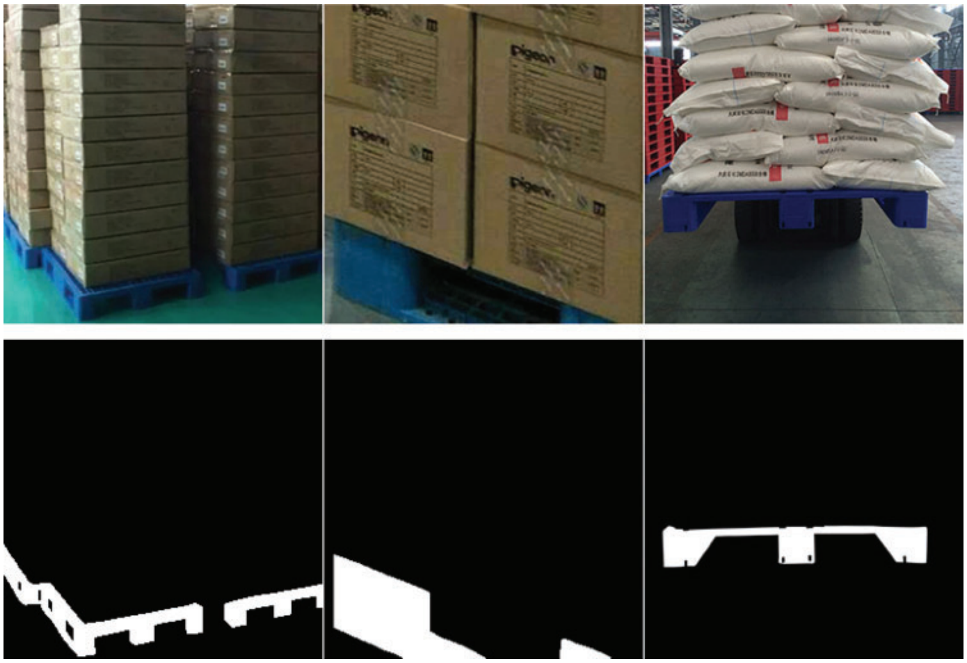


Figure 5: Examples of original and segmented logistics pallet images

Table 1: Setting the hyperparameters

Item	Value
OS	Windows 10
GPU	NVIDIA GTX1080
Deep learning framework	Keras 2.1.5
Batch size	2000
Epochs	400
Learning rate	0.01

4.3 Evaluation Methods

This paper uses three evaluation indicators—precision (P), recall (R), and F1 score (F1)—to evaluate the segmentation performance of the proposed MFMBA algorithm comprehensively. The following are the calculation formulas for Precision, Recall, and F1 score:

$$\mathrm{P} = \frac{TP}{TP + FP}, \tag{26}$$

$$\mathrm{R} = \frac{TP}{TP + FN}, \text{and} \tag{27}$$

$$F1 = \frac{2 \times P \times R}{P + R}, \tag{28}$$

where TP represents a true positive (the number of pixels of the logistics pallet that were correctly detected), FP represents a false positive (the number of pixels of the logistics pallet that were incorrectly detected), and FN represents a false negative (the number of pixels of the logistics pallet that were incorrectly detected).

4.4 Experimental Results of Different Methods

In this section, a comparative experiment is conducted to demonstrate the superiority of the proposed algorithm. Furthermore, all experiments were carried out in the same environment and with the same hyperparameters. We compared AlexNet [18], Res-Net [19], DenseNet [20], Unet [21], and DeepLab-v3 [2] with the proposed MFMBA model. The comparative experimental results of various methods are shown in Tables 2 and 3 and Fig. 6. Only 1% and 20% of the training samples were chosen in separate experiments.

Table 2: The result of comparison with five different models under 1% of the training samples

Methods	1% of the training samples		
	P	R	F1
AlexNet [35]	0.8950	0.8555	0.8790
ResNet [36]	0.9188	0.8602	0.8885
DenseNet [37]	0.8487	0.9376	0.8910
Unet [38]	0.8497	0.8915	0.8579
DeepLab-v3 [39]	0.8512	0.9013	0.8755
MFMBA (Proposed method)	**0.9237**	**0.9417**	**0.9275**

Table 3: The result of comparison with five different models under 20% of the training samples

Methods	20% of the training samples		
	P	R	F1
AlexNet [18]	0.9481	0.9624	0.9552
ResNet [19]	0.9517	0.9675	0.9595
DenseNet [20]	0.8917	0.9580	0.9478
Unet [21]	0.9258	0.9389	0.9524
DeepLab-v3 [22]	0.9387	0.9221	0.9599
MFMBA (Proposed method)	**0.9564**	**0.9793**	**0.9626**

Due to the wide variety of pallets, shape complexity, strong regularity, and complex environments (e.g., pallets being occluded in the industrial production environment and changing lighting conditions), the semantic segmentation of the pallet segmentation image can be an arduous task. The training set was made up of either 1% or 20% of the total number of samples. Table 2 shows that the overall residual network outperforms the dense network and AlexNet, as evidenced by the experimental results. Because the residual network preserves many shallow features, and the residual calculation and deep features are better merged to gain additional features, the residual calculation and deep features are better integrated to obtain more features. Fig. 6 shows that ResNet has a high level of accuracy on positive samples.

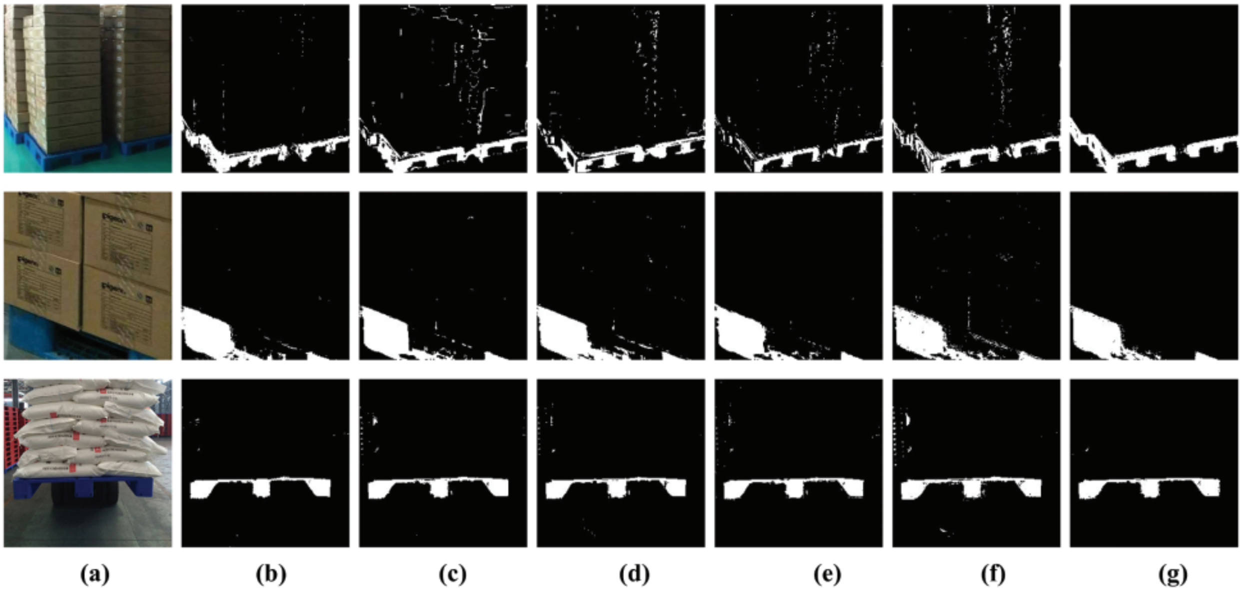


Figure 6: Segmentation results under 1% training sample. (a) Input image; (b) AlexNet; (c) ResNet; (d) DenseNet; (e) Unet; (f) DeepLab-v3; and (g) MFMBA (Ours)

Furthermore, the residual network has better performance for mining features, as evidenced by the precision index. On the same training sample, our proposed MFMBA algorithm outperforms other methods in P (0.5%–8.1% higher than the others), R (0.4%–9.1% higher than the others), and F1 score (compared to the other five groups of models, 3.9%–7.5% higher), demonstrating its feasibility. Fig. 6 depicts the outcomes of the experiment (using 1% of the samples for training). Table 3 shows that, with the increase in the number of training samples, the performance of our algorithm is significantly improved and outperforms the other five methods. This fully demonstrates the effectiveness of our model.

4.5 Ablation Experiment on the MFMBA Sub-Module

The sub-modules of the proposed algorithm were subjected to ablation experiments and are described in this section. The multi-feature fusion (MFF) module, multiscale network (MSN), and bidirectional attention (BA) mechanism are acronyms for multi-feature fusion module, multiscale network, and bidirectional attention mechanism, respectively. We combined them and ran separate experiments to see which sub-modules have the greatest impact on segmentation performance. Table 4 summarizes the findings of the ablation experiment.

Table 4 and Fig. 7 clearly show that any two sub-modules perform better segmentation than a single module. MFF outperforms a single MS and BA, demonstrating the utility of multi-feature extraction. Moreover, the MFF MS combination is superior to the MS and BA combination because the multi-feature extraction and fusion module can extract richer semantic information. Furthermore, the combined model outperforms a single module, demonstrating that the proposed algorithm's MFF extraction, MSN, and BA mechanism are effective. As a result, MFMBA's effectiveness is also demonstrated.

Table 4: Ablation experiment results of the sub-module of MFMBA under 1% and 20% training samples

Methods	1% training samples			20% training samples		
	P	R	F1	P	R	F1
MFF	0.9087	0.9254	0.9187	0.9364	0.9589	0.9488
MS	0.8847	0.9258	0.9005	0.9258	0.9210	0.9187
BA	0.8547	0.8825	0.8974	0.9105	0.9055	0.8945
MS-BA	0.8854	0.9187	0.9174	0.9415	0.9574	0.9478
MFF-MS	0.9199	0.9409	0.9250	0.9501	0.9714	0.9614
MFF-BA	0.9199	0.9411	0.9201	0.9514	0.9647	0.9601
MFMBA (Ours)	**0.9237**	**0.9417**	**0.9275**	**0.9564**	**0.9793**	**0.9626**

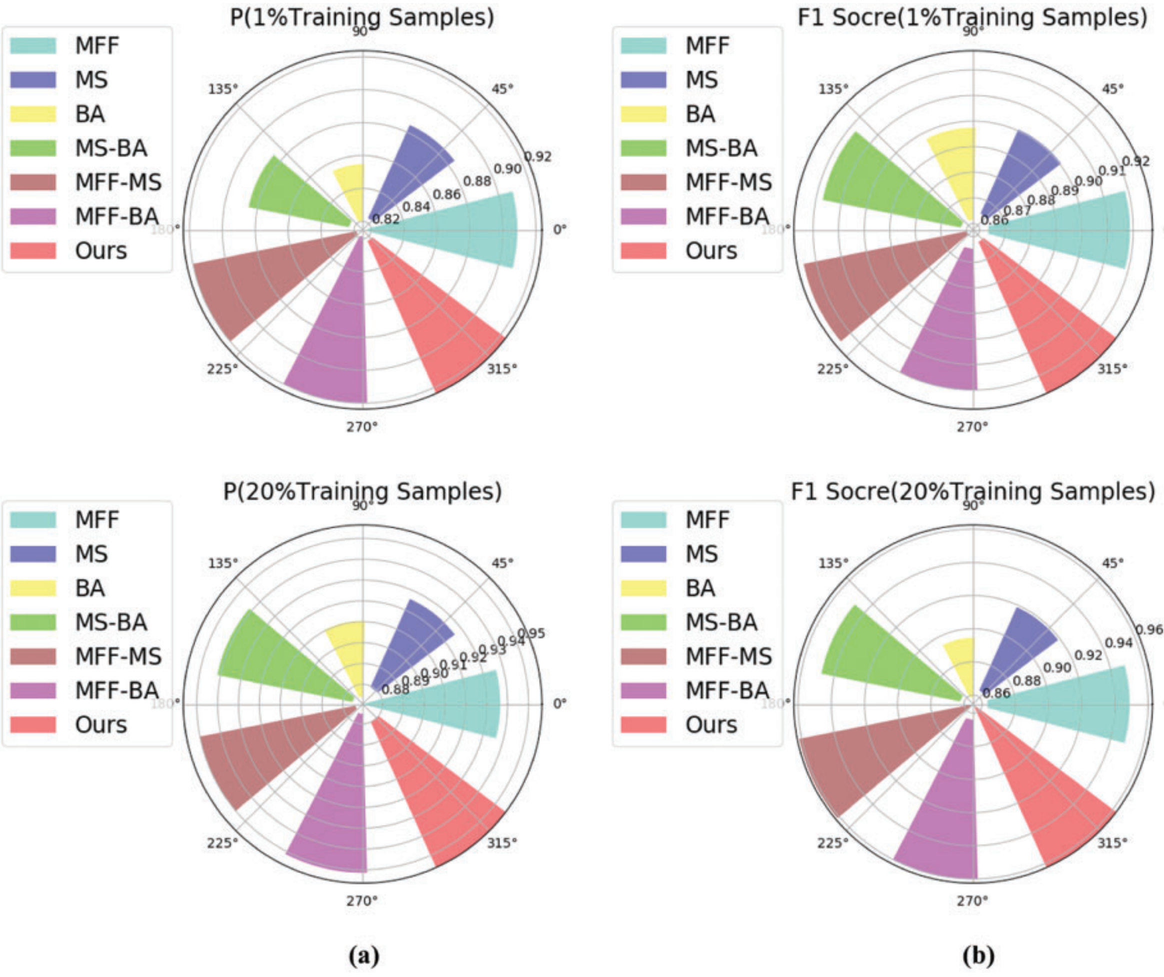


Figure 7: Visualized results of ablation experiments. (a) Precision; and (b) F1 score

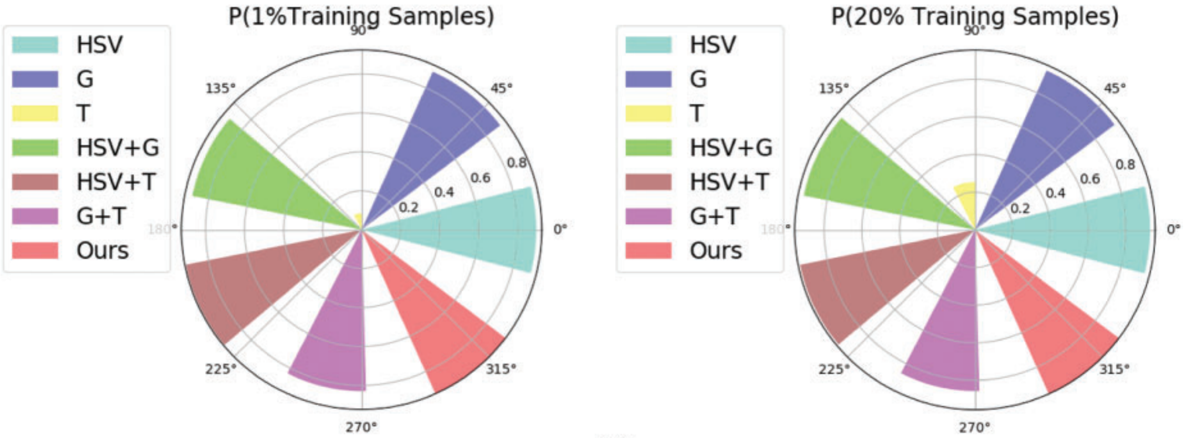
4.6 Ablation Experiment of Multi-Feature Fusion

In the previous section's ablation experiment, we discovered that the MFF module performs exceptionally well in the proposed algorithm. As a result, this section sets up an ablation experiment to investigate the impact of various features on the experimental outcomes. The three extracted features were abbreviated as HSV, T, and G, and ablation experiments were performed on combinations of these three features. Table 5 presents the results of the experiment.

Table 5: Experimental results of ablation studies on multi-feature fusion under 1% and 20% training samples

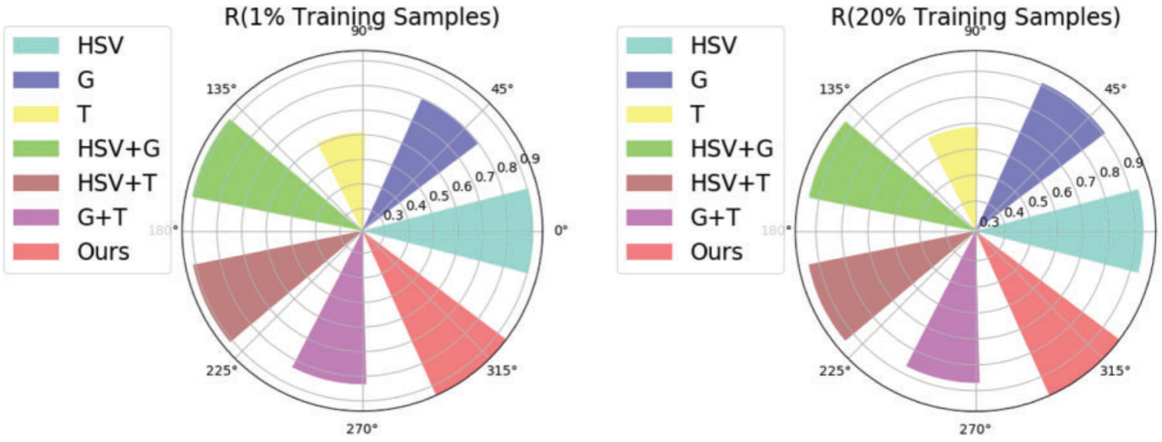
Methods	1% training samples			20% training samples		
	P	R	F1	P	R	F1
HSV	0.8945	0.9043	0.8994	0.9285	0.9317	0.9366
G	0.8901	0.7971	0.8455	0.9284	0.9105	0.8958
T	0.0847	0.6096	0.1488	0.2545	0.6854	0.3922
HSV + G	0.8842	0.9141	0.8989	0.9258	0.9399	0.9287
HSV + T	0.9184	0.9111	0.9147	0.9458	0.9412	0.9587
G + T	0.8297	0.8322	0.8310	0.8574	0.8695	0.8717
MFMBA (Ours)	**0.9237**	**0.9417**	**0.9275**	**0.9564**	**0.9793**	**0.9626**

From Fig. 8 and Table 5, we see that the segmentation performance using texture features is the worst, while the performance using HSV features is the best. HSV features contain more semantic information, while logistics pallets' grayscale features do not. Local features can also be described in greater detail with greater accuracy. Furthermore, the fusion of any two groups of features exceeds the utility of a single feature, meaning that integrating several characteristics provides more semantic information than using a single feature. The results suggest that the MFF of the algorithm is effective.



(a)

Figure 8: (Continued)



(b)

Figure 8: Visualized results of multi-feature fusion ablation experiment. (a) Precision. (b) Recall

5 Conclusions

This paper proposes a novel MFMBA neural network for logistics pallet segmentation. To better predict the foreground category (the pallet) and background category (the cargo) of the pallet image, three types of features (grayscale, texture, and HSV) are extracted and fused. Experimental results demonstrate that all three features improve the segmentation performance of the model, especially the HSV feature. Also, we demonstrated the superiority of the multiscale architecture, which extracts more semantic features than other architectures used to date. In addition, since the traditional attention mechanism only allocates attention from a single direction, we also designed a two-way attention mechanism that can assign cross-attention weights to each feature from two directions (horizontally and vertically). This mechanism improves the segmentation performance of the proposed algorithm, which is also demonstrated by comparison and ablation experiments.

Funding Statement: This work was supported by the Postgraduate Scientific Research Innovation Project of Hunan Province under Grant QL20210212 and the Scientific Innovation Fund for Postgraduates of Central South University of Forestry and Technology under Grant CX202102043.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Dangelmaier, W., Fahrentholz, M., Franke, H., Mueck, B. (2001). A demand-driven logistics concept for the fully automated rail system NBP. *World Congress o Railway Research*, Cologne.
2. Sheu, J. B. (2007). An emergency logistics distribution approach for quick response to urgent relief demand in disasters. *Transportation Research Part E: Logistics and Transportation Review, 43(6),* 687–709. DOI 10.1016/j.tre.2006.04.004.
3. Zielske, M., Held, T. (2021). Application of agile methods in traditional logistics companies and logistics startups: Results from a German delphi study. *Journal of Systems and Software, 177,* 110950. DOI 10.1016/j.jss.2021.110950.

4. Rodríguez Cornejo, V., Cervera Paz, Á., López Molina, L., Pérez-Fernández, V. (2020). Lean thinking to foster the transition from traditional logistics to the physical internet. *Sustainability, 12(15),* 6053. DOI 10.3390/su12156053.
5. Kawa, A. (2012). SMART logistics chain. *Asian Conference on Intelligent Information and Database Systems*, pp. 432–438. Springer, Berlin, Heidelberg
6. Lee, C. K., Lv, Y., Ng, K. K. H., Ho, W., Choy, K. L. (2018). Design and application of Internet of Things-based warehouse management system for smart logistics. *International Journal of Production Research, 56(8),* 2753–2768. DOI 10.1080/00207543.2017.1394592.
7. Humayun, M., Jhanjhi, N. Z., Hamid, B., Ahmed, G. (2020). Emerging smart logistics and transportation using IoT and blockchain. *IEEE Internet of Things Magazine, 3(2),* 58–62. DOI 10.1109/MIoT.8548628.
8. Wu, Z., Chu, W. (2021). Sampling strategy analysis of machine learning models for energy consumption prediction. *IEEE 9th International Conference on Smart Energy Grid Engineering*, pp. 77–81. Oshawa, Canada, IEEE.
9. Cai, W., Wei, Z., Liu, R., Zhuang, Y., Wang, Y. et al. (2021). Remote sensing image recognition based on multi-attention residual fusion networks. *ASP Transactions on Pattern Recognition and Intelligent Systems, 1(1),* 1–8. DOI 10.52810/TPRIS.
10. Li, X., Qiu, J. (2021). A multi-parameter video quality assessment model based on 3D convolutional neural network on the cloud. *ASP Transactions on Internet of Things, 1(2),* 14–22. DOI 10.52810/TIOT.2021.100063.
11. Hofbauer, F., Putz, L. M. (2019). Can gamification reduce the shortage of skilled logistics personnel?. *Artificial Intelligence and Digital Transformation in Supply Chain Management: Innovative Approaches for Supply Chains. Proceedings of the Hamburg International Conference of Logistics*, vol. 27, pp. 331–354, Berlin: epubli GmbH.
12. Hidayat, R. D. R., Firdaus, M. I., Lesmini, L., Purwoko, H. (2017). Logistics bonded center as new customs facility breakthrough for reducing logistics time and cost. *Global Research on Sustainable Transport*, pp. 47–58. Hall of Trisakti Institute of Transportation, Atlantis Press.
13. Cai, W., Wei, Z., Song, Y., Li, M., Yang, X. (2021). Residual-capsule networks with threshold convolution for segmentation of wheat plantation rows in UAV images. In: *Multimedia tools and applications*, vol. 80, pp. 1–17.
14. Liu, D., Zheng, Z. (2021). Research on logistics transportation of detection and segmentation based on deep learning. *2021 IEEE International Conference on Artificial Intelligence and Industrial Design*, pp. 356–359. Guangzhou, China, IEEE.
15. Jia, F., Tao, Z., Wang, F. (2021). Wooden pallet image segmentation based on otsu and marker watershed. *Journal of Physics: Conference Series, 1976(1),* 012005. DOI 10.1088/1742-6596/1976/1/012005.
16. Zhao, M., Jha, A., Liu, Q., Millis, B. A., Mahadevan-Jansen, A. et al. (2021). Faster mean-shift: GPU-accelerated clustering for cosine embedding-based cell segmentation and tracking. *Medical Image Analysis, 71,* 102048. DOI 10.1016/j.media.2021.102048.
17. Cui, G. Z., Lu, L. S., He, Z. D., Yao, L. N., Yang, C. X. et al. (2010). A robust autonomous mobile forklift pallet recognition. *2nd International Asia Conference on Informatics in Control, Automation and Robotics*, vol. 3, pp. 286–290. Wuhan, China, IEEE.
18. Chen, G., Peng, R., Wang, Z., Zhao, W. (2012). Pallet recognition and localization method for vision guided forklift. *8th International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 1–4. Shanghai, China, IEEE.
19. Syu, J. L., Li, H. T., Chiang, J. S., Hsia, C. H., Wu, P. H. et al. (2016). An assisted forklift pallet detection with adaptive structure feature algorithm for automated storage and retrieval systems. *International Conference on Frontier Computing*, pp. 251–260. Springer, Singapore.
20. Seelinger, M., Yoder, J. D. (2006). Automatic visual guidance of a forklift engaging a pallet. *Robotics and Autonomous Systems, 54(12),* 1026–1038. DOI 10.1016/j.robot.2005.10.009.

21. Garibotto, G., Masciangelo, S., Ilic, M., Bassino, P. (1997). Service robotics in logistic automation: Robolift: Vision based autonomous navigation of a conventional fork-lift for pallet handling. *8th International Conference on Advanced Robotics*, pp. 781–786. Monterey, CA, USA, IEEE.
22. Zheng, Y. Y., Kong, J. L., Jin, X. B., Wang, X. Y., Su, T. L. et al. (2019). CropDeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors, 19(5),* 1058. DOI 10.3390/s19051058.
23. Vaira, R., Pietrini, R., Pierdicca, R., Zingaretti, P., Mancini, A. et al. (2019). An IoT edge-fog-cloud architecture for vision based pallet integrity. *International Conference on Image Analysis and Processing*, pp. 296–306. Springer, Cham.
24. Fooladivanda, A., Chehrerazi, N., Sadri, S., Amirfattahi, R., Montazeri, M. A. (2010). Automatic segmentation of pallet images using the 2-D wavelet transform and YUV color space. *18th Iranian Conference on Electrical Engineering*, pp. 209–214. Isfahan University of Technology-Iran, IEEE.
25. Cai, W., Wei, Z. (2020). Remote sensing image classification based on a cross-attention mechanism and graph convolution. *IEEE Geoscience and Remote Sensing Letters*. (in Press). DOI 10.1109/LGRS.8859.
26. Cai, W., Zhai, B., Liu, Y., Liu, R., Ning, X. (2021). Quadratic polynomial guided fuzzy C-means and dual attention mechanism for medical image segmentation. *Displays, 70*, 102106. DOI 10.1016/j.displa.2021.102106.
27. Ghosh, S., Das, N., Das, I., Maulik, U. (2019). Understanding deep learning techniques for image segmentation. *ACM Computing Surveys, 52(4),* 1–35. DOI 10.1145/3329784.
28. Zhang, L., Sun, L., Yu, L., Dong, X., Chen, J. et al. (2021). ARFace: Attention-aware and regularization for face recognition with reinforcement learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science* (in Press). DOI 10.1109/TBIOM.2021.3104014.
29. Gao, M., Cai, W., Liu, R. (2021). AGTH-Net: Attention-based graph convolution-guided third-order hourglass network for sports video classification. *Journal of Healthcare Engineering, 2021*, 8517161, DOI 10.1155/2021/8517161.
30. Najibi, M., Singh, B., Davis, L. S. (2019). Autofocus: Efficient multi-scale inference. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9745–9755. Seoul, Korea.
31. Chandra, S., Kokkinos, I. (2016). Fast, exact and multi-scale inference for semantic image segmentation with deep Gaussian CRFs. *European Conference on Computer Vision*, pp. 402–418. Springer, Cham.
32. Kong, J., Wang, H., Wang, X., Jin, X., Fang, X. et al. (2021). Multi-stream hybrid architecture based on cross-level fusion strategy for fine-grained crop species recognition in precision agriculture. *Computers and Electronics in Agriculture, 185,* 106134. DOI 10.1016/j.compag.2021.106134.
33. Liu, R., Cai, W., Li, G., Ning, X., Jiang, Y. (2021). Hybrid dilated convolution guided feature filtering and enhancement strategy for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters* (in Press). DOI 10.1109/LGRS.2021.3100407.
34. Kingma, D. P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
35. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25,* 1097–1105. DOI 10.1145/3065386.
36. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. Munich, Germany.
37. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708. Honolulu, Hawaii.
38. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, Cham.
39. Chen, L. C., Papandreou, G., Schroff, F., Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.