ISSN:2790-1688 Volume-15-(2025)

Research on image segmentation method of steel slag layer in RH vacuum degassing

Zhijie Yu, Tangyou Liu

College of Information Science and Technology, Donghua University, Shanghai, China

Abstract. In the RH vacuum degassing refining process, the measurement of the slag layer thickness is crucial. This paper provides a light-weight improved model on the basis of DeepLabv3+. Firstly, MobileNetV2 is used as the backbone network for feature extraction. This improves training efficiency and reduces model complexity. Secondly, regular convolutions in the ASPP module are replaced by depth-separable convolutions to increase the speed of the calculation. Finally, efficient multi-scale attention is introduced into high-level features to enhance segmentation accuracy. The experimental results show that, in comparison with the original DeepLabv3+, the, the proposed method reduces model parameters by 89.37%, increases mIoU by 1.89%, and balances model complexity and accuracy, demonstrating high practicality.

Keywords: Image segmentation; DeepLabv3+; Attention mechanisms; Lightweight.

1. Introduction

In the RH vacuum degassing [1,2] refining method, the dip tube is fixed above the ladle, and the ladle must be raised to an appropriate height to enable an effective degassing reaction. The dip tube needs to penetrate through the slag layer on the surface of the molten steel after the ladle is raised, allowing its lower end to come into contact with the molten steel. The measurement of the slag layer thickness is exploratory; the ladle is lifted into the upper dip tube, and after a period of contact, it is lowered. Due to the temperature difference between the molten steel layer and the slag layer, different colored marks are left on the outside of the dip tube. The thickness of the slag layer can be inferred from the area of contact with the slag. Traditionally, experienced workers visually observe this process, but manual observation not only lacks accuracy but also poses certain safety risks. Therefore, developing an accurate and efficient algorithm for slag layer segmentation is of great significance.

In the early stages, researchers primarily used traditional image processing and computer vision techniques for image segmentation. These methods typically classify pixels based on features such as texture, edges, and shapes. However, they exhibited limitations in segmentation accuracy, handling complex scenes, and generalization capability. With the rapid development of deep learning, convolutional neural network (CNN)[3] based image segmentation techniques have made significant advances. CNNs have increasingly been applied in the field of image segmentation by many researchers. Long et al. [4]pioneered the first end-to-end fully convolutional network (FCN) for pixel-level prediction. FCN achieved full-image segmentation by replacing fully connected layers with convolutional layers. Ronneberger et al.[5] proposed the U-Net network, which adopts a U-shaped structure with an encoder-decoder framework, introducing the concepts of upsampling and skip connections. In the decoder part, U-Net uses transposed convolution layers for upsampling, improving segmentation accuracy. However, by using skip connections, this model contains many parameters and calculations, making it rather large. Google's team introduced DeepLabv3+[6], which is regarded as a new milestone in image segmentation and has been widely applied across various domains.

Compared to traditional image processing methods, deep learning models are better at detection of slag areas, but they are often too complex for practical applications. To address the issues of high computational complexity and loss of detail during feature extraction in DeepLabV3+, this paper provides an improved DeepLabV3+ model based on MobileNetV2. Using the lightweight backbone feature extraction network MobileNetV2 and optimizing the ASPP module, the model's computational complexity is reduced. Additionally, a high-scale attention EMA module without

dimensionality reduction is introduced to generate better pixel-level attention for high-level features. Finally, the effectiveness of our method is demonstrated by theoretical and experimental results.

2. Methods

2.1 Overall architecture

DeepLabV3+ adopts an overall encoder-decoder architecture. In this paper we mainly improve the encoder part of the DeepLabV3+ network. We replace the original Xception backbone feature extraction network with the lightweight MobileNetV2, and substitute the 3x3 standard convolution in the ASPP module with depthwise separable convolution to reduce computational cost and model parameters, thereby meeting the real-time requirements of industrial applications. Additionally, the high-level features output from the MobileNetV2 backbone are processed through the efficient multi-scale attention (EMA) module, which provides more pixel-level attention to the feature maps extracted from the backbone. Figure 1 shows the structure of the enhanced DeepLabV3+ network.

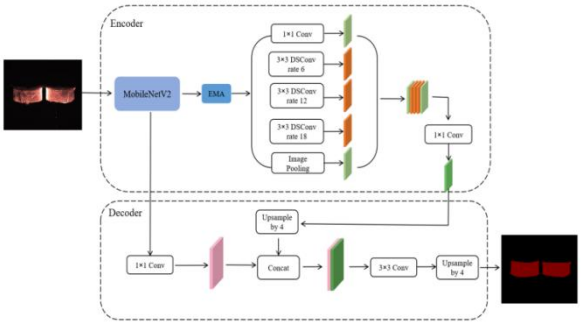

Figure.1 Improved DeeplabV3+ network

2.2 Lightweight backbone feature extraction network MobileNetV2

MobileNetV2 is a lightweight network model proposal from the Google team. Like the Xception network[7], it uses separable convolutions as its core operation. However, whereas Xception increases the number of parameters when using depth-separable meshes, MobileNet uses depth-separable meshes to compress the model and improve detection speed, making the model lighter.

The inverted residual structure[8] is the key innovation of MobileNetV2. In contrast to traditional residual networks, the inverted residual structure in MobileNetV2 begins with a 1×1 convolutional layer to increase the number of channels in the feature map, followed by a 3×3 depthwise separable convolution for feature extraction, and concludes with a linear bottleneck layer for dimensionality reduction. This inverted residual structure not only preserves the advantages of residual connections but also significantly reduces computational complexity. Figure 2 illustrates the structure of the residual block in the MobileNetV2 network.

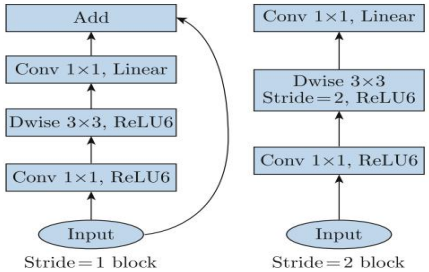

Figure.2 Residual block structure of MobileNetV2

2.3 Depthwise separable convolutions

The MobileNetV2 network reduces computational complexity through the use of depthwise separable convolutions. The ASPP module, commonly employed for multi-scale feature extraction, involves a substantial number of convolution operations, resulting in a significant computational load. To optimize the computational cost associated with the convolution operations in the ASPP module, we can also use depth-separable meshes to increase the training speed of the network.

Depthwise separable convolution is an efficient convolution operation that decomposes the traditional convolution process into two distinct steps: depthwise convolution and pointwise convolution. In depthwise convolution, each input channel is processed with an independent convolutional kernel, preserving spatial features, while pointwise convolution merges the feature maps obtained from depthwise convolution along the channel dimension. This separation not only effectively reduces the computational load and the number of parameters associated with convolutions but also maintains the ability to extract spatial information from the input features, as illustrated in Figure 3.

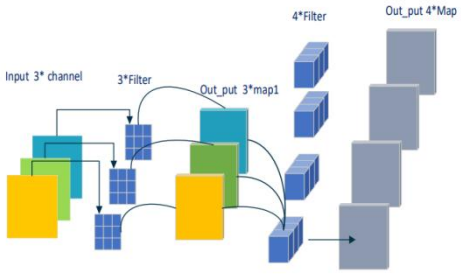

Figure.3 depthwise separable convolutions

Assuming the input feature map has a size of $D_K * D_K$ with M channels and N output channels, and a convolution kernel size of $D_K * D_K$, the computational cost for traditional convolution is $D_K * D_K * M * N * D_F * D_F$. In contrast, the computational costs for depthwise convolution and pointwise convolution are $D_K * D_K * M * D_F * D_F$ and $M * N * D_F * D_F$, respectively. From Equation (1), it can be derived that depthwise separable convolution has a computational cost of $\frac{1}{N}+\frac{1}{D_K^2}$ times that of traditional convolution, significantly reducing the computational and memory requirements of the network model while improving runtime speed.

$$\frac{D_K*D_K*M*D_F*D_F+M*N*D_F*D_F}{D_K*D_K*M*N*D_F*D_F}=\frac{1}{N}+\frac{1}{D_K^2} \tag{1}$$

2.4 Efficient multi-scale attention

The EMA attention mechanism module network [9,10] is illustrated in Figure 4. Initially, for any input $X \in R^{C\times H\times W}$, the EMA divides it along the channel dimension into G sub-features, represented as $X=[X_0, X_i, ..., X_{G-1}]$, $X \in R^{C\times H\times W}$, to capture various semantic information. Subsequently, attention weights for different feature maps are extracted through two parallel paths on the 1×1 branch and one path on the 3×3 branch. In the 1×1 branch, one-dimensional global average pooling operations in two different directions are employed to encode the channels, facilitating effective cross-channel information interaction. Conversely, the 3×3 branch omits the one-dimensional global average pooling operation and GroupNorm, aiming to achieve multi-scale feature representation. Following this, a two-dimensional global average pooling operation is applied to encode the global spatial information of the outputs from both the 1×1 and 3×3 branches. The formula for the two-dimensional global average pooling operation is as follows:

$$Z_C=\frac{1}{H\times W}\sum_j^H\sum_i^W X_C(i,j) \tag{2}$$

In Equation (2), Z_C represents the output value of the Cth channel after pooling, where H and W denote the spatial dimensions of the input features, and C signifies the number of channels. $X_C(i,j)$ indicates the input of the Cth channel at width i and height j. The output feature maps for each group are obtained by aggregating the two generated spatial attention weight values. Finally, the Sigmoid activation function is employed to capture pixel-level pairwise relationships, thereby acquiring global contextual information.

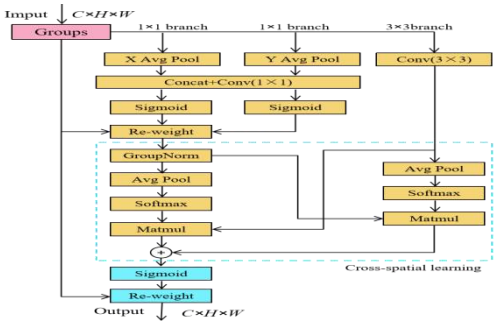

Figure4. EMA Network Architecture

3. Experimental results and analysis

3.1 Dataset

The dataset used in the experiment consists of image data collected from industrial cameras at a steel plant in HandanCity. Based on 500 original images, we performed horizontal flipping, contrast adjustment, brightness adjustment, and Gaussian blurring to increase the volume of training data. The samples were manually annotated using the Labelme tool and categorized into two main classes: immersion tubes and slag layer areas. Following the construction methodology of the Pascal VOC 2007 dataset, we created the RH slag layer image segmentation dataset. The dataset was split into training and testing sets in a 9:1 ratio, with an image resolution of 512×512 pixels.

3.2 Experimental environment and parameters

The operating system used in this experiment is Ubuntu 18.04, with an Intel(R) Xeon(R) Platinum 8255C CPU and an NVIDIA RTX 3080 GPU. The experiment utilizes version 1.7.1 of the PyTorch framework, with CUDA version 11.3. The training parameters are presented in Table 1.

Table.1 Training parameters

Keys	Values
Epoch	150
Batch_size	4
Init_lr	7e-3
Min_lr	7e-5
Optimizer_type	sgd
Momentum	0.9
Weight_decay	1e-4

3.3 Comparative analysis of experimental results

In order to evaluate the performance of the improved DeepLabv3+ model on RH slag layer image segmentation, multiple metrics were selected as criteria for model assessment. These metrics include mean Intersection over Union (mIoU), mean Pixel Accuracy (mPA), and the number of parameters. The calculation methods for mIoU and mPA are shown in formulas (3) and (4), where k+1 represents the number of classes (k object classes plus one background class). P_{ii} denotes the number of correctly classified pixels, P_{ij} indicates the number of pixels belonging to class i but

classified as class j, and P_{ji} represents the number of pixels belonging to class j but classified as class i.

$$mIOU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}+\sum_{j=0}^{k}P_{ji}-P_{ii}} \tag{3}$$

$$mPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}} \tag{4}$$

3.3.1 Ablation experiment

To ascertain the efficacy of the proposed enhancements, ablation experiments were conducted on the backbone feature extraction network and the attention mechanism. Keeping other experimental parameters constant, DeepLabv3+ served as the main framework, with MobileNetV2 and Xception used as backbone feature extraction networks, along with the implementation of an efficient multi-scale attention mechanism. The experimental results are presented in Table 2.

Table.2 Ablation experiment

Model	Backbone	Attention Mechanism	MIoU%	MPA%	Params/M
DeepLabv3+	Xception	-	88.57	92.38	208.7
DeepLabv3+	MobileNetV2	-	87.84	91.74	22.18
DeepLabv3+	MobileNetV2	EMA	90.46	94.90	22.24

Based on the experimental results presented in Table 2, replacing Xception with MobileNetV2 as the backbone feature extraction network led to a reduction in the model's parameter count by 89.37%, thereby decreasing model complexity. Additionally, while ensuring the model's lightweight nature, the incorporation of an efficient multi-scale attention mechanism resulted in improvements of 1.89% and 2.52% in mean Intersection over Union (mIoU) and mean Pixel Accuracy (mPA), respectively. The results suggest that the proposed approach reduces the complexity of the model and improves the accuracy of image segmentation.

3.3.2 Model comparison experiment

In order to verify the superiority of the improved method proposed in this paper, the enhanced DeepLabv3+ model was compared with classical semantic segmentation models, including FCN, U-Net, and the original DeepLabv3+ model. The experimental results are presented in Table 3.

Table.3 Model comparision experiment

Model	MIOU%	MPA%	Params/M
FCN	85.64	89.57	25.86
U-Net	90.87	95.43	94.95
DeepLabv3+	88.57	92.38	208.70
Ours	90.46	94.90	22.24

The results in Table 3 indicate that, compared to the FCN model using ResNet50 as the feature extraction network, our method achieves a 4.82% improvement in mean Intersection over Union (mIoU) while maintaining a similar model size. Additionally, when compared to the U-Net model using VGG as the feature extraction network, our method has only 23.42% of its parameter count, despite having comparable accuracy. These comparisons reveal that the improved DeepLabv3+ model strikes the best balance between complexity and accuracy.

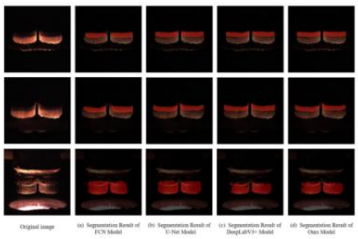

Figure.5 Comparison of segmentation results of different models

Figure 5 illustrates the segmentation results of different models. From the detection outcomes, it can be observed that our model exhibits higher segmentation accuracy in the thickness area of the steel slag layer compared to the FCN and DeepLabv3+ models. Furthermore, in terms of segmenting certain edge information, our model outperforms the U-Net model.

4. Summary

In the RH degassing refining process, image segmentation requires high accuracy and real-time performance. This paper proposes an improved method based on DeepLabv3+, replacing the backbone feature extraction network with MobileNetV2 to enhance training efficiency and reduce model complexity. Additionally, conventional convolutions in the ASPP module are substituted with depthwise separable convolutions to improve computation speed. A high-scale attention EMA module, which does not involve dimensionality reduction, is introduced to strengthen feature representation capabilities. Experiments have shown that it provides a good balance between model complexity and accuracy, and meets production requirements. Future work will focus on further optimizing the network to make it more lightweight while achieving more accurate predictions and enhanced real-time capabilities, thereby advancing the construction of intelligent steelmaking.

Reference

[1] Pope, M. C., P. Norbury, and J. Nicholson. "Developments in RH Vacuum Degassing at BSC General Steel's Scunthorpe Works." In International Conference Secondary Metallurgy (ICS'87).(Preprints), pp. 220-230. 1987.

[2] Wahlster, M. and Reichel, H.H., 1967. Some Technical and Metallurgical Aspects of the Application of the RH Process. Steel Times, 195(5179), p.459.

[3] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", Proc. Adv. Neural Inf. Process. Syst., vol. 25, pp. 1097-1105, Sep. 2012.

[4] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3431-3440, Jun. 2015.

[5] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation", Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., pp. 234-241, 2015.

[6] L. Chen, G. Papandreou, I. Kokkinos et al., "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets Atrous Convolution and Fully Connected CRFs", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, pp. 834-848, 2018.

[7] K. Shaheed, A. Mao, I. Qureshi, M. Kumar, S. Hussain, I. Ullah, et al., "DS-CNN: A pre-trained xception model based on depth-wise separable convolutional neural network for finger vein recognition", Expert Syst. Appl., vol. 191, Apr. 2022.

[8] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov and L. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks Conference on Computer Vision and Pattern Recognition, pp. 4510-4520, 2018.

[9] D.L. Ouyang, S. He, G.Z. Zhang, M.Z. Luo, H.Y. Guo et al., "Efficient Multi-Scale Attention Module with Cross-Spatial Learning", 2023 IEEE International Conference on Acoustics Speech and Signal Processing. Rhodes Island., pp. 1-5, 2023.

[10] A G Howard, M Zhu, B Chen et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]", arXiv preprint, 2017.