ISSN: 2688-8645

MSBG-DeepLabV3+ Conveyor Belt Idler Roller Segmentation Algorithm

Guoxing Wang *

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan, 454000, China

* Corresponding author: Guoxing Wang

Abstract

Idler rollers play a crucial supporting and guiding role in the operation of conveyor belts, and their state changes significantly affect conveyor belt misalignment. Studies have found a significant difference in the exposed area of the left and right idler rollers under misalignment and non-misalignment conditions. Based on this characteristic, this paper focuses on idler rollers and proposes a semantic segmentation algorithm based on MSBG-DeepLabV3+ idler rollers to achieve high-precision segmentation of the idler roller region, providing technical support for subsequent research on conveyor belt misalignment. Building upon the original DeepLabV3+ framework, firstly, MobileNetV2 is used to replace the Xception backbone network, achieving model lightweighting. Secondly, a strip perception DenseASPP is introduced in the encoding stage to enhance the multi-scale contextual modeling capability of the slender directional structure of the idler rollers. Finally, in the decoding stage, a boundary enhancement module (BEM) and a global directional attention mechanism (GDAM) are combined to further improve the segmentation boundary accuracy and the perception capability of key regions and directional features. Experimental results show that the proposed MSBG-DeepLabV3+ model achieves an mIoU of 97.16% and an mPA of 98.62% on a self-made dataset. While ensuring high detection efficiency, this method achieves high-precision and high-efficiency segmentation of idler roller targets in complex industrial environments, verifying the effectiveness and practicality of the proposed algorithm.

Keywords

Conveyor Belt Misalignment; Idler Rollers; Semantic Segmentation; Backbone Network.

1. Introduction

Belt conveyor systems are widely used in heavy industries such as mining and ports, and their operational stability and safety have a significant impact on production efficiency and operational safety[1]. If the conveyor belt deviates during operation, it is highly susceptible to abnormal contact with structural components such as the frame, which not only accelerates belt wear and causes material spillage, but in severe cases, the high temperatures generated by friction can also trigger fires. Therefore, continuous and effective monitoring of the conveyor belt's operating status has become a crucial aspect of industrial safety management. Idler rollers, as important load-bearing and support components of belt conveyors, directly affect the stability and overall service life of the conveying system. However, due to the large number and wide distribution of idlers, traditional manual inspection methods are generally inefficient, lack real-time performance, and are highly subjective, making them insufficient to meet the demands of modern industry for safe equipment operation and intelligent maintenance[2]. Therefore, research on the automatic sensing and precise analysis of conveyor belt idler status in complex industrial scenarios, achieving reliable and high-precision idler detection, has

ISSN: 2688-8645

become an important foundation for conveyor belt fault diagnosis and condition-based maintenance.

Currently, monitoring the misalignment of conveyor belts mainly employs traditional image processing methods and deep learning-based methods. Traditional image processing methods typically use edge detection, threshold segmentation, and geometric feature extraction to identify the edge features of rollers, belts, or other components, thereby determining the conveyor belt's misalignment status. Yang et al.[3] proposed a machine vision-based conveyor belt misalignment detection method, extracting edges and laser center lines through an improved Canny operator and Hough transform, and combining corner information to achieve misalignment judgment. Sun et al. [4]proposed a dual-baseline localization method through ROI extraction and image preprocessing, improving Canny edge detection and combining Hough transform and least squares fitting to achieve conveyor belt misalignment identification. Wu et al.[5] enhanced image quality through high dynamic range imaging, combined Canny and Hough to extract straight line features, and achieved belt deflection identification. Wang et al.[6] used an improved HED algorithm to achieve accurate detection of belt edges, obtained single-pixel contours through edge refinement, combined with optical flow for camera vibration compensation, and finally completed misalignment identification by comparing with the labeled real edges. Liu et al.[7] used OpenCV to process industrial videos, extracting conveyor belt edges through image preprocessing, Canny edge detection, and Hough transform, and comparing the results with the idler roller baseline to determine the conveyor belt deviation status. Sun et al.[4] proposed a curve conveyor belt deviation assessment method based on ARIMA-LSTM, integrating machine vision and mechanical analysis to achieve deviation prediction and early warning. However, these methods are sensitive to interference factors such as lighting changes and image noise, and are prone to generating redundant or false edges under complex working conditions. When the image clarity is low or the background environment is complex, it is often difficult to stably and accurately extract the straight line features of the conveyor belt edge, resulting in a decrease in the accuracy of deviation detection.

With the rapid development of deep learning technology, many researchers have proposed deep learning-based methods for belt misalignment detection. These methods can be broadly categorized into: methods based on direct identification through detection or segmentation; methods based on 3D vision and spatial geometric modeling; and methods based on the fusion of deep learning and traditional vision. Wang et al.[8] proposed a GES-YOLO-based idler detection algorithm, introducing GSConv and EMA attention to achieve efficient and accurate idler positioning in low-light conditions. Li et al.[9] used the Unet model to segment the belt line and combined it with the straight-line position of the belt line extracted by probabilistic Hough transform to identify misalignment. An et al.[10] directly detected a group of idlers by improving RT-DETR, identifying misalignment based on the exposure degree of the left and right idlers. Zeng et al.[11] performed idler segmentation based on an improved DeeplabV3+, introducing MobileNetV3, GAM attention, and Lovasz loss to achieve conveyor belt misalignment discrimination. Ni et al.[12] detected idlers and conveyor belts based on an improved YOLOv8, introducing multi-head self-attention to enhance position feature modeling to achieve misalignment discrimination. Zhao et al.[13] proposed an online belt misalignment identification method based on binocular vision, utilizing binocular edge features and centerline distance to determine misalignment. Zhang et al.[14] fused a segmentation network with YOLOv5 to extract conveyor belt edge features and model diagonal information, achieving misalignment status determination and offset measurement. Wang et al.[15] segmented the conveyor belt edge using DeeplabV3+, extracted the centerline through morphological processing, and calculated the deviation distance, achieving full-position misalignment detection. While these methods have achieved certain results, the overall process is complex and heavily reliant on the conveyor belt itself and edge features, making them susceptible to

ISSN: 2688-8645

interference affecting detection accuracy in complex industrial environments. Furthermore, simplifying misalignment into a binary classification, while structurally simple, makes it difficult to obtain geometric information such as the misalignment direction, limiting model interpretability and engineering applicability, and resulting in insufficient generalization ability to changes in operating conditions.

Addressing the aforementioned issues, this paper focuses on the crucial role of idlers in conveyor belt operation, discovering a significant difference in the exposed area of the left and right idlers under conveyor belt deviation and non-deviation conditions. Based on this characteristic, this paper proposes an MSBG-DeepLabV3+ idler segmentation algorithm to achieve high-precision segmentation and detailed modeling of the idler region, thereby providing effective technical support for intelligent monitoring and safety management of conveyor belt operation. The main contributions of this paper are summarized as follows: Based on the original DeepLabV3+ framework, this paper replaces the Xception backbone with MobileNetV2 for lightweighting and employs the Strip Perception DenseASPP (SP-DenseASPP) module to enhance the contextual modeling capability of the slender directional structure of the idler; in the decoding stage, the Boundary Enhancement Module (BEM) is introduced to improve boundary segmentation accuracy, while the Global Directional Attention Mechanism (GDAM) is combined to strengthen the perception of key regions and directional features. The MSBG-DeepLabV3+ model achieves high-precision segmentation of idler targets in complex industrial environments.

2. Methodology

Belt conveyors play a crucial role in industrial production and equipment condition monitoring. With the continuous improvement of industrial intelligence, utilizing inspection robots to replace manual inspections and combining artificial intelligence technology to achieve automatic perception of conveyor belt operating status has become a current trend in research and engineering applications. In existing deep learning-based methods for conveyor belt deviation detection, most studies use object detection or semantic segmentation models to directly determine the deviation state. However, these methods struggle to effectively obtain the key geometric information needed for deviation direction, and the interpretability and engineering practicality of the models remain somewhat insufficient, with limited generalization ability under complex operating conditions. To overcome these shortcomings, some studies have attempted to combine deep learning methods with traditional visual analysis, typically relying on the detection and geometric analysis of the conveyor belt itself and its edge features. However, in actual industrial settings, factors such as changes in lighting, dust obstruction, and background interference can easily degrade the edge features of the conveyor belt, thus limiting further improvements in deviation detection accuracy.

To address the aforementioned issues, this paper proposes a conveyor belt idler segmentation method based on MSBG-DeepLabV3+. By performing high-precision pixel-level segmentation of the idler area, the proposed method can more accurately characterize the spatial distribution features of the idler, thereby improving the accuracy and stability of idler segmentation and providing reliable technical support for subsequent deviation geometry analysis.

The MSBG-DeepLabV3+ algorithm is a semantic segmentation model improved upon the original DeepLabV3+ model. DeepLabV3+ employs a typical Encoder–Decoder structure, where the Encoder consists of a backbone feature extraction network and a dilated spatial pyramid pooling module (ASPP). Xception is used by default as the backbone network to extract multi-level semantic features. The backbone network output includes high-level semantic features and intermediate low-level features. High-level semantic features are input to the ASPP module to enhance multi-scale contextual information, while low-level features are directly passed to

ISSN: 2688-8645

the Decoder for detail recovery. The ASPP module models contextual information through a multi-branch parallel structure, containing four dilated convolutional branches with different dilation rates and a global average pooling branch. The outputs of each branch are concatenated along the channel dimension and then fused using 1×1 convolutions before being input to the Decoder. In the Decoder, low-level features are first compressed along the channel dimension using 1×1 convolutions to reduce computational complexity and preserve spatial details. Subsequently, the ASPP output is upsampled to match the resolution of the low-level features and then concatenated along the channel dimension. The concatenated features are refined and fused through two 3×3 convolutional blocks, and finally, the output segmentation result is upsampled to the same size as the input image.

MSBG-DeepLabV3+ is a lightweight, high-precision semantic segmentation model improved upon the original DeepLabV3+ framework. This model replaces the Xception backbone network with MobileNetV2, effectively reducing the number of model parameters and computational complexity, thus achieving network lightweighting. In the encoding stage, SP-DenseASPP, which incorporates a strip pooling module (SPM)[16], is introduced to enhance the global multi-scale context modeling capability for the slender and directional structures of idler rollers. In the decoding stage, a boundary enhancement module (BEM) is introduced after fusing low-level and high-level features to strengthen boundary refinement before prediction, thereby improving segmentation boundary accuracy and overall mIoU. Furthermore, GDAM is introduced before high-level features enter the decoder to further enhance the network's ability to focus on key regions and directional structural features. Experiments show that this model can achieve high-precision segmentation of idler roller targets in complex industrial environments, and its overall structure is shown in Figure 1.

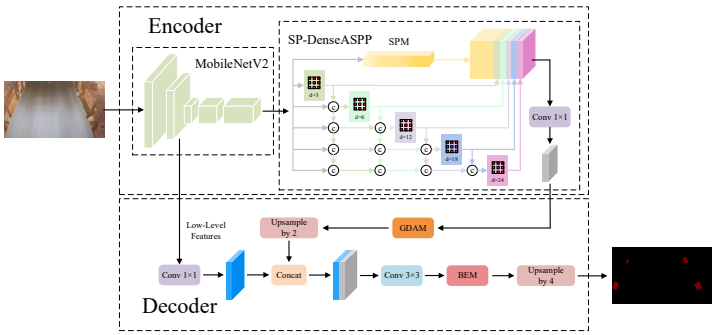

Figure 1. MSBG-DeepLabV3+ network structure

3. Network Structure and Improvement

3.1. MobileNetV2

MobileNetV2 is a high-efficiency, lightweight convolutional neural network proposed by Google. It aims to significantly reduce the number of model parameters and computational complexity while maintaining strong feature representation capabilities, and is widely used in resource-constrained scenarios such as mobile devices and embedded systems. Building upon MobileNetV1, this network introduces two key design features: an inverted residual structure and a linear bottleneck, which structurally improve the performance and stability of the lightweight network.

Overall, MobileNetV2 consists of standard convolutional layers and multiple stacked inverse residual modules. By progressively reducing the spatial resolution of the feature maps and increasing the number of channels, it achieves efficient extraction from low-level texture features to high-level semantic features. In the inverse residual module, 1×1 pointwise convolutions are first used to expand the channels of the input features to enhance feature

ISSN: 2688-8645

representation. Then, 3×3 depthwise separable convolutions are used to extract spatial features at a lower computational cost. Finally, 1×1 pointwise convolutions without nonlinear activations are used to map the features back to the low-dimensional space, thus avoiding information loss caused by nonlinear transformations of low-dimensional features. When the input and output feature sizes are the same, residual connections are introduced to enhance feature propagation and gradient propagation capabilities. By extensively employing depthwise separable convolutions, MobileNetV2 significantly reduces the number of parameters and computational complexity while maintaining good feature representation capabilities, thus improving the model's inference efficiency. Therefore, MobileNetV2 is very suitable as the backbone feature extraction network for dense prediction tasks such as semantic segmentation, and can meet the real-time and accuracy requirements of idler roller detection on industrial conveyor belts. The backbone network structure of MobileNetV2 in this paper is shown in Table 1 below.

Table 1. Three Scheme comparing

Model	expand	in_channel	out_channel	number	stride
Conv2d	—	3	32	1	2
InvertedResidual	1	32	16	1	1
InvertedResidual	6	16	24	2	2
InvertedResidual	6	24	32	3	2
InvertedResidual	6	32	64	4	1
InvertedResidual	6	64	96	3	1
InvertedResidual	6	96	160	3	1
InvertedResidual	6	160	320	1	1

3.2. SP-DenseASPP

In industrial conveyor belt monitoring scenarios, idler rollers typically exhibit structural features such as significant directionality and large scale variations, while also being susceptible to interference from coal dust obstruction, uneven lighting, and complex backgrounds. Although traditional ASPP (Automatic Spatial Component Pooling) can acquire rich contextual information through multi-scale dilated convolutions, its convolutional operations are spatially isotropic, limiting its ability to model slender targets like idler rollers with pronounced directional structures and failing to fully characterize their overall continuous morphology and long-range structural features. To address these issues, this paper proposes SP-DenseASPP, which introduces a strip pooling mechanism based on DenseASPP, to enhance the network's ability to model the directional structure and global contextual information of idler rollers.

SP-DenseASPP inherits the multi-branch, densely connected design philosophy of DenseASPP in its overall architecture. It consists of an original input branch, multiple dilated convolutional branches with different dilation rates, and a feature fusion layer. Each dilated convolutional branch uses a different dilation rate for feature extraction, and multi-scale features are fused stepwise through dense connections. This allows for the acquisition of multi-scale receptive fields without significantly increasing the number of parameters, effectively alleviating the semantic inconsistency problem between features of different scales, enhancing the ability to transmit contextual information, and enabling the network to simultaneously consider both the local details and the overall semantic structure of the idler roller.

Building upon this, SP-DenseASPP improves upon the original input branch of DenseASPP by introducing SPM (Spatial Pooling). This module effectively aggregates long-range spatial information of the roller along the main direction by performing global pooling along both the horizontal and vertical directions, enhancing the network's ability to perceive slender, directional structures. Subsequently, the obtained direction-aware features are fused with the

ISSN: 2688-8645

output of the multi-scale dilated convolution branch, thus forming a highly discriminative feature representation that combines multi-scale semantic information with directional structure perception capabilities. SP-DenseASPP is shown in Figure 2.

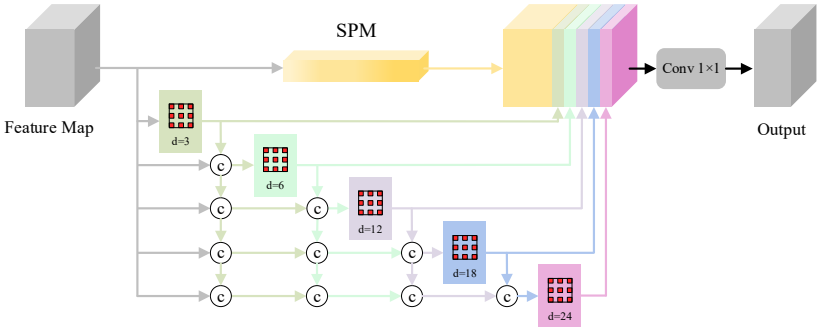

Figure 2. SP-DenseASPP

The strip pooling module takes the input feature map $X \in R^{C\times H\times W}$ as input and aims to enhance the network's ability to represent features of slender structural targets through multi-scale context aggregation and orientation-aware modeling. As shown in Figure 4-2, the module first uses two parallel 1×1 convolutions to perform channel compression and feature splitting on the input features, reducing the number of channels from C to C/4, and obtaining intermediate features X_1 and X_2 , so as to reduce computational complexity and provide dedicated feature representations for different branches. In the multi-scale context branch, standard 3×3 convolution and two different scales of adaptive average pooling operations are applied to the pooled features. After convolution mapping and upsampling to the original spatial resolution, the pooled features are fused with local features. This process can be represented as follows:

$$F_m = \phi(f_{3\times3}(X_1)) + u(f(p_s(X_1))), \tag{1}$$

Where $p_s(\cdot)$ represents adaptive pooling operations at different scales, $u(\cdot)$ is the upsampling operator, and $\phi(\cdot)$ is the nonlinear activation function, thus achieving effective aggregation of multi-scale contextual information. In the orientation-aware branch, the module performs global strip pooling on along both the horizontal and vertical directions, i.e.:

$$p_h(X_2) \in \mathrm{R}^{C/4\times1\times W}, \quad p_v(X_2) \in \mathrm{R}^{C/4\times H\times1}, \tag{2}$$

Furthermore, directional features are modeled using 1×3 and 3×1 asymmetric convolutions to enhance the network's ability to perceive long-range directional dependencies. Subsequently, multi-scale contextual features and direction-aware features are concatenated along the channel dimension and restored to the original channel dimension using 1×1 convolution. Finally, these are fused with the input features using a residual method, and the output features can be represented as:

$$Y = \sigma(X + f_{1\times1}([F_m, F_d])), \tag{3}$$

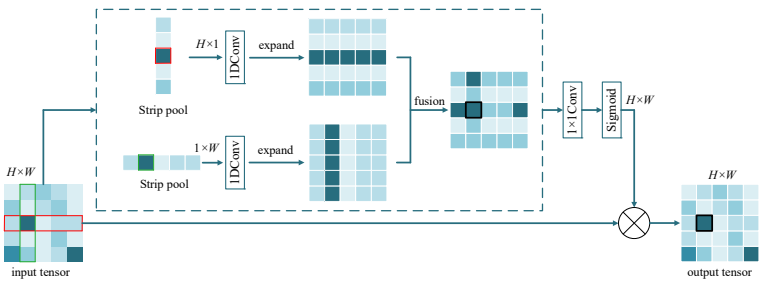

Figure 3. Strip Pooling Module Structure

ISSN: 2688-8645

Where F_d represents the orientation-aware feature and $\sigma(\cdot)$ is the ReLU activation function. Through the above design, SPM effectively enhances the network's ability to model the global structure of targets with obvious directional and scale-changing characteristics, such as idler rollers, while maintaining computational efficiency. The SPM structure is shown in Figure 3.

3.3. BEM

The boundary enhancement module aims to address the issue of weakened boundary information after high- and low-level feature fusion, thereby improving the boundary accuracy and structural integrity of idler segmentation in complex industrial conveyor belt environments. This module uses Depthwise Separable Convolution (DSC) as the basic computational unit, constructing local feature modeling branches and hole feature modeling branches in parallel to achieve joint modeling of contextual information at different scales. The local branch employs standard depthwise separable convolution for fine-grained feature modeling, focusing on capturing local texture and contour details of the idler edges; the hole branch introduces Dilated Depthwise Separable Convolution (DDSC) to expand the receptive field without increasing the number of additional parameters, enhancing the contextual consistency between the boundary region and the target object. Subsequently, the two feature paths are concatenated along the channel dimension, and adaptive feature fusion is achieved through 1×1 convolution, mitigating interference from differences in feature distribution at different scales. Finally, the module uses residual connections to add the fused features to the input features, ensuring stable gradient propagation while enhancing boundary response capabilities while maintaining the original semantic structure. Through the above design, BEM can effectively improve the boundary clarity and structural continuity of the idler roller segmentation results with lightweight computational overhead, and has good robustness to industrial scenarios such as coal dust obstruction, uneven lighting, and complex backgrounds. The BEM module is shown in Figure 4.

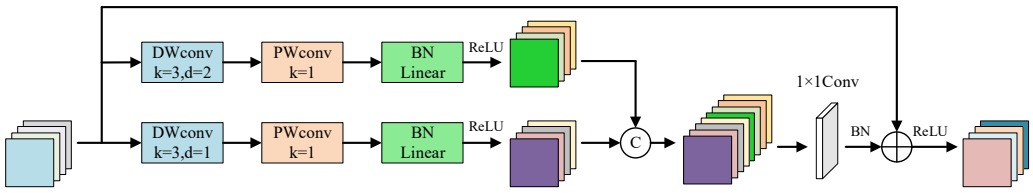

Figure 4. BEM structure

Dilated depth-separable convolution is a highly efficient convolution operator that organically combines dilated convolution and depth-separable convolution. Its core idea is to significantly reduce the number of parameters and computational complexity while expanding the effective receptive field of the convolution kernel by introducing a dilation rate, thereby enhancing the network's ability to model multi-scale contextual information and long-range dependencies. This operator first applies a depthwise convolution with a dilation rate of 2 and a kernel size of 3×3 to each channel of the input features, achieving the perception of a larger range of spatial information without introducing additional parameters or reducing the spatial resolution of the feature map. Subsequently, a 1×1 pointwise convolution is used to complete the fusion and re-encoding of cross-channel features. Compared with standard convolution operations, dilated depth-separable convolution significantly reduces computational overhead while maintaining strong feature representation capabilities. It is beneficial for depicting fine local structures and effectively improves the perception of mid- to long-range contextual information.

3.4. GDAM

To enhance the model's ability to perceive the directional structural features of idler rollers, this paper introduces a Global Directional Attention (GDAM) mechanism between the encoder

ISSN: 2688-8645

and decoder. This module jointly models high-level semantic features by fusing global channel attention and direction-aware spatial attention, suppressing background redundancy while strengthening the direction-related semantic expression of the idler roller region. This two-stage attention design improves the model's segmentation accuracy and robustness for idler roller targets in complex industrial environments. GDAM is a hybrid attention mechanism composed of a channel attention submodule and a direction-aware spatial attention submodule connected in series. It adopts a channel-first, then spatial attention modeling approach, progressively enhancing and recalibrating the input features. This module introduces strip convolution while maintaining low computational overhead to adapt to the modeling requirements of target features with obvious directional structures. The GDAM structure is shown in Figure 5.

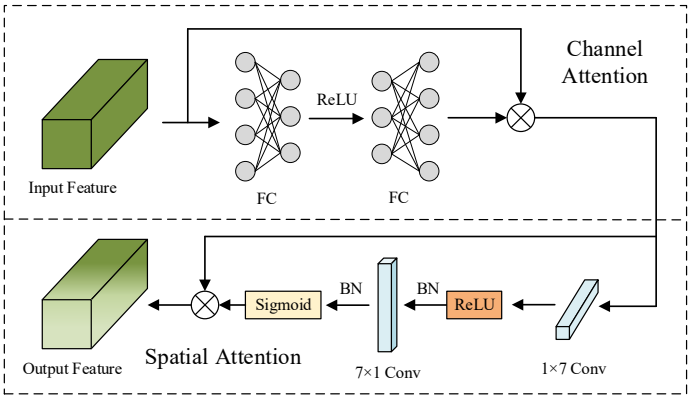

Figure 5. GDAM structure

4. Experiment

The experimental environment in this paper is the Ubuntu system, and the hardware and software configurations used in the experiment are shown in Table 2.

Table 2. Hardware and software configuration

name	version
GPU	NVIDIAGeForce RTX3090
Video memory	24G
Python	3.9
Pytorch	1.12.1
Cuda	11.3

4.1. Datasets and Evaluation Indicators

4.1.1. Datasets

The experimental training and performance evaluation in this paper are based on a self-built conveyor belt scenario dataset. The raw data was acquired in real-time by industrial cameras in a real industrial conveyor belt operating environment, covering multiple different industrial application scenarios. To reflect the complexity of actual working conditions, the data acquisition process comprehensively considered different operating environments, background conditions, and conveyor belt operating states, thus obtaining highly representative raw video data. In the data preprocessing stage, the acquired raw video data was first subjected to frame extraction, using an automated frame extraction tool to extract a large number of static images from continuous video. Subsequently, the extracted images underwent manual screening and quality control, removing blurry and severely occluded samples, retaining effective images with clear target structures and discriminative features.

ISSN: 2688-8645

Finally, a dataset containing 450 high-quality images was constructed, a portion of which is shown in Figure 6.

Figure 6. shows a partial dataset

Given the relatively limited number of usable images in the original dataset, this paper introduces various data augmentation strategies to expand the dataset and further mitigate the potential impact of insufficient sample size on model training stability and generalization performance. Without altering the semantic information and target structural features of the images, random horizontal flipping, random vertical flipping, random brightness adjustment, and random contrast transformation are applied to the original samples to effectively simulate conveyor belt scenes under different shooting angles, lighting conditions, and environmental changes. These data augmentation methods significantly improve the diversity and robustness of the dataset, enabling the model to more fully learn the discriminative features of the target during training. Ultimately, the dataset size is expanded from the original 450 images to 720 images, which are then used for model training and validation.

4.1.2. Evaluation Indicators

To evaluate the performance of the improved algorithm, this paper uses Mean Intersection Over Union (mIoU), Mean Pixel Accuracy (mPA), number of model parameters (Params), Floating Point Operations (FLOPs), and Frames Per Second (FPS) as evaluation metrics. mIoU measures the overlap between the model's predictions and the ground truth annotations. This metric, by averaging the intersection-over-union ratios (IoU) for each class, comprehensively reflects the overall segmentation performance of the model across different classes; a higher value indicates better segmentation. mPA measures the average pixel classification accuracy of the model across each class. This metric is calculated by averaging the pixel accuracy across all classes, with correctly predicted pixels out of the total number of true pixels in each class. The formulas for mIoU and mPA are as follows:

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{i=0}^{k}p_{ij}+\sum_{i=0}^{k}p_{ji}-p_{ii}}, \tag{4}$$

$$mPA = \frac{1}{k+1}\frac{p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}}, \tag{5}$$

In the formula, k+1 represents the total number of categories, P_{ii} represents the number of correctly predicted pixels, P_{ij} represents the number of pixels whose true category is i but are predicted as j, and P_{ji} represents the number of pixels whose true category is j but are predicted as i.

ISSN: 2688-8645

Params represents the total number of learnable parameters in the network, used to measure model size and its storage overhead; FLOPs represents the number of floating-point operations required by the model in one forward inference process, used to measure the model's computational complexity; FPS represents the number of image frames per second that the model can process in actual operation, used to measure the model's inference efficiency and real-time performance. Table 3 shows the parameter settings during model training.

Table 3. Model Training Experiment Parameter Settings

Parameter name	Value
Image resize	512×512
Batch size	4
Epoch	300
downsample_factor	8
Initial learning rate	7e-3
Minimum learning rate	(7e-3) * 0.01
optimizer_type	SGD
momentum	0.9
Weight decay rate	1e-4

4.2. Ablation Experiment

To verify the impact of different module improvements on model performance, ablation experiments were conducted on a self-made dataset under the same experimental environment, as shown in Table 4. The modules are represented by letters: A represents MobileNetV2; B represents SP-DenseASPP; C represents BEM; and D represents GDAM.

Table 4 shows that replacing the DeepLabV3+ model's backbone network with MobileNetV2 resulted in slight improvements in both mIoU and mPA. Simultaneously, the model's parameter count and floating-point computation cost significantly decreased, to 10.62% and 16.42% of the original model, respectively, while inference speed (FPS) was also greatly improved. This indicates that introducing a lightweight backbone network can effectively improve the overall detection efficiency of the model while maintaining segmentation accuracy. Introducing SP-DenseASPP based on the improved backbone network, although the model's FPS decreased slightly, improved mIoU by 0.41% and mPA by 1.02%. Furthermore, the model's

Table 4. Ablation Experiment Results

Model	A	B	C	D	mIoU/%	mPA%	Paras/M	FLOPs/G	FPS
DeepLabV3+					95.32	97.21	54.71	243.31	27.46
1	✓				95.44	97.23	5.81	39.96	83.96
2	✓	✓			95.85	98.25	4.36	33.75	74.54
3	✓		✓		96.13	98.08	6.08	44.38	79.12
4	✓			✓	96.38	98.07	6.08	41.04	76.43
5	✓	✓	✓		96.76	98.49	4.63	38.17	66.90
6	✓	✓		✓	96.75	98.34	4.62	34.83	65.05
7	✓		✓	✓	96.87	98.28	6.34	45.46	71.45
8	✓	✓	✓	✓	97.16	98.62	4.89	39.25	60.91

parameter count and floating-point computation cost were further reduced, indicating that this module enhances multi-scale feature representation capabilities while helping to further compress the model size, thereby improving the model's detection performance. Subsequently, introducing C (BEM) based on modules A and B, although the model's parameter count and computation cost increased to some extent, improved mIoU by 0.91% and mPA by 0.24%. This indicates that BEM, in synergy with the aforementioned improved modules, effectively

ISSN: 2688-8645

enhances the model's ability to model the features of target boundaries, thereby significantly improving segmentation accuracy. Finally, the introduction of the GDAM attention mechanism strengthens the model's ability to focus on key target regions, further improving mIoU by 0.40% and mPA by a slight increase.

In summary, the improved MSBG-DeepLabV3+ model achieves significant improvements in both segmentation performance and computational efficiency compared to the original DeepLabV3+ model. Specifically, mIoU is improved by 1.84% to 97.16%, and mPA is improved by 1.41% to 98.62%. Simultaneously, the number of model parameters is reduced by approximately 91.06%, floating-point operations are reduced by approximately 83.87%, and inference speed is improved by approximately 121.9%. Experimental results clearly demonstrate that these improvements work synergistically, not only enhancing the model's segmentation performance but also significantly reducing the number of parameters and floating-point operations, fully validating the effectiveness and practical value of the proposed method.

4.3. Comparative Experiment

To verify the performance of the proposed algorithm MSBG-DeepLabV3+ on a self-made dataset, the proposed method was compared with classic segmentation models such as U-Net[17], HRNetv2[18], PSPNet[19], and SegFormer[20]. The experimental results are shown in Table 5.

Table 5. Comparison results of different algorithms

Algorithm	mIoU/%	mPA%	Paras/M	FLOPs/G	FPS
DeepLabV3+	95.32	97.21	54.71	243.31	27.46
U-Net	97.48	98.84	24.89	225.84	41.56
HRNetv2	96.91	98.29	29.54	45.46	20.23
PSPNet	95.03	97.10	49.07	194.4	43.5
SegFormer	95.19	97.31	3.71	6.77	82.27
MSBG-DeepLabV3+	97.16	98.62	4.89	39.25	60.91

As shown in Table 5, the MSBG-DeepLabV3+ model proposed in this paper achieves the highest mIoU and mPA compared to mainstream segmentation models such as DeepLabV3+, HRNetV2, PSPNet, and SegFormer. Simultaneously, this model has the lowest parameter count and floating-point computation cost, and the highest inference speed (FPS), indicating that MSBG-DeepLabV3+ achieves excellent idler segmentation accuracy while possessing higher detection efficiency and real-time performance. Compared to U-Net, although the mIoU and mPA of the proposed method are slightly lower than U-Net by 0.32% and 0.22% respectively, both mIoU and mPA reach 97.16% and 98.62%, meeting the accuracy requirements for idler target segmentation in complex industrial conveyor belt environments. Furthermore, the proposed method has significantly lower parameter count and floating-point computation cost than U-Net, and outperforms U-Net in inference speed, demonstrating superior lightweight characteristics and real-time processing capabilities. Overall, the MSBG-DeepLabV3+ model proposed in this paper achieves a good balance between segmentation accuracy, model complexity, and inference speed. It not only demonstrates better performance advantages in comparison with various mainstream models, but also significantly reduces model size and computational overhead, and improves detection efficiency. Therefore, it is more suitable for real-time and stable segmentation of idler targets in industrial conveyor belt scenarios, providing reliable technical support for subsequent conveyor belt deviation detection.

4.4. Visualization of Results

To verify the segmentation effect of the proposed algorithm in real-world scenarios, this paper presents a visual comparative analysis of the image segmentation results and compares the

ISSN: 2688-8645

proposed method with various classic semantic segmentation models. Figure 7 shows the segmentation visualization results of different algorithms in three typical industrial conveyor belt scenarios. In the figure, columns (1), (2), and (3) correspond to three different scenarios; (a) is the original image, (b) is the manually annotated Ground Truth, and (c) to (h) are the segmentation results of different algorithms in the corresponding scenarios. As can be seen from Figure 7, the proposed method can accurately identify and finely segment the roller target in the image. For target areas with a small pixel ratio in the image, compared with other comparative methods, the proposed method performs better in terms of target integrity and boundary characterization, and can effectively reduce missed detections and boundary blurring. As shown in scenario (1) of Figure 7, DeepLabV3+ failed to segment the roller target, HRNetv2 had a deviation in target boundary segmentation, PSPNet and SegFormer's segmentation was incomplete, while U-Net and the proposed method could segment the target area relatively accurately. Under complex lighting conditions in scenario (3), the segmentation of the idler roller target near the light source area becomes more difficult. DeepLabV3+, HRNetV2, PSPNet, and SegFormer all failed to segment the target in this area, and the segmentation of the idler roller boundary was also relatively blurry. However, the method in this paper can still accurately identify and completely segment the idler roller target, demonstrating strong robustness. Although U-Net also achieved good segmentation results in this scenario, considering the detection efficiency and lightweight characteristics of the model, the method in this paper has more advantages in practical industrial applications.

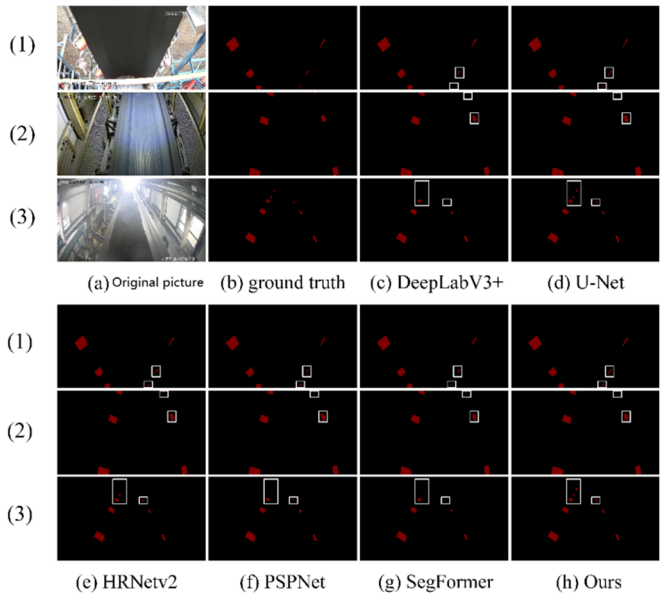

Figure 7. Visualization results of different algorithms

5. Conclusion

The intelligent development of belt conveyors is of great significance for monitoring conveyor belt misalignment in coal mines and industrial settings. Idler rollers, as key load-bearing and support components of belt conveyors, directly affect system stability through their operating status. This paper proposes an idler roller segmentation method based on MSBG-DeepLabV3+, achieving efficient and accurate idler roller segmentation and providing technical support for conveyor belt misalignment detection and analysis.

1. Based on the original DeepLabV3+ framework, this paper achieves model lightweighting by replacing the Xception backbone with MobileNetV2, and introduces the strip perception SP-DenseASPP module in the encoding stage to enhance the multi-scale context modeling

capability of the slender directional structure of the idler roller. In the decoding stage, the boundary enhancement module BEM and the global direction attention mechanism GDAM are combined to further strengthen the perception of key regions and directional features, thereby effectively improving the segmentation accuracy and robustness of the idler roller.

2. To more realistically reflect actual industrial conditions, this paper constructs a self-built dataset for conveyor belt scenarios. The original data was collected in real-time by industrial cameras in a real conveyor belt operating environment, covering various typical industrial application scenarios, resulting in 450 valid samples. By introducing data augmentation strategies to expand the scale of training data and improve the model's training stability and generalization ability, a final dataset of 720 samples was formed for model training.
3. Experimental results show that compared with mainstream semantic segmentation algorithms, MSBG-DeepLabV3+ has a lighter model structure, lower parameter count, and lower computational complexity. On the self-built transport belt dataset, this algorithm achieves better segmentation results while maintaining high segmentation accuracy, and significantly improves the model's inference speed, demonstrating stronger real-time performance and engineering application potential.

The complex environment of industrial transport belts, including lighting conditions and camera positioning, can easily affect image quality, thus limiting detection accuracy. Future work will focus on optimizing data acquisition strategies to improve the model's robustness and generalization ability in complex scenarios.

References

[1] ZHANG S, XIA X. Modeling and energy efficiency optimization of belt conveyors [J]. Applied energy, 2011, 88(9): 3061-71.

[2] WANG H, WANG H. Status and prospect of intelligent key technologies of belt conveyor [J]. Coal science and technology, 2022, 50(12): 225-39.

[3] YANG J, LI Z, GAO L, et al. Research on conveyor belt deviation detection method based on machine vision [J]. Journal of Physics: Conference Series, 2024, 2786(1).

[4] SUN X, WANG Y, MENG W. Evaluation System of Curved Conveyor Belt Deviation State Based on the ARIMA–LSTM Combined Prediction Model [J]. Machines, 2022, 10(11): 1042.

[5] WU X, WANG C, TIAN Z, et al. Research on Belt Deviation Fault Detection Technology of Belt Conveyors Based on Machine Vision [J]. Machines, 2023, 11(12): 1039.

[6] WANG Haoyu, WANG Xipeng. Machine vision-based mistracking detection of conveyor belts [J]. China Coal, 2024, 50(S2): 158-63.

[7] LIU Feng, BAI Jinniu. Vision-based method for detecting belt conveyor misalignment [J]. Shanxi Coal Coking Technology, 2023, 47(04): 25-8+31.

[8] WANG H, KOU Z, WANG Y. GES-YOLO: A Light-Weight and Efficient Method for Conveyor Belt Deviation Detection in Mining Environments [J]. Machines, 2025, 13(2): 126.

[9] LI Nanyan, MIAO Hui, ZHAO Long, et al. Intelligent Detection Method for Conveyor Belt Misalignment Based on Semantic Segmentation [J]. Technology Wind, 2025, (03): 59-61.

[10] AN Longhui, WANG Manli, ZHANG Changsen. Fault detection algorithm for underground conveyor belt deviation based on improved RT-DETR [J]. Journal of Mine Automation, 2025, 51(03): 54-62.

[11] ZENG F, FENG S, TAO Y, et al. Conveyor belt deviation detection method based on improved Deeplabv3+; proceedings of the Tenth International Conference on Mechanical Engineering, Materials, and Automation Technology (MMEAT 2024), F, 2024 [C]. SPIE.

[12] NI Y, CHENG H, HOU Y, et al. Study of conveyor belt deviation detection based on improved YOLOv8 algorithm [J]. Scientific Reports, 2024, 14(1): 26876.

[13] ZHAO Fangbing, XU Xuedong. Online Detection Method for Coal Conveyor Belt Misalignment Based on Binocular Vision Image Recognition [J].Mechanical Management and Development, 2024, 39(10): 329-31.

[14] ZHANG M, JIANG K, CAO Y, et al. A new paradigm for intelligent status detection of belt conveyors based on deep learning [J]. Measurement, 2023, 213: 112735.

[15] WANG Z, LI J, YANG X, et al. Automatic detection method of conveyor belt deviation based on DeepLabv3+; proceedings of the International Conference on Internet of Things and Machine Learning (IoTML 2022), F, 2023 [C]. SPIE.

[16] HOU Q, ZHANG L, CHENG M-M, et al. Strip pooling: Rethinking spatial pooling for scene parsing; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2020 [C].

[17] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation; proceedings of the International Conference on Medical image computing and computer-assisted intervention, F, 2015 [C]. Springer.

[18] WANG J, SUN K, CHENG T, et al. Deep high-resolution representation learning for visual recognition [J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(10): 3349-64.

[19] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2017 [C].

[20] XIE E, WANG W, YU Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers [J]. Advances in neural information processing systems, 2021, 34: 12077-90.