IMPROVING DEEP LEARNING-BASED INSTANCE SEGMENTATION FOR CRYSTALLIZATION TRAILS WITH RESIDUAL MULTI-SCALE FEATURE AND ATTENTION MECHANISM

Xiaoyan ZHANG[1], Bo LAN[2,*], Lugang ZHANG[3]

The process of monitoring crystallization is significant for understanding the formation of crystals and assisting in the screening of high-throughput crystal forms. Crystal process image recognition is faced with small contours, dense targets, and scale changes. Therefore, this paper proposes a network model for the segmentation of crystallographic images. Firstly, the model's generalization ability is improved by data augmentation methods such as contrast adjustment and Gaussian blur. Afterwards, a residual feature enhancement module is inserted between the encoder and decoder to strengthen the feature extraction for the top layer of the feature pyramid network, effectively utilizing high-level information and enhancing multi-scale feature extraction. Lastly, attention gates are applied for automatic learning to focus on targets of varying sizes and shapes, emphasizing valuable features while suppressing irrelevant areas in the input image. Experimental outcomes indicate that the optimized approach achieves an IoU of 0.8521, an F1-score of 0.9185, and a sensitivity of 0.9012, outperforming other methods. The method can meet the requirements of image segmentation in the crystallization process and provide a reliable reference for the automatic process of crystal form screening.

Keywords: crystal, instance segmentation, transfer learning, multi-scale feature, attention mechanism.

1. Introduction

The development cycle of new drugs is long and expensive. Thousands of compounds often need to be screened. The screening of crystal forms is an essential part of the whole process. Crystallization is an operation that separates substances from solutions in a crystalline state. It can deal with many problems that operations cannot solve, including distillation, extraction, and adsorption,

[1] School of Information Engineering, Beijing Institute of Petrochemical Technology, Beijing China, e-mail: zhangxiaoyan@bipt.edu.cn
[2] School of Information Engineering, Beijing Institute of Petrochemical Technology, Beijing China, e-mail: lanbo@bipt.edu.cn
[3] School of Information Engineering, Beijing Institute of Petrochemical Technology, Beijing China, e-mail: zhanglugang@bipt.edu.cn

which is widely used to separate new products [1]. Assessing crystal morphology and size distribution is critical for managing and optimizing product quality and production efficiency in crystallization. Various process analysis tools, including attenuated total reflection infrared spectroscopy, Raman spectroscopy, focused beam reflectometry, and ultrasonic attenuation, have been utilized in the crystallization process. These process analysis techniques help to monitor the growth behavior during the crystallization process, which can control the crystallization process and develop its product quality [2].

In contrast, image signals represent more intuitive information and are regarded as one of the most promising techniques for measuring crystal form and size [3]. The changes in crystal morphology and imaging size during crystallization are monitored using visual sensing devices. A variety of parameters such as crystal morphological characteristics, growth rate, and growth state are obtained during the crystallization process [4]. It is possible to further study the crystal growth mechanism, identify the crystal habit ratio, and provide an adequate basis for regulating the crystal growth process. To investigate the crystallization results, it is still necessary for researchers to observe the samples through a microscope in most cases. Crystals are manually found to determine if the crystallization process is complete. In high-throughput crystal form screening, manual crystal identification has become a bottleneck factor affecting the experimental progress and automation.

Conventional methods in image processing have allowed for the calculation of crystal form and distribution of dimensions while crystals are forming, utilizing visual imaging systems [5]. Vancleef introduced an adaptable and dependable image analysis approach that merges edge detection, intensity thresholding, and an advanced watershed algorithm for monitoring particle size distribution, quantity, and solid concentration via a flow microscope [6]. Offiler created a new automated technique that merges a flow cell for crystal growth with image analysis, using the Hough transform to identify crystal facets and calculate growth rates, which was confirmed through experiments with α-glycine crystals [7]. Additionally, the advancement in computer science, particularly deep learning within machine learning, has significantly progressed, enhancing image characteristic information extraction and recognition accuracy through big data.

Deep learning has found applications in monitoring the crystallization process in pharmaceutical and chemical fields. Huo employed a binocular system to measure the distribution and size of crystals based on the fusion of binocular images captured by two miniature cameras at different angles [8]. Vagenknecht suggested an approach using generative adversarial networks to analyze particle size in low-resolution online microscope images, allowing for real-time monitoring of particle size distribution in crystallization processes [9]. Manee suggested a strategy for controlling crystallization that combines a measurement

sensor based on a convolutional neural network (CNN) with a reinforcement learning (RL) framework. This strategy aims to improve control loops during crystallization, improve the quality of crystal size measurements, and decrease the time required for image processing [15]. Salami created an image analysis model using deep learning and convolutional neural networks (CNN) to detect unwanted crystals in real-time during crystallization, with the goal of improving product purity and process efficiency in the production of Cephalexin [16]. Xin introduced two techniques to enhance Mask R-CNN: merged sampling and random proposal augmentation, which improve training on small datasets and reduce overfitting. These techniques were applied to the measurement of zeolite catalyst particles in SEM images, significantly increasing measurement accuracy [17]. Fan presented a new method for in-situ measurement using binocular stereo imaging to track the distribution of crystal length and width while cooling crystallization occurs. This method utilizes a stereo vision imaging system combined with deep learning algorithms, enhancing the accuracy and efficiency of crystal size distribution measurements [18].

The proposed method offers several key advancements in crystallization image analysis, distinguishing it from existing approaches. Traditional deep learning models face challenges with high-density targets, small contours, and scale variations in crystal images. To address these issues, our study introduces the following specialized optimizations:

(1) Custom-designed Data Augmentation: To combat data imbalance and prevent model overfitting, tailored augmentation techniques are implemented, ensuring robustness across various scenarios.

(2 Residual Feature Enhancement Module: Added at the top layer of the Feature Pyramid Network (FPN), this module improves multi-scale feature extraction, which is particularly beneficial for detecting and segmenting small crystals that are often overlooked by traditional models.

(3) Attention Mechanism in Feature Fusion: This mechanism helps refine the learning of relevant features while suppressing noise, leading to more accurate and context-aware feature representation.

(4) Emphasis on Skip Connections: By focusing on skip connections within the instance segmentation network, the model restores full spatial resolution and reduces semantic gaps between the encoder and decoder, which enhances segmentation performance.

Together, these targeted optimizations contribute to superior segmentation accuracy and adaptability, setting this method apart from existing solutions in the analysis of crystallization images.

2. Related Work

In recent years, deep learning-based models have gained significant attention for their application in the analysis of crystallization processes, particularly in monitoring and controlling crystal size, morphology, and distribution in real-time. These models have been developed to tackle challenges such as high-density crystal slurries, overlapping particles, and high-throughput image analysis, all of which are commonly encountered in industrial crystallization processes.

Several studies have advanced the state of crystallization image analysis through the use of deep learning. Li proposed a deep learning-based strategy that effectively analyzes in-situ microscopy images of high-density crystallization slurries by incorporating image and data augmentation techniques. This method enhances the ability to monitor and control the crystallization process in real-time, enabling better management of the crystal growth dynamics and product quality[10]. Wang introduced a Nonlinear Model Predictive Control (NMPC) approach aimed at real-time control of crystal size and standard deviation during cooling crystallization processes. This approach improves the precision of crystallization control, ensuring more reliable and consistent crystal formation[11]. Zong explored deep learning techniques, particularly the Mask R-CNN-based online image analysis, to address the challenge of accurately segmenting images of high solid concentration and overlapping particles in continuous industrial crystallization processes. By improving segmentation accuracy, this method contributes to better analysis of crystallization behavior in large-scale settings[12]. He leveraged deep learning for the analysis of how microporous plate sizes influence the crystalline growth of active pharmaceutical ingredients (APIs). The integration of microscopy imaging techniques allowed for the successful high-throughput analysis of large volumes of crystal data, providing insights into the crystallization process[13]. Wu developed an advanced image analysis method that integrates the S2A-Net model, a state-of-the-art deep learning architecture for object detection. This method enables real-time, accurate determination of crystal size distribution and quantity, significantly aiding in the observation and analysis of crystallization processes such as taurine crystallization[14].

Additionally, several well-established deep learning architectures have been applied to image segmentation tasks in crystallization analysis. U-Net, originally designed for biomedical image segmentation, has proven effective in leveraging limited annotated samples through data augmentation strategies. This allows for end-to-end training on small datasets, which is particularly useful in crystallization image analysis where data may be limited[19]. FCN-8s, another popular model, facilitates efficient semantic segmentation by employing fully

convolutional architectures. This allows for the transformation of arbitrarily sized inputs into correspondingly dimensioned outputs, making it suitable for various crystallization image sizes[30]. SegNet, with its deep fully convolutional encoder-decoder architecture, enhances segmentation accuracy and efficiency by using pooling indices for nonlinear upsampling between the encoder and decoder, which improves the model's ability to handle complex image features such as overlapping crystals and varying particle sizes[28].

These advancements in deep learning-based models have contributed significantly to the field of crystallization process analysis. However, challenges remain in dealing with complex images, overlapping particles, and high-density slurries, indicating that further innovations in model design, feature extraction, and data augmentation are required to fully optimize the analysis of crystallization processes.

3. Model Design

The research employs a comprehensive deep learning approach to segment crystallization pictures. The fundamental model framework adopts an encoder-decoder structure for feature extraction and image segmentation. Utilizing the U-net structure as its backbone, the model integrates a FPN and an attention mechanism for the segmentation task. Moreover, the encoder generates characteristics for FPN input at various scales, with a residual feature enhancement module implemented at the highest level of the FPN. The FPN output, in conjunction with the decoder through the attention gate, generates high-level semantic features as well as fine-grained image features. Finally, a mask is output to signify the crystallization. Fig. 1 illustrates the visual structure of the optimized model.

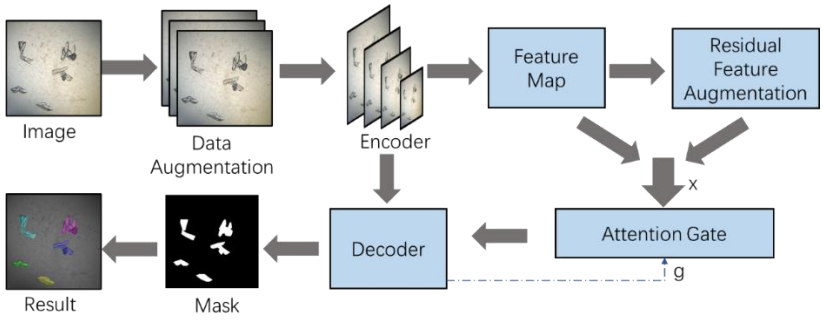

Fig. 1. The structure of the optimized model

3.1. Basic network structure

The basic network structure utilizes the U-net design as its core, consisting of both an encoder and a decoder [19]. The encoder facilitates the propagation of

contextual information through skip connections, thereby enhancing the extraction of complex hierarchical features. The decoder assimilates features of varying complexity for the reconstruction process. The U-net's encoder and decoder are connected with long-range connections, allowing for the integration of various hierarchical features into the decoder, leading to a network that is both more precise and scalable.

The U-net's encoder structure is a conventional convolutional network architecture. In the encoder, every stage includes two 3x3 convolutions with mish activation, then a 2x2 max pooling layer is used for downsampling. The feature channels are doubled at each step. In the decoder, the feature maps undergo up-sampling via 2×2 convolutions at each stage. The encoder's feature maps from each step are combined and incorporated into the decoder. Subsequently, mish activation is performed after two 3×3 convolutions. Lastly, a 1x1 convolution is utilized on the feature map to decrease it to the required channel quantity, resulting in segmented images.

3.2. Residual feature augmentation

Throughout the process of crystallization, crystals undergo continuous growth and change in both shape and size. Some subtle information in crystallization imaging is often missed. Feature pyramid can be applied to fuse multi-scale features. The feature pyramid to the convolutional neural network enables the extracted features to better represent the multi-dimensional information of input images. In the FPN, the feature information at the top level is lost due to the reduction of feature channels [20]. The information provided is limited to a single scale context and does not align with features at other levels. In this paper, the residual feature augmentation module [21] is applied to adaptively pool the top-level information of FPN, as shown in Fig. 2. Different contextual information is extracted and the loss is reduced at the highest level in the feature pyramid using a residual manner.

Within the FPN framework, the higher-level feature maps are transmitted in a downward direction and slowly combined with the lower-level ones. On the one hand, low-level feature maps are improved by high-level semantic information, so that features are naturally assigned to different contextual information. On the other hand, the top layer contains only single-scale contextual information incompatible with other features and suffers the loss of information due to reduced feature channels.

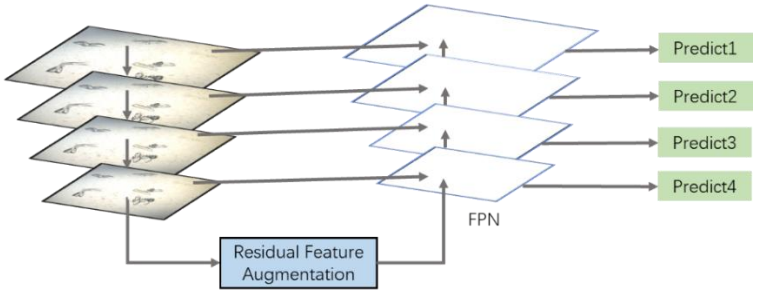

Fig. 2. The FPN with residual feature augmentation module

Fig. 3 displays the overall framework of the residual feature augmentation module. At the highest level, the module conducts adaptive pooling that is invariant to ratios in order to acquire feature maps of various scales. 1x1 convolution is applied to modify the feature maps, while bilinear interpolation is used to up-sample all feature maps of varying scales to match the size of the original top layer. Contextual features are combined using adaptive spatial fusion, rather than simply adding them adaptively. The adaptive spatial fusion module takes the up-sampled features as input, generates a spatial weight for each feature, and aggregates the contextual features using the weight. The features developed by the adaptive feature fusion module are combined with the top layer of the pyramid by summation, and the feature fusion is performed by propagation to other layers.

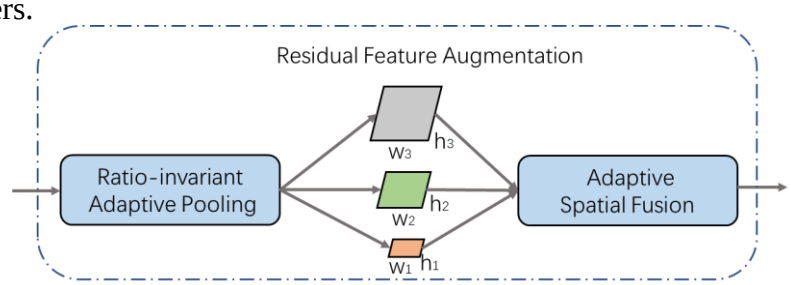

Fig. 3. The overall framework of the residual feature augmentation module

3.3. Attention mechanism

The attention mechanism is recognized as a successful technique for improving the efficiency of CNNs. It emulates the biological process of observation by suppressing redundant information and emphasizing intricate details of the desired target. Vaswani demonstrated dependencies on machine translation input by leveraging self-attention [22]. Simultaneously, the attention mechanism has been integrated into computer vision. Wang and Lu [23,24] utilized spatial attention for image captioning and classification, while Fu [25] employed a dual attention mechanism to capture global features in semantic segmentation. The attention mechanism is used in different digital image

segmentation tasks to achieve better results. Generally, attention modules can augment existing CNN models. They use spatial regions and channel interrelations to assist CNNs in concentrating on the target's more pragmatic features.

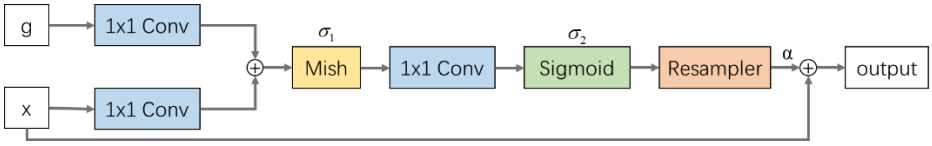

Fig. 4.The structure of the attention gate

The attention gate is an attention mechanism that can automatically focus on object regions and suppress responses in irrelevant regions. Its structure is shown in Fig. 4. The attention gate can be described as:

$$q_{att}^{l} = \psi^{T}(\sigma_1(W_x^T x_i^L + W_g^T g_i + b_g)) + b_\psi \tag{1}$$

$$\alpha_i^l = \sigma_2(q_{att}^l(x_i^l, g_i; \Theta att)) \tag{2}$$

The Mish activation is denoted by when x is the input feature and g is the gating signal. The term represents the sigmoid function:

$$\sigma_2(x) = \frac{1}{1+\exp(-x)} \tag{3}$$

At first, g and x are subjected to a parallel 1×1 Conv operation, and then the resulting characteristics are combined. Following this, a series of Mish activation, 1×1 Convolution, and Sigmoid function calculations are performed to generate the attention coefficient α by resampling. In the end, the input encoding matrix x is multiplied by the attention coefficient α to yield the ultimate output. Attention gates demand minimal computational power and a limited number of model parameters. They can augment the sensitivity and precision of dense label prediction models.

3.4. Mish activation function

The activation function plays a critical role in the training and evaluation of deep neural networks and increases the nonlinearity of the model in the neural network. Swish, Leaky ReLU, Sigmoid, and ReLU are the widely used activation functions. The method in this paper uses the Mish activation function. In terms of challenging datasets, it works better than Swish and ReLU. Additionally, the simplicity of Mish enables its smooth implementation in neural network.

Mish is a non-monotonic smooth neural network activation function that is defined by a specific formula.

$$f(x) = x \cdot tanh(\omega(x)) \tag{4}$$

The ω(x) is defined as $\ln(1+e^x)$. A graph of the Mish activation function is shown in Fig. 5.

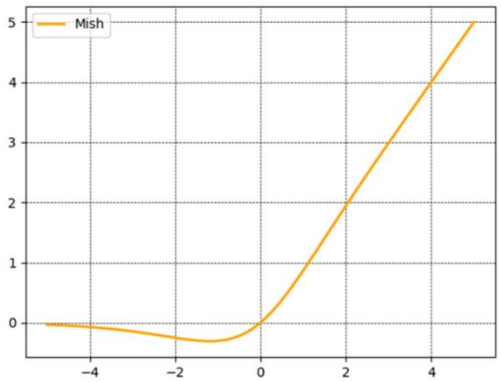

Fig. 5. The graphical of Mish activation function

A gating function is implemented by Mish, in which the input to the gate is a scalar. The gating features helps replace activation functions such as ReLU. Because the input to the gating function is a scalar, there is no need to modify the network parameters.

3.5. Loss function

The purpose of model training is to increase the ability of the model to identify different classifications. In order to accomplish this, a loss function of weighted binary cross-entropy is utilized. For the weighted binary cross-entropy implementation, positive pixels are weighted by the ratio of positive to negative voxels in the training group. The size of the negative class in the crystallized image is relatively larger than that of the positive class. Therefore, the weight can be adjusted so that the network is not biased towards a specific class when training.

The formula for the loss of weighted binary cross-entropy is as shown below:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[\omega_p y_i \log\left(\hat{y}_i\right) + \left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right] \tag{5}$$

Where y_i is the actual label (0 or 1), $\hat{y}_i$ is the prediction of the model, ω_p is the positive prediction weight, and *N* is the number of samples.

4. Results and Analysis

4.1. Data source

The datasets used in this paper come from two sources. One sample is sourced from the MARCO dataset [26] for machine recognition of crystallization

outcomes, while the other was collected in our own laboratory. The MARCO is a dataset marked for the classification of crystallization images. The dataset needs to be re-annotated when it is applied to the segmentation of crystallization images. Adding the laboratory data to the MARCO dataset can enhance the algorithm's performance during segmentation. It is more beneficial to improve the generalization ability under different conditions.

To improve the recognition ability of the model, it is firstly pre-trained on the COCO dataset, which is a large dataset proposed by Microsoft for object detection or image segmentation. Pre-training the model on the COCO dataset enables it to acquire fundamental image feature extraction capabilities. After that, the dataset is trained using the pre-trained model. The re-annotated images are shown in Fig. 6.

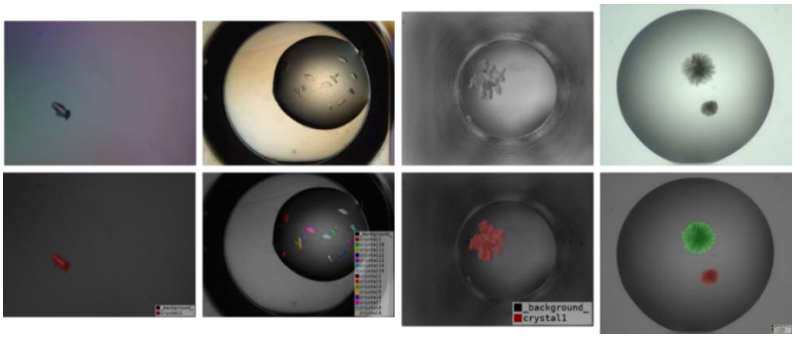

Fig. 6. The re-annotated images of the dataset

4.2. Data augment

The training images are extended using the data augmentation method to avoid the recognition bias and model overfitting caused by the small number of samples in the dataset. Data augmentation is a technique to enrich the number of samples in the dataset by modifying existing samples or generating new synthetic data. In image segmentation applications, the most commonly used data augmentation techniques include adjusting brightness or contrast, zooming in/out, cropping, rotating, noise or flipping. In addition to the above methods, the following methods are used in this paper. In view of the problem that the illumination intensity of crystallization images collected in practical applications may vary due to the different environments, image brightness enhancement or reduction is used to simulate the light change in practical applications. There may be some blur and jitter in the actual captured image, and the Gaussian blur normal distribution method is used to process images. The augmented data are shown in Fig. 7.

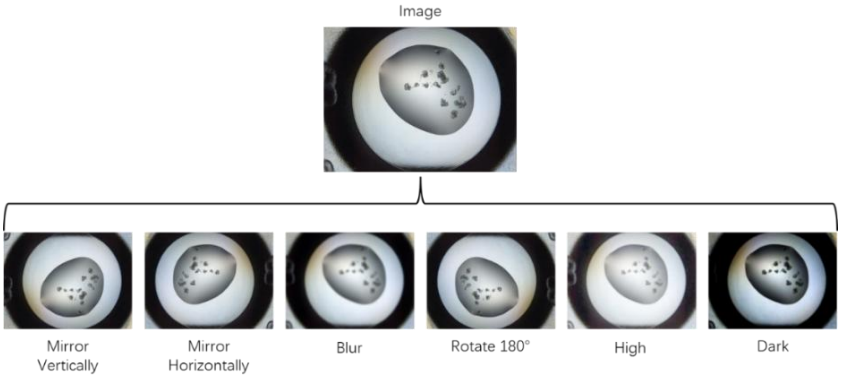

Fig. 7. Original image and augment image

4.3. Network training

Using the PyTorch deep learning platform, the study in this article was carried out on a 64-bit Ubuntu 16.04 LTS system.The computer hardware configuration comprises 32 GB of memory, an Intel Xeon E5-2680 CPU, and an RTX3090 GPU.

Model training utilizes the K-fold cross-validation method [27], which partitions the raw training dataset into k subsets, followed by k iterations of model training and validation. Cross-validation maximizes dataset efficiency, guaranteeing that the assessment outcomes accurately represent the model's effectiveness on the test dataset.

In the phase of training the model, the Adam optimization algorithm is set up with these parameters: a decay rate of 1e-6, an initial learning rate of 0.0001, $\beta\_1 = 0.99$, and $\beta\_2 = 0.999$. Moreover, an early stopping training strategy [28] is implemented to prevent overfitting during model training. In the testing phase, the model inference results are visually presented, with the segmentation outcomes overlaid on the crystallized image.

4.4. Evaluation metrics

For an image segmentation model, a proper assessment of its performance is crucial. This section presents several well-established and commonly employed evaluation metrics for image segmentation. The metrics derived from the confusion matrix pertain to false negatives (FN), false positives (FP), true negatives (TN), and true positives (TP).

The IoU is calculated by dividing the intersection of two regions by the total area they cover. It is used to measure the accuracy of the prediction compared to the actual data.

Defined in terms of the confusion matrix variables, it is:

$$IoU = \frac{TP}{TP + FN + FP} \tag{6}$$

The ratio of correctly predicted samples to the total number is the accuracy. These samples usually refer to pixels or voxels in object detection or image segmentation. In practical applications, accuracy is rarely used alone due to the uneven distribution of different categories in the dataset, and it generally needs to be used in conjunction with other metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Precision is defined as the proportion of true positive samples among all positive predictions. Likewise, the specificity is defined as the proportion of accurately predicted negative instances to all negative predictions. Both precision and specificity help assess the number of false positive pixels in the image.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

Recall is the proportion of positive samples correctly identified. Evaluating a model or algorithm is often done by balancing precision and recall.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

The F1-score, often utilized to assess the overall effectiveness of a model, is calculated as the harmonic mean of precision and recall.

$$F1\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{11}$$

4.5 Experiment results and analysis

The instance segmentation results of our network are shown in Fig. 8. It can be seen that even with many precipitates, most protein crystals are still recognized by the network. The results of instance segmentation are more consistent with the shape of protein crystals.

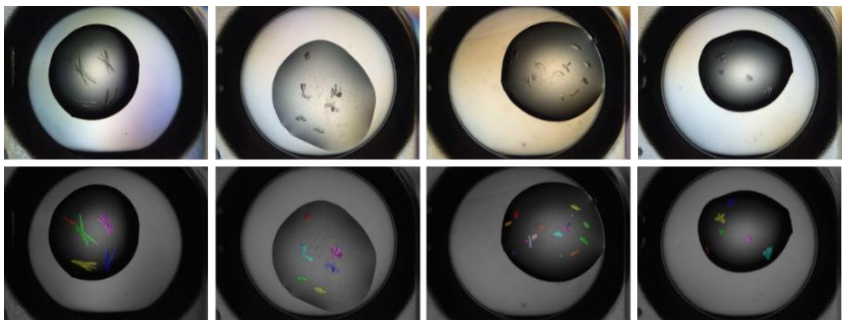

Fig. 8. Test results of the optimized network

Table 1

Experiment results and comparison against other networks.

Network	IoU	F1-score	Sensitivity
U-net[19]	0.8123	0.8989	0.8683
FCN-8s[30]	0.6879	0.8137	0.7043
SegNet[28]	0.8063	0.8939	0.8413
Reference [10]	0.8201	0.9005	0.8702
Reference[11]	0.8354	0.9053	0.8857
Reference[12]	0.8256	0.9021	0.8725
Reference[13]	0.8407	0.9104	0.8903
Reference[14]	0.8309	0.9087	0.8806
Our method	0.8521	0.9185	0.9012

In Table 1, there is a quantitative comparison of our network with other networks in the dataset. In terms of performance, our network outperforms all other models across each evaluation metric. Reference [13] produced the most reliable results among the compared networks but still trails our method by 1.14% in Intersection over Union (IoU), 0.81% in F1-score, and 1.09% in sensitivity. This suggests that while it excels in integrating microscopy imaging techniques for high-throughput analysis, our method's architecture is better suited for processing complex crystallization images, especially in handling high-density slurries and overlapping particles.

References [11] and [14], which innovate in real-time control and observation, respectively, showed commendable performance but still fell short of the overall efficacy of our method. On the other hand, the FCN-8s model [30] underperformed, likely due to its inability to effectively handle the complexity and diversity inherent in crystallization images. Its fully convolutional architecture may not capture detailed features of the crystallization process adequately. Additionally, while U-Net [19] and SegNet [28] are both successful in biomedical image segmentation, they may require further optimization for the specific challenges presented by crystallization images.

Through the analysis of the table, it can be found that different network models have different performances in crystallization image segmentation. This difference in performance can be explained by the application of the attention module, leading to a more refined policy for generally better segmentation results. At the same time, the multi-scale architecture samples different granularities of the crystallization images. Basic characteristics offer greater clarity and include additional location and specific details, yet they lack depth and include more interference. High-level characteristics may have lower resolution, yet contain valuable semantic details. Effectively combining the two features will significantly improve the segmentation task. These differences suggest that attention and multi-scale mechanisms can improve the performance of segmentation networks.

In the table, showing the difference in performance by quantitative evaluation is not enough to fully demonstrate the strength of the model. Therefore, in Fig. 9, some intuitive comparison examples of several different methods are also given for crystallographic image segmentation. The proposed network that combines attention and multi-scale mechanisms achieves better results than other segmentation networks.The visual outcomes indicate that the suggested technique is more effective in isolating the finer elements within images.

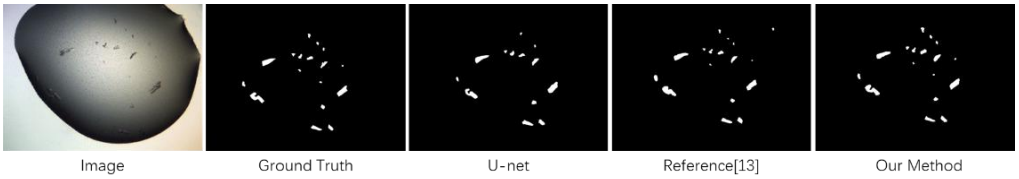

Fig. 9. The comparison examples of several different methods

5. Discussion

The deep learning model proposed in this article achieved significant results in the image segmentation task for crystal crystallization. To enhance the model's practicality and extend its application, future research should delve into and improve upon the following key areas:

1) Integration of Transformer Models: Transformers have shown excellent performance in natural language processing, with their capability to capture long-distance dependencies. Considering images as sequences to some extent, future work could apply Transformer models to image segmentation. The self-attention mechanism within Transformers could process global contextual information in images, potentially improving the model's ability to recognize long-range features in crystal images, especially with varying crystal sizes and shapes.

2) Model Reduction and Optimization: To fit edge computing environments, the model needs to be streamlined and optimized. Techniques such as model pruning can remove unimportant weights to reduce the number of model parameters without compromising performance. Weight quantization can decrease memory usage and speed up inference, making the model more suitable for devices with limited computing resources. Knowledge distillation can transfer knowledge from large, complex models to smaller ones, improving efficiency while maintaining high performance levels.

3) Enhancement of Real-time Performance: Real-time processing capability is crucial for industrial applications. Future work should focus on optimizing the model's inference speed through algorithmic optimization, parallel computing, and hardware acceleration. Algorithmic optimization can be achieved by improving network structures and reducing computational load; parallel computing can exploit the parallel processing capabilities of GPUs or TPUs;

hardware acceleration may involve custom hardware designs like FPGAs or ASICs to further enhance running speed.

4) Robustness Enhancement: Although the model performs well on specific datasets, its robustness under different conditions needs strengthening. Future work will involve more stress testing and exploring new data augmentation techniques to improve the model's adaptability to various lighting conditions, noise, and crystal types. Also, introducing adversarial training and regularization techniques can effectively enhance model robustness. Adversarial training trains models with adversarial samples to generalize better to unseen samples, while regularization techniques prevent overfitting by adding regular terms to the loss function.

Testing Model Generalization: Assess the model's generalization by testing it across different datasets and real-world scenarios to evaluate its performance with unseen crystal types and crystallization conditions. To achieve this, it is necessary to gather and create a wider range of data sets, such as crystal pictures taken from various sources and environments, along with more extensive cross-validation. Furthermore, techniques like transfer learning could further improve model generalization.

Through in-depth research in these directions, we anticipate further enhancements in model performance, enabling a greater role in practical industrial applications. These studies will also contribute to technological progress in the field of image segmentation.

6. Conclusions

This study successfully introduced a deep learning-based model for crystal crystallization trajectory image segmentation, significantly enhancing performance by incorporating a residual multi-scale feature enhancement module and attention mechanism. The model outperformed existing methods in key metrics such as IoU, F1-score, and sensitivity, offering a novel solution for image segmentation tasks in crystallization processes. This accomplishment is important for the efficient screening of crystal shapes in the pharmaceutical and chemical sectors, which could lead to improved automation of crystal screening and better quality control and production efficiency.

Moreover, the model demonstrated advantages in handling complex image characteristics like small contours, dense targets, and scale variations by employing data augmentation and attention mechanisms, providing fresh perspectives for applying deep learning in similar complex image analysis tasks in the future. The model's generalizability and robustness could still be improved, and its adaptability to different crystal types and crystallization processes requires further validation and research.

Future work may focus on further optimizing the model structure for diverse crystallization processes and industrial scenarios; exploring advanced technologies like generative adversarial networks and recurrent neural networks to handle more complex crystal images and dynamic processes; conducting extensive practical application tests to confirm the model's effectiveness and robustness across varied crystallization scenarios; and considering computational efficiency and real-time capabilities to support real-time monitoring and online control. These initiatives are expected to lead to improved monitoring of the crystallization process and better control over product quality, ultimately driving technological progress in the pharmaceutical and chemical sectors.

REFERENCES

[1] Z. Gao, S. Rohani, J. Gong, et al. Recent developments in the crystallization process: toward the pharmaceutical industry. Engineering, 2017, 3(3): 343-353.

[2] X. Z. Wang, K. J. Roberts, C. Ma. Crystal growth measurement using 2D and 3D imaging and the perspectives for shape control. Chemical Engineering Science, 2008, 63(5): 1173-1184.

[3] S. Verma, P. J. Shlichta. Imaging techniques for mapping solution parameters, growth rate, and surface features during the growth of crystals from solution. Progress in crystal growth and characterization of materials, 2008, 54(1-2): 1-120.

[4] P. Larsen, J. Rawlings, N. Ferrier. Model-based object recognition to measure crystal size and shape distributions from in situ video images. Chemical Engineering Science, 2007, 62(5): 1430-1441.

[5] C. Xiouras, F. Cameli, G. L. Quillo, et al. Applications of artificial intelligence and machine learning algorithms to crystallization. Chemical Reviews, 2022, 122(15): 13006-13042.

[6] A. Vancleef, D. Maes, G. T. Van, L. C. J. Thomassen, L. Braeken. Flow-through microscopy and image analysis for crystallization processes. Chemical Engineering Science, 2022, 248: 117067

[7] C. A. Offiler, A. J. Cruz-Cabeza, R. J. Davey, T. Vetter. Crystal Growth Cell Incorporating Automated Image Analysis Enabling Measurement of Facet Specific Crystal Growth Rates. Crystal Growth & Design, 2022, 22: 2837-2848.

[8] Y. Huo, T. Liu, Y. X. Yang, et al. In Situ Measurement of 3D Crystal Size Distribution by Double-View Image Analysis with Case Study on L-Glutamic Acid Crystallization. Industrial & Engineering Chemistry Research, 2020, 59(10): 4646-4658.

[9] M. Vagenknecht, J. Soukup, A. Chen, R. Irizarry . A deep learning solution for particle size analysis in low resolution inline microscopy images based on generative adversarial network. Powder Technology, 2023, 426: 118641

[10] M. Li, J. Liu, T. Yao, Z. Gao, J. Gong . Deep-learning based in-situ micrograph analysis of high-density crystallization slurry using image and data enhancement strategy. Powder Technology, 2024, 437: 119582.

[11] L. Wang, Y. Zhu, C. Gan. Nonlinear model predictive control of crystal size in batch cooling crystallization processes. Journal of Process Control, 2023, 128: 103020.

[12] S. Zong, G. Zhou, M. Li, X. Wang. Deep learning-based on-line image analysis for continuous industrial crystallization processes. Particuology, 2023, 74: 173-183.
[13] J. He, J. Zhou, J. Dong, Z. Su, L. Huang. Revealing the effects of microwell sizes on the crystal growth kinetics of active pharmaceutical ingredients by deep learning. Chemical Engineering Journal, 2022, 428: 131986.
[14] Y. Wu, Z. Gao, S. Rohani. Deep learning-based oriented object detection for in situ image monitoring and analysis: A process analytical technology (PAT) application for taurine crystallization. Chemical Engineering Research and Design, 2021, 170: 444-455.
[15] V. Manee, R. Baratti, J. A. Romagnoli. Learning to navigate a crystallization model with Deep Reinforcement Learning. Chemical Engineering Research and Design, 2022, 178: 111–123.
[16] H. Salami, M. A. McDonald, A. S. Bommarius, R. W. Rousseau, M. A. Grover. In Situ Imaging Combined with Deep Learning for Crystallization Process Monitoring: Application to Cephalexin Production. Organic Process Research & Development, 2021, 25: 1670-1679.
[17] J. Xin, Z. Wei, M. Yang, X. Peng, W. Du. Merged-Sampling Mask R-CNN With Random Proposal Expansion for Particle Measurement of SEM Images of Molecular Sieve Catalysts. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 5019413.
[18] J. Fan, T. Liu, Y. Huo, Y. Tan, J. Chen. In Situ Measurement of 2-D Crystal Size Distribution During Cooling Crystallization Process via a Binocular Telecentric Imaging System. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 5015115.
[19] O. Ronneberger, P. Fischer, T. Brox. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, 2015, 234-241.
[20] T. Y. Lin, P. Dollár, R. Girshick, et al. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 2117-2125.
[21] C. Guo, B. Fan, Q. Zhang, et al. Augfpn: Improving multi-scale feature learning for object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 12595-12604.
[22] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.
[23] F. Wang, M. Jiang, C. Qian, et al. Residual attention network for image classification. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 3156-3164.
[24] J. Lu, C. Xiong, D. Parikh, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 375-383.
[25] J. Fu, J. Liu, H. Tian, et al. Dual attention network for scene segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, 3146-3154.
[26] A. E. Bruno, P. Charbonneau, J. Newman, et al. Classification of crystallization outcomes using deep convolutional neural networks. PLOS one, 2018, 13(6): e0198883.
[27] Y. Bengio, Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. Advances in neural information processing systems, 2003, 16.

[28] V. Badrinarayanan, A. Kendall, R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.

[29] H. Liu, X. Shen, F. Shang, et al. CU-Net: Cascaded U-Net with Loss Weighted Sampling for Brain Tumor Segmentation. Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy, 2019, 11846.

[30] J. Long, E. Shelhamer, T. Darrell. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, 3431-3440.