Industry Science and Engineering Vol. 1 No. 6, 2024

MCL_NET: Multi-Scale Collaborative Learning Network for Pellet Facies Microstructure Image Segmentation

Chun Zhang[1,2], Yanbing Liang[2*]
[1]*Hebei Key Laboratory of Data Science and Application, North China University of Science and Technology, Tangshan, Hebei, China*
[2]*College of Science, North China University of Science and Technology, Tangshan, Hebei, China*
**Corresponding Author*

Abstract: The segmentation model based on traditional convolution focuses more on local features during training, while Transformers excel in long-range modeling tasks but lack direct advantages in short-range feature analysis. To achieve fine-grained segmentation of pellet microstructure images more effectively, we propose a Multi-Scale Collaborative Learning Network for Pellet facies microstructure image Segmentation (MCL_NET). The model integrates information from different scales to achieve pixel recovery and precise boundary segmentation of pellet microstructure images. In the segmentation task of pellet ore microstructure images, MCL-NET demonstrates superior experimental results, allowing for a more accurate representation of the shapes and boundaries of the pellet microstructures.

Keywords: Microstructure Image Segmentation; Encoding and Decoding Structure; Attention Mechanism; Feature Cross Complementary; Pyramid Structure

1. Introduction

With the accelerated research in deep learning and the rapid advancement of computational power, utilizing Convolutional Neural Networks (CNN) for deep semantic information extraction in images has become a significant focus in the field of computer vision. the Fully Convolutional Network (FCN)[1] is designed with a fully convolutional structure to accomplish pixel-level image classification tasks. the U-Net series models employ a symmetrical encoder-decoder structure, enhancing the segmentation performance at material boundaries [2]. the Deeplab series, developed by Google's team, aims to provide more precise boundary segmentation solutions [3–5]. However, due to the limited receptive field of convolution, these models struggle to establish long-range spatial dependencies and lack the capability to perceive global information in images.

Transformer has gained widespread attention in the field of NLP for its ability to capture long-range dependencies in sequences [6]. Many researchers have extended the Transformer to the computer vision domain. By dividing images into patches and mapping them into sequences, it enhances the spatial modeling capabilities of images [7, 8]. Howeverthe Transformer introduces substantial computational complexity. To reduce computational costs, image resolution needs to be progressively lowered during training, which is a key reason for the loss of some texture information in images.

Pellets have uniform particle size, high iron content, high strength, and good reducibility, making them high-quality raw materials for blast furnace ironmaking [9]. the segmentation and quantitative characterization of pellet microstructure images are crucial steps in studying the mechanisms affecting pellet performance. However, the complex distribution of foreground and background in microstructure images, along with the similar material structures, uneven spatial distribution, and class imbalance, pose challenges. Achieving accurate segmentation of pellet microstructure images, especially the precise delineation of microstructural boundaries, is of great significance.

We propose model named Multi-Scale Collaborative Learning Network for Pellet Facies Microstructure Image Segmentation (MCL-NET). In the model's encoder, we employ a dual-backbone assisted fusion encoding strategy, combining the local feature

extraction capability of convolution with the global modeling strength of Transformer. This approach enhances attention to the target in both spatial and channel dimensions. Additionally, a bidirectional pyramid structure is used to model global context and multi-scale feature information, which serves as supplementary features in the decoding phase. During decoding, shallow image detail textures are progressively fused with deep abstract features to achieve pixel-level classification of pellet microstructure images.

2. Related Work

Semantic segmentation is a fundamental task in computer vision, and the introduction of U-Net has highlighted the potential of skip connection structures in preserving sensitive detail information. However, in convolution-based encoder architectures, a single convolution operation can only capture local pixel relationships within the size of the convolutional kernel. To obtain global features, continuous down-sampling operations are required, which inevitably lead to the loss of some information.

The Vision Transformer can utilize richer global information early in the encoding phase, fundamentally overcoming the limitations of convolution. the Swin Transformer uses a pure Transformer structure for feature extraction and has achieved excellent results across various computer vision tasks. TransUNet [10] is an innovative segmentation model where both convolution and Transformer work together in the encoding phase. To address the limitations of convolution in local feature extraction and the computational complexity of the Transformer, TransUNet uses features extracted by convolution as input for the Transformer to model contextual feature relationships during the encoding phase.

3. Methods

3.1 MCL-NET Model Structure

This paper presents a multi-scale collaborative learning model for pellet microstructure image segmentation, referred to as MCL-NET. the model structure is illustrated in **Figure 1**.

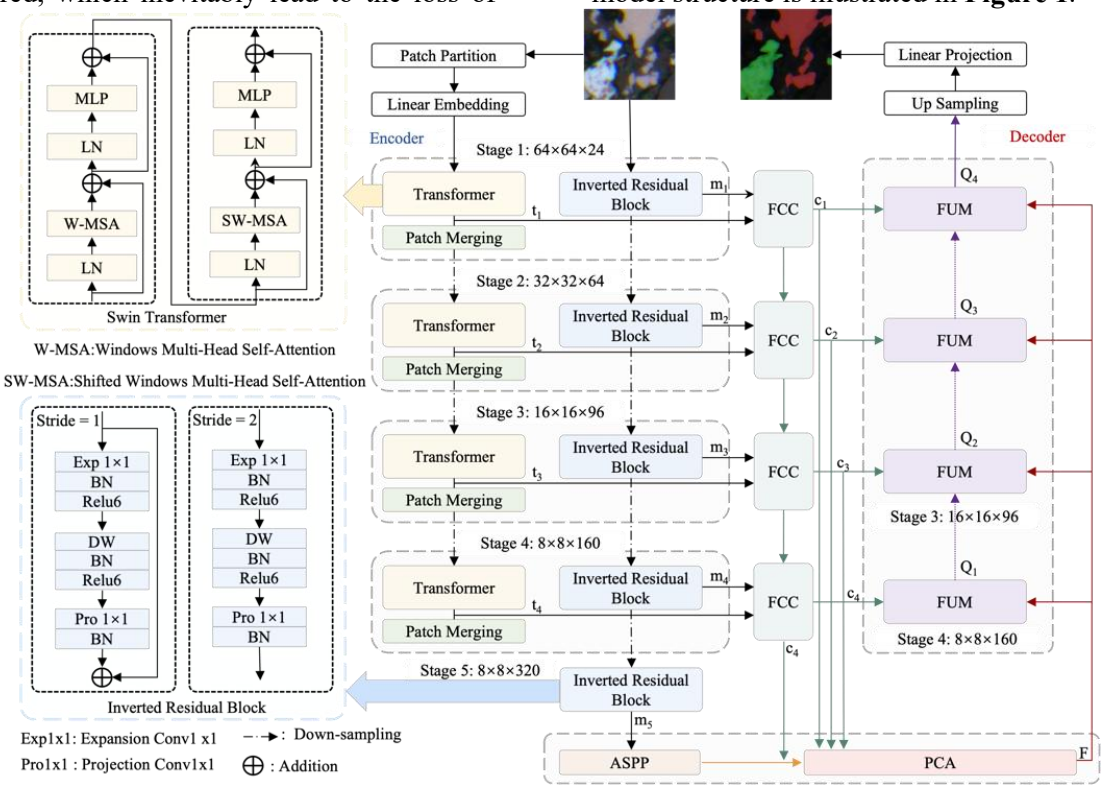

Figure 1. Structure of the MCL-NET Model

The model is primarily composed of three modules. In the encoding phase, the FCC module and DPFA module are employed to complement local features and spatial features, as well as to supplement contextual and multi-scale features. In the decoding phase, the FUM module utilizes shallow features as supplementary information to achieve accurate image segmentation.

MCL-NET employs a parallel backbone

feature extraction encoder consisting of the convolutional structure Mobilenetv2[11] with a reverse residual structure and the Swin Transformer. the Mobilenetv2 architecture introduces the reverse residual structure and linear bottleneck layers. the Swin Transformer incorporates a sliding window attention mechanism. the reverse residual structure utilizes depthwise separable convolutions for feature extraction, while replacing the original nonlinear activation layer with Relu6 to perform linear activation transformations. the Swin Transformer adopts a hierarchical approach and sliding window mechanism. In the encoder, the Swin Transformer is divided into four stages based on the resolution, with feature resolutions of 1/2, 1/4, 1/8, and 1/16 of the input image resolution, respectively.

3.2 Feature Cross-Compensation Module

To address the complex spatial information distribution and the requirements for fine segmentation of local edges in pellets, this model incorporates a feature cross-compensation module. This module enhances the model's attention to complex image boundaries and textures, and its structure is illustrated in **Figure 2**.

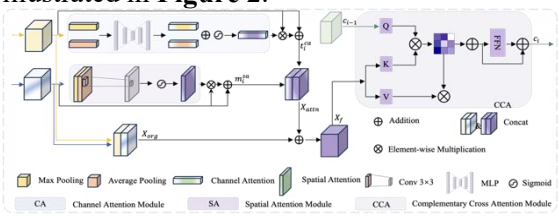

Figure 2. Structure of the FCC

This module first receives the features obtained from the encoder. It employs spatial attention and channel attention to enhance the feature extraction capabilities of both convolution and Transformer in the spatial and channel dimensions. the input results before and after applying attention are concatenated, and a residual structure is used to maintain feature integrity. To recover the representation of detailed features in the image, a new Complementary Cross Attention (CCA) mechanism is introduced in this module. This mechanism uses low-level spatial information to re-weight the higher-level pixel positions, guiding the higher-level feature learning towards more important and finer targets.

3.3 Dual-direction Pyramid Feature Aggregation

The model includes a horizontal Atrous Spatial Pyramid Pooling (ASPP) and a vertical Pyramid Feature Aggregation (PFA) module to achieve cross-scale feature aggregation. Additionally, grouped convolutions and channel shuffling are utilized for improved feature extraction. Finally, attention calibration is performed on the channel and spatial dimensions for each feature group, with a residual connection to reduce feature loss. the PFA module facilitates cross-scale feature aggregation, as illustrated in **Figure 3**.

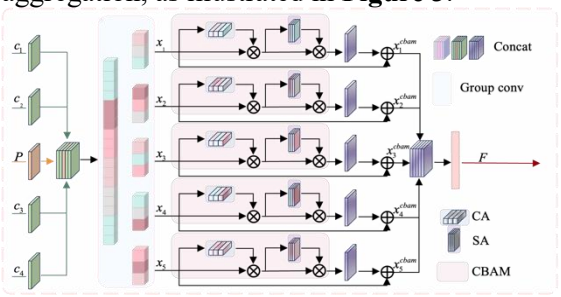

Figure 3. Structure of the PFA

This module receives the output results from the FCC module at different stages and the output from the ASPP module as inputs. It aggregates feature information from different receptive fields and scales, utilizing channel shuffling and grouped convolutions to achieve information fusion across various scales. Attention calibration is performed on each group of features in both the channel and spatial dimensions, and a residual connection is used to reduce feature loss. Ultimately, multiple groups of features are fused to achieve normalization and grouped weighting, resulting in the fused features F. the above process is defined as follows.

$$x_i = GConv\left(Concat_{\xi}\left(P, c_1, c_2, c_3, c_4\right)\right)$$

$$x_i^{ca} = x_i \otimes CA\left(x_i\right) + x_i$$

$$x_i^{cbam} = x_i^{ca} \otimes SA\left(x_i^{ca}\right) + x_i^{ca}$$

$$F = Sof\,t\max\left(BN\left(x_1^{cbam} \oplus, \cdots, \oplus x_5^{cbam}\right)\right)$$

Here, $Concat_{\xi}(\cdot)$ represents the feature fusion and channel shuffling operations, while $GConv(\cdot)$ denotes the grouped convolution. $\otimes$ and $\oplus$ represent element-wise multiplication and addition operations, respectively. $BN(\cdot)$ indicates batch normalization, and $Sof\,tmax(\cdot)$ is used as the nonlinear activation function.

4. Experiments and Results Analysis

4.1 Dataset and Experimental Setup

First, the experimental pellets were collected,

and the pellets were sliced to prepare thin sections for observation. Polarizing microscopy was used to examine the thin section samples, and image data were collected. A slider was used to cut the image samples, and Labelme was employed to annotate the image data. A total of 17, 100 training images and 4, 500 validation images were obtained.

We build the MCL-NET model based on the Python 3.9 environment and the PyTorch framework, training the model on an Nvidia RTX 3090 GPU with 24 GB of memory. During training, the learning rate is set to 0.007 and linearly decays with the number of iterations. the model is trained using the SGD optimizer with a momentum of 0.9, with a batch size of 4 for 700 epochs.

4.2 Experiment Result

To validate the performance of the model in the pellet microstructure image segmentation task, we compared the MCL-NET model with six other outstanding semantic segmentation models. We use Pixel Accuracy (PA), Mean Pixel Accuracy (MPA), and Mean Intersection over Union (MIoU) to assess the segmentation accuracy of the model. To validate the performance of the model algorithm, we also introduce three additional metrics: the number of parameters (Params), the number of floating-point operations (FLOPs), and the frames per second (FPS). The experimental results are presented in **Table 1**.

Table 1 Test Result Comparison of Different Networks in Pellet Microstructure Images

Methods	PA	MPA	MIoU	Params	FLOPs/G	FPS
U-Net [2]	90.23	86.64	77.43	**35.89**	**3.314**	20.2
Deeplabv3+[5]	90.46	88.16	78.22	37.22	4.729	19.3
TransUNet [10]	91.18	87.93	78.57	39.48	3.482	19.5
Swin-Unet [12]	90.88	88.07	77.89	42.68	4.745	20.5
MISS-Former [13]	92.54	88.36	79.76	40.58	4.903	**21.2**
MCL-NET	**93.14**	**89.31**	**81.56**	41.34	4.158	20.9

From Table 1, it can be seen that for the pellet microstructure image segmentation task, the MCL-NET model improves the PA by 0.6 to 2.91 percentage points. the MPA is improved by 0.95 to 2.67 percentage points, and the Mean Intersection over Union MIoU is enhanced by 1.8 to 7.01 percentage points.

This model has a large number of parameters, however, compared to TransUNet, Swin-Unet, and MISSFormer, MCL-NET demonstrates better performance in terms of FLOPs. the FPS of MCL-NET is 0.3 lower than that of the optimal model.

Five images (a) to (e) were randomly selected from the validation set to compare the original images, the segmentation results of seven models in the pellet microstructure image segmentation task, and the annotated images. the segmentation results are shown in Figure 4.

By observing the results of different models in the pellet microstructure image segmentation task shown in **Figure 4**, MCL-NET demonstrates superior performance in complex spatial segmentation tasks. It achieves more precise segmentation of small objects even in cases of extreme spatial distribution imbalance and data label imbalance. Additionally, MCL-NET provides more refined processing of the image's edge textures in the microstructural material.

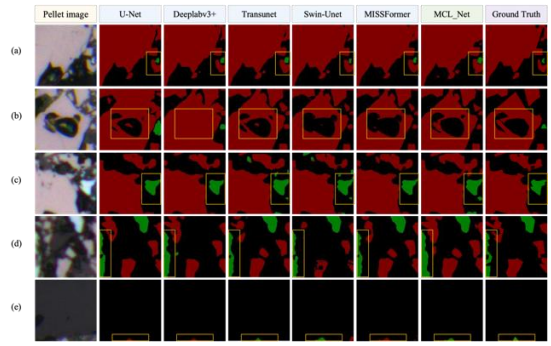

Figure 4. Segmentation Visualization Result Examples for Pellet Microstructure

5. Conclusion

This paper presents a multi-scale collaborative learning model for pellet microstructure image segmentation, designed based on an encoder-decoder architecture that integrates the advantages of dual backbone feature extraction and multi-scale feature information. the encoder employs the FCC module to enhance the feature extraction capabilities of the backbone network, while a new attention mechanism improves the model's focus on complex regions. the design of the bidirectional pyramid structure enables the model to achieve better spatial modeling capabilities. Experimental results show that the

model significantly improves segmentation accuracy for material boundaries and small particle substances.

Reference

[1] LONG J, SHELHAMER E, DARRELL T. Fully convolu-tional networks for semantic segmenta-tion [C]//Proceedings of the IEEE conference on com-puter vision and pattern recognition. 2015:3431-3440.

[2] RONNEBERGER O, FISCHER P, BROX T. U-net: Convo-lutional networks for biomedical image segmenta-tion [C]//Medical image computing and comput-er-assisted intervention–MICCAI 2015:18th interna-tional conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Pub-lishing, 2015:234-241.

[3] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs [J]. IEEE transactions on pattern analysis and machine in-telligence, 2017, 40(4):834-848.

[4] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Re-thinking atrous convolution for semantic image seg-mentation [J]. arXiv preprint arXiv:1706.05587, 2017.

[5] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encod-er-decoder with atrous separable convolution for se-mantic image segmentation [C]//Proceedings of the Eu-ropean conference on computer vision (ECCV). 2018:801-818.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information pro-cessing systems, 2017, 30.

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020.

[8] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted win-dows [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021:10012-10022.

[9] CASTELLANOS R M, IGLESIAS J C Á, GOMES O F M, et al. Characterization of iron ore pellets by multimodal microscopy and image analysis [J]. REM-International Engineering Journal, 2018, 71:209-215.

[10] CHEN J, LU Y, YU Q, et al. Transunet: Transformers make strong encoders for medical image segmentation [J]. arXiv preprint arXiv:2102.04306, 2021.

[11] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottle-necks [C]//Proceedings of the IEEE conference on com-puter vision and pattern recognition. 2018:4510-4520.

[12] CAO H, WANG Y, CHEN J, et al. Swin-unet: Unet-like pure transformer for medical image segmenta-tion [C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022:205-218.

[13] HUANG X, DENG Z, LI D, et al. Missformer: An effective medical image segmentation transformer [J]. arXiv pre-print arXiv:2109.07162, 2021.