

## Article

# Fine-Grained Identification of Benthic Diatom Scanning Electron Microscopy Images Using a Deep Learning Framework

Fengjuan Feng<sup>1</sup>, Shuo Wang<sup>2</sup>, Xueqing Zhang<sup>2</sup>, Xiaoyao Fang<sup>1</sup>, Yuyang Xu<sup>1</sup> and Jianlei Liu<sup>2,\*</sup> <sup>1</sup> School of Life Sciences, Qufu Normal University, Lucheng Street, Jining 273165, China; fengfj@qfnu.edu.cn (F.F.); fangxiaoyao@qfnu.edu.cn (X.F.); 19736372249@qfnu.edu.cn (Y.X.)<sup>2</sup> School of Cyber Science and Engineering, Qufu Normal University, Lucheng Street, Jining 273165, China; shuowang@qfnu.edu.cn (S.W.); xueqingzhang@qfnu.edu.cn (X.Z.)

\* Correspondence: jianleiliu@qfnu.edu.cn

**Abstract:** Benthic diatoms are key primary producers in aquatic ecosystems and sensitive bioindicators for water quality monitoring; for example, the Yellow River Basin exhibits high diatom species diversity. However, traditional microscopic identification of such species remains inefficient and inaccurate. To enable automated identification, we established a benthic diatom dataset containing 3157 SEM images of 32 genera/species from the Yellow River Basin and developed a novel identification method. Specifically, the knowledge extraction module distinguishes foreground features from background noise by guiding spatial attention to focus on mutually exclusive regions within the image. This mechanism allows the network to focus more on foreground features that are useful for the classification task while significantly reducing the interference of background noise. Furthermore, a dual knowledge guidance module is designed to enhance the discriminative representation of fine-grained diatom images. This module strengthens multi-region foreground features through grouped channel attention, supplemented with contextual information through convolution-refined background features assigned low weights. Finally, the proposed method integrates multi-granularity learning, knowledge distillation, and multi-scale training strategies, further improving the classification accuracy. The experimental results demonstrate that the proposed network outperforms comparative methods on both the self-built diatom dataset and a public diatom dataset. Ablation studies and visualization further validate the efficacy of each module.



Received: 14 April 2025

Revised: 24 May 2025

Accepted: 28 May 2025

Published: 30 May 2025

**Citation:** Feng, F.; Wang, S.; Zhang, X.; Fang, X.; Xu, Y.; Liu, J. Fine-Grained Identification of Benthic Diatom Scanning Electron Microscopy Images Using a Deep Learning Framework. *J. Mar. Sci. Eng.* **2025**, *13*, 1095. <https://doi.org/10.3390/jmse13061095>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** benthic diatom; fine-grained identification; deep learning; attention mechanism

## 1. Introduction

Benthic diatoms refer to diatoms that have a relatively fixed growth location and inhabit diverse environments. They occur on stones, sediments, aquatic plants, and other substrates in almost all water bodies [1]. Diatoms are microscopic, unicellular or multicellular eukaryotic algae that engage in photosynthesis, serving as one of the most important primary producers in aquatic ecosystems [2]. When habitat destruction, eutrophication, and heavy metal pollution occur, the community composition and relative abundance of benthic diatoms can respond quickly to water quality or environmental changes [3,4]. Therefore, benthic diatoms are often used as bio-indicators to evaluate water quality and monitor the health of aquatic ecosystems [5–7].

The morphology and ornamentation of the diatom frustule are important bases for classification. With advancements in microscopy, researchers have been able to observe diatoms with increasing precision. Features such as raphe morphology, central and axial area

characteristics, striae shape, and girdle band structure have become increasingly important in diatom species identification [8,9]. Consequently, capturing high-resolution electron microscopy images has become an essential task in diatom research. In practice, due to the vast diversity and morphological variability of benthic diatoms, traditional manual identification methods are time-consuming, labor-intensive, and prone to human error [10]. Additionally, when collecting diatoms by scraping samples from various substrates, impurities such as mud and sand will be mixed. This can significantly increase the difficulty of manual identification. As a result, the application of artificial intelligence technology for diatom identification has gained increasing attention [10–14].

The rise of deep learning has provided an important impetus for the development of fine-grained visual recognition algorithms [15]. Among these algorithms, the convolutional neural network (CNN) [16] architecture enables significantly improved recognition accuracy and discrimination ability. At present, CNN-based fine-grained visual classification algorithms are primarily categorized into two paradigms: strongly supervised and weakly supervised learning approaches. In strongly supervised learning frameworks, fine-grained image classification approaches [17–20] typically require manual bounding box annotations or other forms of explicit supervisory information. These supervisory constraints enable the precise localization of discriminative regions, as exemplified by the R-CNN framework proposed by Girshick et al. [20], which incorporates a region proposal mechanism to select optimal detection results through classification scoring. However, the heavy reliance on large-scale manual annotations in such approaches has resulted in substantial labeling costs, consequently limiting their scalability on extensive datasets.

In light of these limitations, researchers have progressively shifted their focus toward weakly supervised learning methods [21–23]. These methods automatically localize discriminative features without requiring additional manual annotations, thereby enhancing the practicality and scalability of such algorithms. Within this context, ResNet [24], a classical deep convolutional neural network, has been extensively employed in fine-grained visual classification tasks. Its deep residual architecture effectively mitigates the vanishing gradient problem and has demonstrated remarkable performance across various computer vision tasks. However, certain limitations persist when ResNet is applied to tasks involving complex background interference and subtle inter-class variations. Specifically, ResNet primarily relies on global features for classification, lacking precise attention to discriminative local regions, which makes it susceptible to background noise interference. Furthermore, the fixed-scale feature extraction approach adopted in ResNet utilizes only the final layer's features for classification. It is, therefore, difficult to fully capture the target information at different scales, limiting its performance in the fine-grained diatom recognition task.

Researchers have made significant efforts in recent years to address the limitations of deep learning methods in diatom image classification tasks. Pu et al. [25] proposed a novel diatom microscopy image recognition approach that combines static data augmentation, transfer learning, and prediction methods incorporating cosine similarity to enhance the classification performance of networks. Li et al. [13] improved upon the YOLOv5 model by introducing attention mechanisms and spatial pyramid pooling (SPP) to strengthen feature extraction capabilities, along with the K-means++ algorithm to obtain more accurate bounding boxes, thereby enhancing the model's detection performance.

In this study, a novel fine-grained diatom recognition network is proposed. The network incorporates a knowledge extraction module (KEM) and a dual knowledge guidance module (DKGM) specifically designed to mitigate the interference of background noise on classification accuracy in diatom images. Furthermore, a multi-granularity learning strategy is implemented to capture hierarchical feature representations, complemented by knowledge distillation techniques to ensure the accuracy and robustness of multi-granularity

features. To further enhance the effectiveness of feature extraction, the network introduces a multi-scale training strategy. Using a foreground cropping method based on the attention mechanism, a purer target diatom region can be obtained, enhancing the model's ability to focus on key features. The entire network is trained in an end-to-end manner, ensuring the coordinated optimization of all modules. The experimental results on a self-built diatom dataset demonstrate significant improvements in classification accuracy. Additionally, ablation studies and visual analysis further validate the effectiveness of each module and their respective contributions to classification performance.

In summary, this study makes the following key contributions:

- (1) We propose a novel fine-grained diatom recognition network that integrates a KEM and a DKGM to enhance target diatom features and suppress background noise. The network combines knowledge distillation, multi-granularity learning, and multi-scale feature extraction, optimizing feature representation for more accurate diatom classification.
- (2) We design a new KEM that generates foreground attention maps to strengthen diatom feature representation and reduce background noise interference. These attention maps also serve as prior knowledge to guide model training.
- (3) We designed a novel DKGM that employs a grouped channel attention mechanism to reinforce multi-region foreground features while extracting and refining background information as auxiliary cues.

## 2. Materials

### 2.1. Benthic Diatom Dataset

Benthic diatom samples were collected from the main stem of the Yellow River and its primary tributaries within Shandong Province, China, including the Dawen River, Beidasha River, and Yufu River. Sampling was conducted during the summer and autumn of 2022, as well as the spring of 2023. In the laboratory, the samples were treated with nitric acid (Tianjin Damao Chemical Reagent Factory, Tianjin, China), and only the siliceous cell walls of the diatoms were retained. Diatom imaging and identification were performed using a field emission scanning electron microscope (FESEM, Zeiss Sigma 500, Carl Zeiss, Oberkochen, Germany). In total, 3157 scanning electron microscope images were collected and classified into 32 diatom taxa (genera or species) (see Table 1 and Figure 1). Notably, considering that environmental monitoring samples often contain various impurities, no measures were taken to reduce or eliminate the interference from sediment when establishing the image dataset.

**Table 1.** Information on the diatom dataset.

| (Genera/Species)               | Number of Images | Percentage (%) |
|--------------------------------|------------------|----------------|
| <i>Melosira varians</i>        | 49               | 1.552          |
| <i>Aulacoseira granulata</i>   | 85               | 2.692          |
| <i>Thalassiosira lacustris</i> | 37               | 1.172          |
| <i>Cyclotella meneghiniana</i> | 288              | 9.123          |
| <i>Cyclotella meduanae</i>     | 61               | 1.932          |
| <i>Discostella</i>             | 25               | 0.792          |
| <i>Stephanodiscus</i>          | 107              | 3.389          |
| <i>Diatoma</i>                 | 118              | 3.738          |
| <i>Fragilaria</i>              | 104              | 3.294          |
| <i>Tabularia</i>               | 99               | 3.136          |
| <i>Punctastriata</i>           | 46               | 1.457          |
| <i>Staurosira</i>              | 43               | 1.362          |
| <i>Pseudostaurosira</i>        | 107              | 3.389          |
| <i>Cocconeis</i>               | 76               | 2.407          |

Table 1. Cont.

| (Genera/Species)                         | Number of Images | Percentage (%) |
|------------------------------------------|------------------|----------------|
| <i>Achnanthyidium laticephalum</i>       | 49               | 1.552          |
| <i>Achnanthyidium minutissimum</i>       | 235              | 7.444          |
| <i>Planothidium</i>                      | 90               | 2.851          |
| <i>Amphora</i>                           | 54               | 1.710          |
| <i>Cymbella</i>                          | 90               | 2.851          |
| <i>Encyonema</i>                         | 99               | 3.136          |
| <i>Encyonopsis</i>                       | 100              | 3.168          |
| <i>Gomphonella</i>                       | 85               | 2.692          |
| <i>Gomphonema parvulum</i>               | 70               | 2.217          |
| <i>Gomphonema constrictum</i>            | 37               | 1.172          |
| <i>Navicula</i>                          | 350              | 11.086         |
| <i>Sellaphora</i>                        | 50               | 1.584          |
| <i>Craticula subminuscule</i>            | 51               | 1.615          |
| <i>Nitzschia</i>                         | 249              | 7.887          |
| <i>Nitzschia sinuata var. tabellaria</i> | 27               | 0.855          |
| <i>Nitzschia soratensis</i>              | 86               | 2.724          |
| <i>Nitzschia palea</i>                   | 119              | 3.769          |
| <i>Nitzschia amphibia</i>                | 71               | 2.249          |
| Total                                    | 3157             | 100.00         |

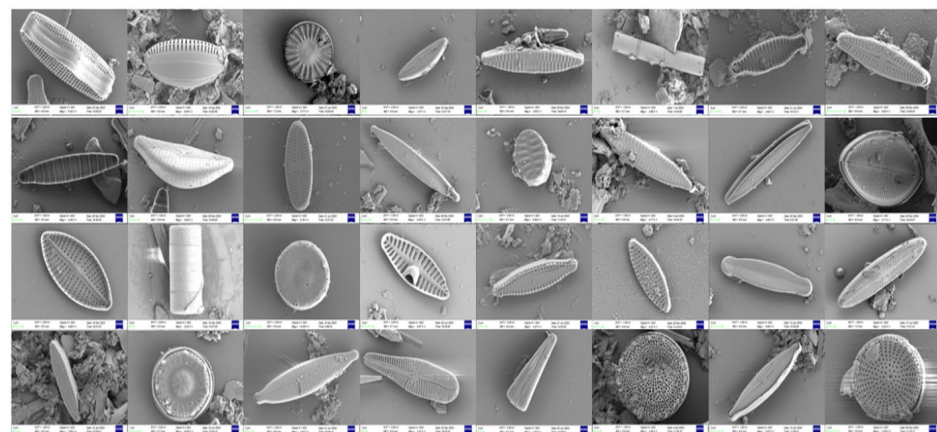


**Figure 1.** Examples of samples from the 32 representative categories in the self-built diatom dataset (ALGAE).

## 2.2. Data Preprocessing

In the training and testing phases of the model, we performed uniform preprocessing operations on the input images. This ensures that the data format meets the model input requirements and improves the training effect and generalization ability of the model. Initially, the shorter side of the image was resized to a preset size of 512 while keeping the aspect ratio constant. This avoids deformation that may result from scaling directly to the target size and maximizes the retention of key information in the image. Subsequently, a square region with a fixed size of  $448 \times 448$  was cropped from the center of the adjusted image to ensure consistent input size. This allows the model to focus more on the features of the subject area, thus reducing background interference. The processed images were then converted into a tensor format ( $[C, H, W]$  – channels  $\times$  height  $\times$  width). The pixel values were normalized to the range  $[0, 1]$  to accommodate the numerical computation requirements of the deep learning model. Finally, channel-wise normalization was performed using the mean ( $[0.485, 0.456, 0.406]$ ) and standard deviation ( $[0.229, 0.224, 0.225]$ ) values commonly employed in ImageNet pretrained models. We aimed to make the distribution of the input data consistent with the training data of the pretrained models, thus improving the stability of feature extraction and the effectiveness of migration learning.

This comprehensive preprocessing strategy ensures consistency of the data, optimizes training convergence speed, and enhances model robustness and generalization capability in fine-grained diatom recognition tasks.

### 2.3. Residual Network

The rapid advancement of convolutional neural networks (CNNs) can be largely attributed to the introduction of AlexNet [26] in 2012. Subsequent architectures, including VGG [27] and GoogLeNet [28], have achieved significant improvements in image classification performance through increased network depth and enhanced structural designs. However, as network depth continues to grow, challenges such as vanishing gradients and network degradation have become increasingly prominent. This ultimately limits further performance enhancements.

Proposed by He et al. in 2015, ResNet represents a significant advancement in deep convolutional neural networks through its innovative residual learning mechanism. The core innovation lies in the introduction of skip connections that directly add input features to the output of convolutional layers. It makes the gradient propagate directly in the backpropagation process, effectively alleviating the problem of gradient vanishing in deep networks. When the depth exceeds 20 layers, the traditional convolutional neural network will be saturated, and even the performance will be degraded. In contrast, ResNet can successfully train a 152-layer deep network through its residual learning mechanism. It can also reduce the Top-5 error rate on the ImageNet dataset to 3.57%, which won the ImageNet competition in 2015. The design of the residual block allows the network to learn the residual mapping of the input features instead of directly fitting the full nonlinear mapping. This design reduces the optimization difficulty and significantly improves the feature representation of the network. These architectural innovations mitigate gradient vanishing and network degradation in deep neural networks, enabling deeper network training while providing foundational advances for deep learning development.

ResNet has achieved groundbreaking results on multiple benchmark datasets, including ImageNet [29] and COCO [30], setting new performance records and establishing itself as a classical baseline model for image classification tasks. The core principles of ResNet have not only propelled the development of extremely deep neural networks but also provided significant theoretical insights and practical guidance for subsequent model designs. Within both industrial and academic communities, ResNet's impact has been profound, with widespread applications across various computer vision tasks, such as object detection [31,32], semantic segmentation [33,34], and image classification [35–38]. This architectural innovation stands as a pivotal milestone in the evolution of deep learning.

## 3. Methods

This study proposes a novel fine-grained diatom recognition model (Figure 2) comprising two key components: (1) a KEM and a DKGM that collaboratively enhance foreground feature representation while suppressing background noise, with extracted prior knowledge guiding multi-scale training strategies (MS), and (2) a knowledge distillation mechanism (KD) that provides supervisory signals for multi-granularity feature learning to improve model robustness and generalization. The following sections detail these components' design and implementation.

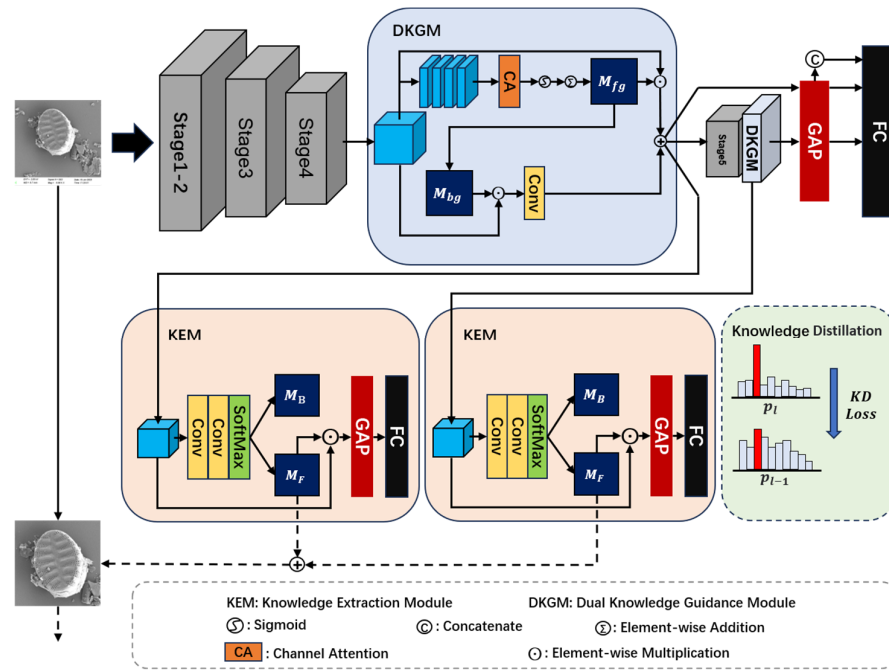


**Figure 2.** The proposed KEGC network architecture, where  $M_B$  is the background attention map,  $M_F$  is the foreground target attention map,  $M_{fg}$  is the foreground features' enhanced matrices,  $M_{bg}$  is the background features refined matrices,  $p_l$  is the prediction probability corresponding to the output of the last layer, and  $p_{l-1}$  is the prediction probability corresponding to the output of the penultimate layer. GAP is global average pooling.

### 3.1. Knowledge Extraction Module

Under electron microscopy imaging conditions, diatom images typically contain complex background information, including diatom debris, sediment, and other non-target particulate matter. The noise information is highly similar to the diatom target in terms of texture, shape, and grayscale features. This leads to the difficulty for traditional networks in effectively distinguishing foreground targets from background noise in the feature extraction process. This confusion leads the model to learn numerous features that are irrelevant to the classification task, significantly compromising classification performance. To address this issue, we propose a KEM. It extracts discriminative foreground features from complex backgrounds while suppressing background noise interference, thereby enhancing the model's feature extraction capabilities and classification accuracy. The implementation description follows.

For a given feature map,  $X \in R^{C \times H \times W}$ , the process begins with its sequential processing through two convolutional, Batch Normalization, ReLU, and Softmax layers to generate mutually exclusive foreground and background attention maps. The input feature map,  $X$ , is then multiplied with the generated foreground attention map,  $M_F$ , to produce an enhanced feature representation. Notably, the final convolutional layer in this module is configured with an output dimension of 2, followed by channel-wise Softmax processing to generate mutually exclusive attention maps. The enhanced feature map, refined through foreground attention, is subsequently fed into pooling layers and classifiers, with optimization performed through cross-entropy loss computation. Meanwhile, the background attention map is discarded, effectively achieving foreground extraction and background noise suppression.

The introduction of the KEM module has significantly increased the computational overhead of the model. Specifically, this module expands the model parameters from 23.57 M to 77.56 M (a 229% increase), raises FLOPs from 16.53 G to 33.39 G (a 102% increase), extends single-sample processing time during inference from 5.90 ms to 12.47 ms (a 111%

increase), and increases GPU memory usage by 82% (from 250.44 MB to 455.95 MB). These elevated computational costs primarily stem from the design of the attention mechanism.

### 3.2. Dual Knowledge Guidance Module

The classification of benthic diatoms presents significant challenges due to high inter-class similarity among subspecies and substantial background noise interference. To address these issues, it is critical to enhance discriminative local features while suppressing irrelevant background noise. Existing attention-based methods predominantly focus on global salient regions. They often neglect both the extraction of discriminative multi-region features and the potential value inherent in background information. To overcome these limitations, a DKGM is proposed. DKGM uses grouped channel attention to decouple foreground and background features. Then, the convolution improvement is carried out on the background features obtained from decoupling. Through low-weight feature fusion, an optimal balance between feature discriminability and diversity is achieved.

Specifically, this module takes the output of the last two layers of the backbone network as input. Firstly, group along the channel dimension, and generate  $k$  regional attention matrices through the parallel pixel attention mechanism. These matrices are fused by a summation operation to obtain the global foreground attention matrix,  $M_{fg}$  (as shown in Equation (1)). Subsequently,  $M_{fg}$  is combined with the input features through element-by-element multiplication. The residual concatenation is used to obtain the foreground enhancement features. To extract the background information,  $M_{fg}$  is thresholded. All elements except the first  $m$  maximum values (replace with the minimum values  $i$ ) are then sigmoid normalized to obtain the background attention matrix,  $M_{bg}$  (as shown in Equation (2)). This matrix is multiplied with the original input features. The spatial correlation is further modeled by a  $1 \times 1$  convolutional layer to ultimately generate the background complementary features. The module output,  $X_{mix}$ , is obtained through the weighted summation of foreground augmented features and background complementary features, as shown in Equation (3).

$$M_{fg} = \sum_k^i M_{fg}^i, \quad (1)$$

$$M_{bg}(i, j) = \sigma \left( \begin{cases} i, & x < I_{Max} \\ x, & x \geq I_{Max} \end{cases} \right), \quad (2)$$

$$X_{mix} = X + X \cdot M_{fg} + \alpha * (X \cdot M_{bg}), \quad (3)$$

The DKGM module demonstrates high computational efficiency. When combined with the MS + KD modules (excluding KEM), the model maintains 30.37 M parameters (a 29% increase over baseline), 37.18 G FLOPs (125% increase), and a 17.54 ms inference time (197% increase). Through efficient global feature modeling, this module achieves an optimal balance between performance and computational cost.

### 3.3. Multiresolution Learning, Multi-Scale Training, and Knowledge Distillation

During feature extraction in the backbone network, shallow-layer features typically capture low-level visual information such as edges or textures. The deep-layer features contain high-level abstract representations, including semantic information and global structures. Traditional networks exclusively utilize the final layer's high-level features for classification. They overlooked the rich information embedded in shallow features. To address this limitation, we implement multi-granularity learning by feature enhancement and fusion on the outputs from the last two layers of the feature extraction network. This

approach enables the comprehensive utilization of multi-level feature information, resulting in more robust feature representations for diatom images.

However, features extracted from intermediate network layers demonstrate high sensitivity to input variations. This issue is particularly pronounced in fine-grained diatom recognition tasks. Due to the substantial intra-class variations caused by diverse diatom orientations, complex backgrounds, and varying imaging conditions. Directly incorporating low-level features as supplementary information may compromise model robustness and generalization capability, as these features are inherently sensitive to noise and variations. To address this challenge, we employ a self-distillation learning approach, where high-confidence semantic information extracted from deep layers guides the refinement of shallow features. This strategy effectively suppresses noise-induced interference in shallow features, thereby enhancing the robustness of feature representation.

The feature map derived from the network's final layer is first processed through global average pooling. This operation compresses spatial information to generate a compact feature vector. The vector is then propagated into the classifier module, where non-linear transformations produce probabilistic outputs. Finally, the resulting predicted probability distribution serves as the teacher signal for subsequent knowledge distillation processes. Simultaneously, the preceding layer's feature map undergoes identical operations to generate its predicted probability distribution as the student signal. By minimizing the distribution discrepancy between the teacher and student signals, high-level semantic information effectively supervises and guides the refinement of shallow features. So as to achieve the purpose of enhancing the generalization ability of the model.

To address the challenge of complex background interference in electron microscopy diatom images, we adopt a multi-scale training strategy. This can further mitigate the negative impact of background noise on the model performance. To further reduce the negative impact of background noise on model performance, inspired by [25], our method implements the multiscale training strategy as follows: Based on the KEM design, during network training, we sum the foreground attention maps generated at each stage along the channel dimension. The aim is to obtain an aggregated foreground attention matrix,  $A_{sum}$ , and then compute its mean,  $A_{mean}$ . Subsequently, binarization is performed using a randomly selected threshold,  $\theta$ , within the predefined range (Equation (5)). Then, the maximum connected region is extracted from the obtained binary matrix by using the 4-neighborhood connectivity. This region is adjusted to form a rectangular bounding box serving as the foreground mask  $M$ . It is used to crop the original input image, thereby obtaining background-suppressed purified images. Notably, bounding boxes are dynamically regenerated each epoch. The cropped images are then fed as input into the next stage network. This approach effectively reduces background noise interference, thereby enhancing feature extraction robustness and classification performance.

The MS module and KEM-guided foreground extraction increase FLOPs by 33.38 G (+100%) and prolong inference time by 13.53 ms (+108%). This is mainly due to the secondary computation of multi-scale features. The KD module enhances robustness through deep-to-shallow distribution alignment. During the training period, only 20% of the computational overhead is increased while maintaining the same inference parameters (77.56 M  $\rightarrow$  77.56 M) and FLOPs. This computational cost distribution demonstrates a decoupling pattern: MS-dominated inference costs versus KD-constrained training costs.

$$M = Crop(M_F^1 + M_F^2) \quad (4)$$

$$Crop = \begin{cases} 1, & A_{sum} > A_{mean} * \theta \\ 0, & otherwise \end{cases}, \quad (5)$$

$$A_{sum}(i, j) = \sum_{c=1}^C (M_F^1 + M_F^2), \quad (6)$$

$$A_{mean} = \sum_{i=1, j=1}^{H \times W} A_{sum}(i, j), \quad (7)$$

### 3.4. Model Evaluation

To evaluate model performance during training and facilitate objective comparison with existing methods, we design a composite loss function consisting of classification loss and distillation loss. Both components employ cross-entropy loss functions, calculated across three factors: (1) classification probabilities obtained from global pooling and the classification of feature maps from the last two layers of the backbone network; (2) classification probabilities derived from the concatenated features of these two layers; and (3) classification probabilities corresponding to feature maps enhanced by the KEM. The distillation loss specifically utilizes the classification probability from the final layer as the teacher signal to supervise the classification probability from the preceding layer, thereby achieving knowledge distillation objectives. Through joint optimization of these loss functions, the model simultaneously enhances classification performance and feature representation robustness. The detailed definitions are as follows:

$$L = L_{cls} + \beta * L_{kd}, \quad (8)$$

$$L_{cls} = \sum_{i=1}^N L_{ce}(p_i) + L_{ce}(p_{cat}) + L_{ce}(p_{KEM}), \quad (9)$$

$$L_{kd} = -p_l \log(p_{l-1}), \quad (10)$$

where  $L$  is the total loss,  $L_{cls}$  is the classification loss,  $L_{kd}$  is the distillation loss, and  $L_{ce}$  is the cross-entropy loss.  $p_i$  denotes the prediction vector at stage  $i$ ,  $p_{cat}$  denotes the vector from the stages after concatenation of the prediction vectors, and  $p_{KEM}$  denotes the prediction vector from the KEM.

Furthermore, we introduce several evaluation metrics to comprehensively assess model performance, including Top-1 Accuracy, Top-5 Accuracy, Recall, and F1 score. The detailed computational formulas are presented as follows:

$$\text{Top} - 1 \text{ Accuracy} = \frac{\text{Number of Correct Top} - 1 \text{ Predictions}}{\text{Total Number of Samples}} \times 100\%, \quad (11)$$

$$\text{Top} - 5 \text{ Accuracy} = \frac{\text{Number of Correct Top} - 5 \text{ Predictions}}{\text{Total Number of Samples}} \times 100\%, \quad (12)$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\%, \quad (13)$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\%, \quad (14)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (15)$$

## 4. Results

### 4.1. Model Training

This experimental study was conducted using the PyTorch 3.9 deep learning framework, with all training and testing performed on NVIDIA RTX 4090 GPU (GAINWARD, Taiwan, China) hardware. ResNet50, pretrained on ImageNet [28], was employed as the backbone network, accompanied by standard image preprocessing operations, in-

cluding resizing and center cropping. Specifically, input images were initially resized to  $512 \times 512$  resolution, followed by center cropping to dimensions of  $448 \times 448$ , with this preprocessing pipeline consistently applied during both training and testing phases. The model was trained for 200 epochs with a batch size of 16. The optimization process utilized Stochastic Gradient Descent with a momentum parameter of 0.9 and weight decay of  $5 \times 10^{-4}$ . For the hyperparameters,  $\alpha$  was set to 0.1,  $\beta$  was set to 2, and  $\theta \in (0.8, 1)$ . Notably, throughout the experiments, only image-level labels were used for supervision. The project code is openly accessible at <https://github.com/ShuoWang00/KEGC-main> (accessed on 27 May 2025).

In deep learning model training, differential learning rates are typically employed for distinct network components to optimize training efficiency. Considering that backbone networks have been pretrained on large-scale datasets and have acquired effective features with strong generalization capabilities across various tasks, we set a relatively low learning rate of  $2 \times 10^{-4}$  for the backbone network. Conversely, for newly added network layers such as classifiers, attention modules, or custom modules, which require learning novel features without pretraining, a higher learning rate of  $9 \times 10^{-3}$  was implemented to facilitate rapid learning and accelerate convergence. This strategy not only expedites the training of new layers but also preserves the valuable features already learned in the backbone network. By implementing differential learning rates across network components, the model can efficiently acquire new features while preventing excessive modifications of existing features, ultimately enhancing both task-specific performance and generalization capability.

In terms of computational efficiency, as presented in Table 2, the KEGC model exhibits 82.81 million parameters, 70.07 GFLOPs, an average inference time of 26.54 ms, and GPU memory usage of approximately 500.08 MB. The results demonstrate that although KEGC shows slightly higher parameter counts and computational requirements than models like ResNet50 and ACC, it maintains efficient operational performance with a controlled inference time of around 26 ms and memory consumption of about 500 MB, enabling effective execution on mid-range GPUs (such as RTX 4090 and above). Compared with baseline models, KEGC achieves significant improvements in classification performance and feature representation capability with only a marginal increase in inference time, demonstrating an optimal balance between performance and efficiency.

**Table 2.** Computational cost analysis on ALGAE dataset.

| Model       | Params (M) | FLOPs (G) | Inference Time (ms) | GPU Memory (MB) |
|-------------|------------|-----------|---------------------|-----------------|
| Vgg16       | 134.39     | 61.51     | 6.90                | 616.47          |
| ResNet50    | 23.57      | 16.53     | 5.90                | 250.44          |
| ACC         | 67.63      | 54.13     | 23.29               | 442.79          |
| KEGC (Ours) | 82.81      | 70.07     | 26.54               | 500.08          |

#### 4.2. Model Comparison

To validate the effectiveness of the proposed method, comprehensive evaluations were conducted on the diatom dataset using the trained model. The experiment compared several representative methods that do not rely on additional annotations, such as bounding boxes, all of which utilize only image-level labels for training. Specifically, we selected the classical convolutional network VGG16 [27], the deep residual network ResNet50 [24], and the recently proposed Microscopic image recognition for diatoms method (CST) [25]. Furthermore, we used an alleviating category confusion (ACC) network [38] for comparative analysis. Table 3 presents the quantitative results for these models across four key metrics: F1 Score, Recall, Top-1 Accuracy, and Top-5 Accuracy.

**Table 3.** Comparative results on self-built diatom dataset (ALGAE).

| Model                | F1 Score | Recall (%) | Top-1 Accuracy (%) | Top-5 Accuracy (%) |
|----------------------|----------|------------|--------------------|--------------------|
| VGG16 (2014) [27]    | 0.4548   | 44.45      | 54.60              | 84.05              |
| ResNet50 (2015) [24] | 0.8515   | 84.45      | 87.59              | 97.72              |
| CST(2023) [25]       | 0.8411   | 84.45      | 84.45              | 98.04              |
| ACC (2025) [39]      | 0.8626   | 85.11      | 88.69              | 97.88              |
| KEGC(Ours)           | 0.8950   | 89.48      | 90.89              | 98.82              |

Table 3 demonstrates that the proposed network achieves superior performance across all metrics on the custom algae dataset compared with existing methods. Specifically, our approach improves Top-1 Accuracy by 36.29%, 3.30%, 6.44%, and 2.20% over VGG16, ResNet50, CST, and ACC, respectively. Significant gains can also be observed for Recall (45.03%, 5.03%, 5.03%, and 4.37% improvements) and F1 Score (0.8950), indicating better precision–recall balance. Notably, the model achieves 98.82% Top-5 Accuracy, demonstrating exceptional generalization capability and robustness for fine-grained diatom recognition tasks.

The performance improvement can be primarily attributed to the synergistic optimization between the KEM and DKGM, which collectively enhance foreground feature representation while effectively suppressing background interference. Furthermore, the KD provides stronger supervisory signals for multi-granularity feature learning, enabling the model to extract more stable and discriminative features across different scales. These innovations work in concert to significantly boost classification performance, granting the model distinct advantages in fine-grained diatom recognition tasks.

To further validate the generalization capabilities of our method, extensive experiments were conducted on a public algae dataset comprising 160 categories with 7845 images (5882 for training and 1963 for testing). This more comprehensive dataset, encompassing broader species diversity, was specifically selected to rigorously evaluate the model’s scalability and robustness in large-scale scenarios. The experimental results are presented in Table 4.

**Table 4.** Comparative results on the public diatom dataset.

| Model           | F1 Score | Recall (%) | Top-1 Accuracy (%) | Top-5 Accuracy (%) |
|-----------------|----------|------------|--------------------|--------------------|
| VGG16 (2014)    | 0.5669   | 56.88      | 71.88              | 88.23              |
| ResNet50 (2015) | 0.7487   | 76.50      | 81.56              | 92.92              |
| CST (2023)      | 0.8498   | 85.23      | 85.43              | 94.06              |
| ACC (2025)      | 0.7422   | 73.66      | 83.18              | 93.32              |
| KEGC (Ours)     | 0.8028   | 82.44      | 86.96              | 94.14              |

The experimental results in Table 3 demonstrate that our proposed network consistently outperforms existing methods (VGG16, ResNet50, CST, and ACC) across all metrics on the public diatom dataset, achieving significant improvements of 15.08%, 5.40%, 1.53%, and 3.78% in Top-1 Accuracy; 5.91%, 1.22%, 0.08%, and 0.82% in Top-5 Accuracy; and an F1 Score of 0.8082, indicating superior precision recall balance. However, CST achieved optimal values in both the F1 Score and Recall metrics. We speculate that this may be due to our model’s current limitations in processing low-resolution images. The model’s exceptional performance is further evidenced by its 82.44 Recall, which substantially surpasses all baselines and confirms its strong generalization capabilities and robustness for fine-grained algae recognition tasks. These comprehensive improvements validate the effectiveness of our proposed architecture in handling complex diatom classification challenges.

To substantiate the proposed method's efficacy from a gradient optimization perspective, comparative training loss curves are presented in Figure 3 across both datasets. The gradient-based analysis of training dynamics reveals distinct optimization patterns: (1) The conventional models (ResNet50/VGG16) with single cross-entropy loss show rapid initial convergence but quickly reach a stable phase, indicating limited optimization capacity for fine-grained tasks. (2) The ACC network's quadruple-loss design demonstrates enhanced feature learning capabilities, yet exhibits noticeable mid-training fluctuations due to multi-objective optimization. (3) Our six-component loss achieves superior optimization dynamics: rapid initial descent followed by stable refinement. These results demonstrate our method's advantage in balancing optimization efficiency with model robustness.

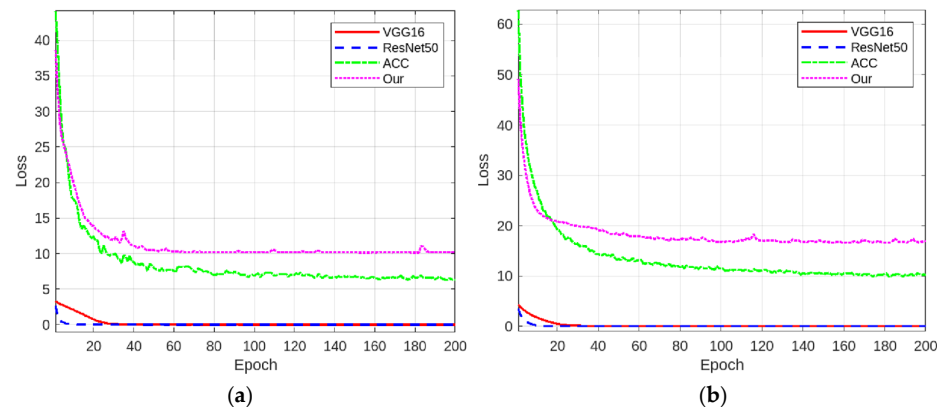


**Figure 3.** Comparative training loss curves of different networks on the self-built diatom dataset (a) and the public diatom dataset (b).

#### 4.3. Ablation Studies

To quantitatively evaluate the contribution of each innovative module to overall performance, we designed systematic ablation studies. Experiments were conducted to decouple and analyze the proposed KEM, DKGM, MS, and KD on a layer-by-layer basis. The independent and synergistic effects of each module on model discrimination ability were explored through the control variable method.

As evidenced by the experimental results in Table 5, the baseline model (without KEM, DKGM, MS, or KD) achieves relatively low performance, with an F1 Score of 0.8515 and a Top-1 Accuracy of 87.59%, indicating that conventional backbone networks struggle to focus on discriminative foreground features without specialized enhancement mechanisms. The introduction of KEM significantly improves all metrics (F1: 0.8735; Top-1: 89.08%), demonstrating its effectiveness in enhancing foreground feature representation while suppressing background interference. Further incorporating MS yields marginal gains (F1: 0.8740; Top-1: 89.47%), suggesting its secondary role in background suppression, though with slight information loss during cropping. Notably, using only KEM + MS causes a minor Top-5 Accuracy drop (97.17%), revealing occasional feature instability from MS. The combination of KEM + KD + MS produces substantial improvements (F1: 0.8842, Recall: 88.26%, Top-1: 90.18%, and Top-5: 98.66%). Finally, the full model (KEM + DKGM + MS + KD) achieves optimal performance (F1: 0.8950, Recall: 89.48%, Top-1: 90.89%, and Top-5: 98.82%), demonstrating synergistic effects among all components. Specifically, KD enhances robustness by guiding shallow features with deep semantic information, while DKGM further refines feature discrimination.

**Table 5.** Results of ablation experiments. KEM is knowledge extraction module. MS is multi-scale training. KD is knowledge distillation. DKEM is dual knowledge guidance module. ✓ indicates the module is included, × indicates the module is excluded.

| KEM | MS | KD | DKGM | F1 Score | Recall (%) | Top-1 Accuracy (%) | Top-5 Accuracy (%) |
|-----|----|----|------|----------|------------|--------------------|--------------------|
| ×   | ×  | ×  | ×    | 0.8515   | 84.45      | 87.59              | 97.72              |
| ✓   | ×  | ×  | ×    | 0.8735   | 87.27      | 89.08              | 98.04              |
| ✓   | ✓  | ×  | ×    | 0.8740   | 87.56      | 89.47              | 97.17              |
| ✓   | ✓  | ✓  | ×    | 0.8842   | 88.26      | 90.18              | 98.66              |
| ✓   | ✓  | ✓  | ✓    | 0.8950   | 89.48      | 90.89              | 98.82              |

To statistically validate the impact of each module, we performed paired *t*-tests. As shown in Table 6, removing any module resulted in a significant performance decline (all *p*-values < 0.05), demonstrating that each component contributes meaningfully to the model's effectiveness. This analysis further corroborates the necessity of our architectural choices.

**Table 6.** Paired *t*-test results for model modules.

| Model                   | <i>p</i> -Value |
|-------------------------|-----------------|
| Full model<br>-w/o KEM  | 0.0058          |
| Full model<br>-w/o MS   | 0.0105          |
| Full model<br>-w/o KD   | 0.0074          |
| Full model<br>-w/o DKGM | 0.0064          |

To further evaluate the model's performance in fine-grained algae recognition tasks, we present the classification accuracy for each category in the ALGAE dataset (Table 7). The results demonstrate that the model achieves satisfactory classification performance for most categories, with a mean class accuracy of 89.41%. It also shows good generalization capability on the public dataset, attaining 81.97% mean class accuracy.

**Table 7.** Classification accuracy by category in the ALGAE dataset.

| Genera/Species                    | Accuracy (%) | Genera/Species                           | Accuracy (%) |
|-----------------------------------|--------------|------------------------------------------|--------------|
| <i>Melosira varians</i>           | 90.00        | <i>Planothidium</i>                      | 94.44        |
| <i>Aulacoseira granulata</i>      | 85.29        | <i>Amphora</i>                           | 90.91        |
| <i>Thalassiosira lacustris</i>    | 100.00       | <i>Cymbella</i>                          | 77.78        |
| <i>Cyclotella meneghiniana</i>    | 98.28        | <i>Encyonema</i>                         | 85.00        |
| <i>Cyclotella meduanae</i>        | 92.00        | <i>Encyonopsis</i>                       | 97.50        |
| <i>Discostella</i>                | 100.00       | <i>Gomphonella</i>                       | 78.57        |
| <i>Stephanodiscus</i>             | 90.70        | <i>Gomphonema parvulum</i>               | 97.06        |
| <i>Diatoma</i>                    | 95.83        | <i>Gomphonema constrictum</i>            | 80.00        |
| <i>Fragilaria</i>                 | 92.86        | <i>Navicula</i>                          | 97.14        |
| <i>Tabularia</i>                  | 87.50        | <i>Sellaphora</i>                        | 90.00        |
| <i>Punctatriata</i>               | 94.74        | <i>Craticula subminuscula</i>            | 100.00       |
| <i>Staurosira</i>                 | 44.44        | <i>Nitzschia</i>                         | 82.00        |
| <i>Pseudostaurosira</i>           | 88.37        | <i>Nitzschia sinuata var. tabellaria</i> | 90.91        |
| <i>Cocconeis</i>                  | 96.77        | <i>Nitzschia soratensis</i>              | 94.29        |
| <i>Achnanthidium laticephalum</i> | 90.00        | <i>Nitzschia palea</i>                   | 77.08        |
| <i>Achnanthidium minutissimum</i> | 98.94        | <i>Nitzschia amphibia</i>                | 82.76        |

Note: Mean accuracy: 89.41%.

Notably, both datasets exhibit certain degrees of class imbalance. For instance, in the ALGAE dataset, while the model achieves 100% accuracy for rare classes like *Discostella* (representing only 0.79% of samples), it struggles with other minority classes such as *Staurosira* (44.44% accuracy). As illustrated in Figure 4, *Staurosira* exhibits significant intra-class morphological variations, which substantially increases the discrimination difficulty. The public dataset contains even more extreme cases, with some classes comprising only 2–3 images, contributing to the observed gap between mean class accuracy (81.97%) and Top-1 Accuracy.

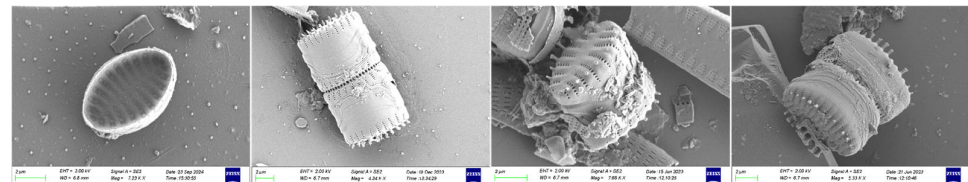


**Figure 4.** Representative samples of *Staurosira*.

Despite these challenges, the model demonstrates robust overall classification performance. Future work will incorporate strategies like class re-weighting, data augmentation, and few-shot learning to enhance performance on long-tailed categories.

Table 8 presents the computational resource consumption of different module combinations on the ALGAE dataset, including parameter count, FLOPs, inference time, and GPU memory usage. The data reveal that the baseline model has the smallest parameter count (23.57 M) and lowest computational complexity (16.53 G FLOPs), with relatively low inference time and memory consumption, making it suitable for resource-constrained environments. The introduction of the KEM module increases the parameters to 77.56 M and FLOPs to 33.39 G, nearly doubling the inference time to 12.47 ms while raising memory usage to approximately 456 MB. This demonstrates the significant computational overhead introduced by the KEM module, though its substantial performance improvements indicate a favorable computation performance trade-off. When incorporating the MS module, while the parameter count remains unchanged, FLOPs nearly double to 66.77 G, and inference time increases to 26.00 ms, with a moderate rise in memory usage. These results confirm that the multi-scale feature processing of the MS module substantially increases computational demands. Subsequent integration of the KD module shows a negligible impact on computational resource consumption, with stable inference time and memory usage, indicating its minimal computational overhead. The complete model, with the additional DKGM module, exhibits a slightly higher parameter count (82.81 M) and computational complexity (70.07 G FLOPs), along with corresponding increases in inference time and memory usage. While representing the highest overall computational cost, this configuration delivers optimal model performance, making it most suitable for applications requiring high classification accuracy.

**Table 8.** Computational efficiency analysis of modules on ALGAE dataset. ✓ indicates the module is included.

| Model    | KEM | MS | KD | DKGM | Params (M) | FLOPs (G) | Inference Time (ms) | GPU Memory (MB) |
|----------|-----|----|----|------|------------|-----------|---------------------|-----------------|
| Baseline |     |    |    |      | 23.57      | 16.53     | 5.90                | 250.44          |
|          | ✓   |    |    |      | 77.56      | 33.39     | 12.47               | 455.95          |
|          | ✓   | ✓  |    |      | 77.56      | 66.77     | 26.00               | 478.90          |
|          | ✓   | ✓  | ✓  |      | 77.56      | 66.77     | 26.00               | 478.90          |
|          | ✓   | ✓  | ✓  | ✓    | 30.37      | 37.18     | 17.54               | 299.46          |
|          | ✓   | ✓  | ✓  | ✓    | 82.81      | 70.07     | 26.54               | 500.08          |

In summary, while the MS module does introduce a considerable computational burden, it provides auxiliary benefits for fine-grained feature extraction. The KEM module remains the primary source of increased computational costs but delivers significant performance gains. In future work, we plan to explore lightweight designs for the MS module to optimize computational efficiency while preserving its performance advantages.

#### 4.4. Visualization

In deep learning model research, interpretability is crucial for revealing decision-making mechanisms, facilitating our understanding of how models process input information and generate predictions. Visualization techniques enable the revelation of network attention distribution across different features, verifying whether decisions are based on semantically relevant information. As a widely adopted visualization method, Grad-CAM [40] generates heatmaps by backpropagating the gradient information of target categories, quantifying the network's attention intensity on critical regions during classification decisions. This approach provides an intuitive visualization of high-level semantic feature focus locations, such as discriminative object parts or contextually relevant regions. Notably, in the heatmaps, warmer colors (red indicating a high response; blue indicating a low response) correspond to stronger feature responses, reflecting higher network attention to those regions [41]. For this reason, this study uses heatmaps generated based on Grad-CAM to visualize and analyze the designed network.

As illustrated in Figure 5, we visualized feature maps from stage 4 (shallow) and stage 5 (deep) of the network. The first column displays the original images, while the second and third columns present grayscale maps and heatmaps of stage 4 features, respectively. Similarly, the fourth and fifth columns show grayscale maps and heatmaps of stage 5 features. The visualization reveals that shallow feature maps primarily focus on local details while lacking global semantic information. By contrast, deep features demonstrate the ability to capture high-level abstract information by emphasizing both global semantics and semantically critical regions of targets. However, deep features tend to overlook fine details, making them sensitive to target pose and occlusion. Consequently, the fusion of shallow-level details with deep semantic features enables the network to simultaneously preserve both global structure and local discriminative information of targets, validating the necessity of multi-granularity learning.

Figure 6 presents a comparative visualization of final-stage feature maps across different networks on both datasets, validating the effectiveness of our proposed modules. The activation patterns exhibit the following distinct characteristics: (1) VGG16 (second row) demonstrates mixed background-target activations due to ineffective feature selection, corresponding to its low recall (44.5); (2) ResNet50 (third row) concentrates excessively on single local regions while neglecting discriminative edge features, explaining its limited fine-grained performance; (3) the ACC network (fourth row) achieves multi-local activation through cross-fusion modules but remains susceptible to background interference. In contrast, our method (fifth row) exhibits globally coordinated activation, with high-response regions uniformly covering key diatom structures (e.g., striae and marginal contours) while effectively suppressing background noise. These visual results directly correlate with KEM's noise-filtering capabilities and DKGM's multi-region feature balancing mechanism, conclusively demonstrating our approach's comprehensive advantages for fine-grained diatom recognition.

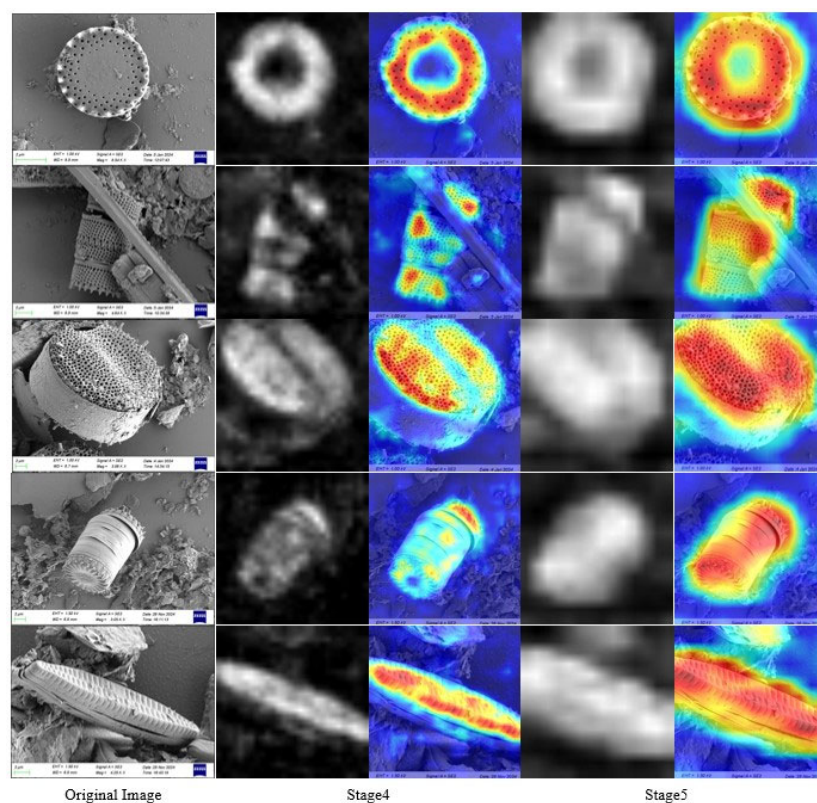


**Figure 5.** Stage 4 and stage 5 visualization of our network. The first column displays original images, while the second and third columns present grayscale maps and heatmaps of stage 4 features, respectively. Warmer colors in the heatmaps (e.g., red and orange) indicate higher activation intensities in those regions, demonstrating that the model prioritizes these areas for classification decisions.

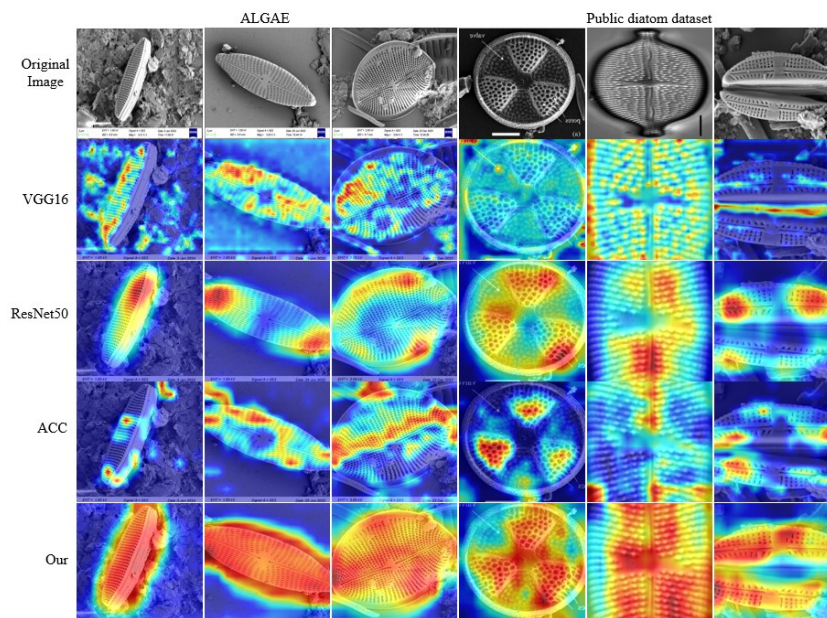


**Figure 6.** Comparative visualization of final-stage feature maps across different networks. The first three columns display images from the ALGAE dataset, while the last three columns show images from the public dataset. The first row presents the original images, and columns 2–5, respectively, show the gradient class activation maps (Grad-CAM) of VGG16, ResNet50, ACC, and our proposed model.

To intuitively demonstrate each module's efficacy, we created Grad-CAM heatmap visualizations to analyze attention mechanisms across different model configurations

(Figure 7). The qualitative results reveal the following: (1) Compared with baseline models, incorporating the KEM substantially enhances focus on discriminative regions. (2) The complete model achieves optimal attention distribution, with significantly superior coverage precision over single-module configurations. These findings confirm that removing either KEM or DKGM degrades fine-grained feature perception, while their synergistic interaction produces marked complementary effects.

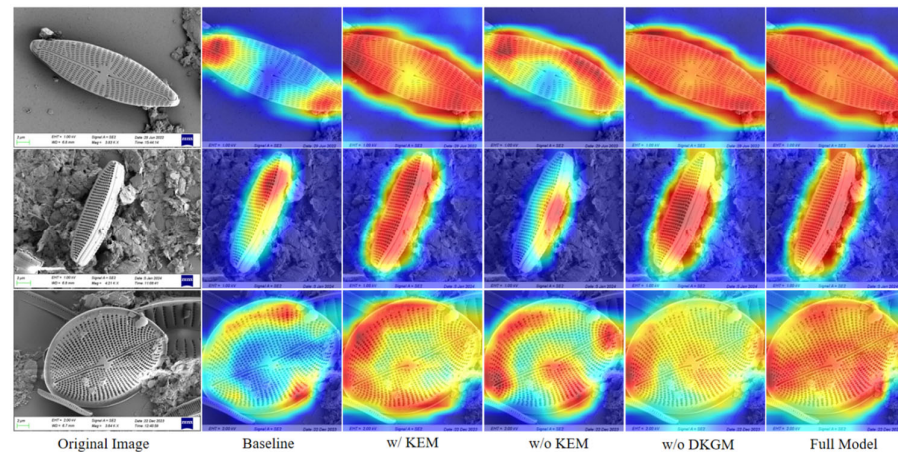


**Figure 7.** Comparative Grad-CAM attention heatmaps of ablation models. The notation w/ indicates inclusion of the corresponding module, while w/o denotes its removal. Warmer colors represent higher attention intensity in the model's focus regions. The results clearly demonstrate that the synergistic interaction between KEM and DKGM allows the network to achieve the most comprehensive and precise attention distribution, significantly outperforming any single-module configuration.

## 5. Discussion

The effectiveness of the proposed method was demonstrated through various experiments. However, several limitations and potential issues remain to be addressed in future work.

### 5.1. Limitations from Class Imbalance

Our dataset exhibits significant class imbalance. Although we standardized the input data through fixed-resolution adjustment and center cropping and evaluated model performance using metrics like Recall and F1 score to mitigate imbalance effects, the classification accuracy of minority classes still has room for improvement. Moreover, this imbalance partially reflects the authentic distribution of diatom communities in the Yellow River Basin. Future studies could incorporate few-shot learning or Generative Adversarial Networks (GANs) to enhance generalization for minority classes while preserving ecological data authenticity.

### 5.2. Model Complexity and Computational Costs

While integrating attention mechanisms, multi-scale training, and knowledge distillation improves classification performance, the resulting model has substantial parameters and may face efficiency challenges when deployed on resource-constrained devices. Future work will explore model light-weighting techniques to improve computational efficiency and reduce hardware requirements.

### 5.3. Technical Challenges for Practical Applications

The current method is well optimized for high-resolution microscopy images ( $1024 \times 768$  px) but achieves suboptimal performance metrics on low-resolution datasets. Future developments will include multi-scale, super-resolution joint training strategies and

semi-supervised learning frameworks to leverage unlabeled watershed data for improved model scalability.

## 6. Conclusions

This study proposed a novel method for fine-grained diatom recognition. In this approach, a KEM is introduced to effectively enhance the expressive capability of foreground features while suppressing background noise interference. By generating mutually exclusive foreground and background attention maps, this module significantly improves both the precision and robustness of feature extraction. Subsequently, the method includes a DKGM that employs a foreground-dominant and background-supplemented fusion strategy. This design not only substantially improves classification accuracy but also maintains the robustness and interpretability of features. Furthermore, the method incorporates a knowledge distillation mechanism that utilizes high-level semantic information extracted from deep networks to supervise shallow features, thereby further enhancing the model's generalization capability. The introduction of multi-granularity learning enables the model to fully leverage multi-level feature information, resulting in a more comprehensive feature representation of diatom images. Finally, through a multi-scale training strategy that combines foreground attention maps for input image cropping, the method further reduces the negative impact of background noise and improves classification performance. The obtained experimental results demonstrated that the proposed method significantly outperforms comparative models on our custom diatom dataset. Ablation studies and visualization analyses further validated the effectiveness of each module. Future work will focus on further optimizing the model's performance to advance research and applications in the field of diatom image classification.

**Author Contributions:** Conceptualization, F.F., S.W., X.Z. and J.L.; Methodology, S.W.; Experiment validation, X.Z.; Benthic diatom dataset construction and results identification, F.F.; Diatom samples preparation and processing, scanning electron microscopy (SEM) images capturing, F.F., X.F. and Y.X.; Writing—original draft preparation, F.F., S.W. and X.Z.; Writing—review and editing, J.L.; Visualization, X.Z.; Supervision, J.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Jining City Key Research and Development Program: “Intelligent Identification System for Benthic Diatom in the Yellow River Basin: Development and Application” (Grant No. 2023HHCG013), and the 2024 Experimental Teaching Reform Research Project of Qufu Normal University “Intelligent Identification of Algal Plants in the Yellow River Basin (Shandong Section) Based on Artificial Intelligence Technology” (Grant No. SJG202432).

**Data Availability Statement:** The public diatom dataset (diatom\_genera\_dataset) used in this study is sourced from [25] (DOI: 10.1111/jpy.13390), accessible via registration at diatom.org. The proposed Yellow River Basin diatom dataset (ALGAE) is available upon request from the corresponding author. Researchers interested in accessing the data should submit a formal application via email, specifying the intended use of the data.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Yu, P.; You, Q.; Pang, W.; Cao, Y.; Bi, Y.; Wang, Q. Development of a Periphytic Diatom—Based Comprehensive Diatom Index for Assessing the Trophic Status of the Lakes in the Lower Reaches of the Yangtze River, China. *Water* **2021**, *13*, 3570. [\[CrossRef\]](#)
2. Parikh, H.S.; Dave, G.; Tiwari, A. Microplastic pollution in aquatic ecosystems: Impacts on diatom communities. *Environ. Monit. Assess.* **2025**, *197*, 206. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Phuyal, P.; Ranabhat, S.; Khatri, S.; Lamichhane, N.; Pant, R.R.; Thapa, L.B.; Yadav, R.K.P. Hydrochemical characteristics, water quality and diatom assemblage in Dordi River, Nepal. *Watershed Ecol. Environ.* **2025**, *7*, 23–35. [\[CrossRef\]](#)

4. Wang, H.; Dong, Y.L.; Jiang, Y.T.; Zhang, N.N.; Liu, Y.; Lu, X.X.; Fan, Y.W. Multiple stressors determine the process of the benthic diatom community assembly and network stability in urban water bodies in Harbin. *Sci. Total Environ.* **2024**, *913*, 169536. [\[CrossRef\]](#)
5. Liu, X.; Zhao, Y.; Yu, X.Q.; Song, Y.; Li, Y.; Yang, G.P.; Li, Y.H.; Zhu, B.H.; Pan, K.H. Seasonal changes of species- and guild-based benthic diatom communities in the transitional water zone of the Yellow River Delta. *Mar. Environ. Res.* **2025**, *203*, 106843. [\[CrossRef\]](#)
6. Kireta, A.R.; Reavie, E.D.; Sgro, G.V.; Angradi, T.R.; Bolgrien, D.W.; Hill, B.H.; Jicha, T.M. Planktonic and periphytic diatoms as indicators of stress on great rivers of the United States: Testing water quality and disturbance models. *Ecol. Indic.* **2012**, *13*, 222–231. [\[CrossRef\]](#)
7. Vidal, T.; Santos, M.; Santos, J.I.; Luís, A.T.; Pereira, M.J.; Abrantes, N.; Gonçalves, F.J.; Pereira, J.L. Testing the response of benthic diatom assemblages to common riverine contaminants. *Sci. Total Environ.* **2021**, *755*, 142534. [\[CrossRef\]](#)
8. Zgłobicka, I.; Gluch, J.; Liao, Z.; Werner, S.; Guttman, P.; Li, Q.; Bazarnik, P.; Plocinski, T.; Witkowski, A.; Kurzydłowski, K.J. Insight into diatom frustule structures using various imaging techniques. *Sci. Rep.* **2021**, *11*, 14555. [\[CrossRef\]](#)
9. Kociolek, J.P.; Theriot, E.C.; Williams, D.M.; Julius, M.; Stoermer, E.F.; Kingston, J.C. Chapter 15—Centric and Araphid Diatoms. In *Aquatic Ecology, Freshwater Algae of North America*, 2nd ed.; John, D.W., Robert, G.S., Kociolek, J.P., Eds.; Academic Press: New York, NY, USA, 2015; pp. 653–708. [\[CrossRef\]](#)
10. Bueno, G.; Deniz, O.; Pedraza, A.; Ruiz-Santaquiteria, J.; Salido, J.; Cristóbal, G.; Borrego-Ramos, M.; Blanco, S. Automated Diatom Classification (Part A): Handcrafted Feature Approaches. *Appl. Sci.* **2017**, *7*, 753. [\[CrossRef\]](#)
11. Deng, J.; Guo, W.; Zhao, Y.; Liu, J.; Lai, R.; Gu, G.; Zhang, Y.; Li, Q.; Liu, C.; Zhao, J. Identification of diatom taxonomy by a combination of region-based full convolutional network, online hard example mining, and shape priors of diatoms. *Int. J. Leg. Med.* **2021**, *135*, 2519–2530. [\[CrossRef\]](#)
12. Yu, W.; Xiang, Q.; Hu, Y.; Du, Y.; Kang, X.; Zheng, D.; Shi, H.; Xu, Q.; Li, Z.; Niu, Y.; et al. An improved automated diatom detection method based on YOLOv5 framework and its preliminary study for taxonomy recognition in the forensic diatom test. *Front. Microbiol.* **2022**, *13*, 963059. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Li, J.; Jiang, C.; Yao, L.; Zhang, S. SC-Diatom Net: An Efficient and Accurate Algorithm for Diatom Classification. *J. Mar. Sci. Eng.* **2024**, *12*, 1862. [\[CrossRef\]](#)
14. Venkataramanan, A.; Faure-Giovagnoli, P.; Regan, C.; Heudre, D.; Figus, C.; Usseglio-Polatera, P.; Pradalier, C.; Laviale, M. Usefulness of synthetic datasets for diatom automatic detection using a deep-learning approach. *Eng. Appl. Artif. Intell.* **2023**, *117*, 105594. [\[CrossRef\]](#)
15. Wei, X.S.; Song, Y.Z.; Mac Aodha, O.; Wu, J.X.; Peng, Y.X.; Tang, J.H.; Yang, J.; Belongie, S. Fine-Grained Image Analysis with Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8927–8948. [\[CrossRef\]](#)
16. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 6999–7019. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Zhang, H.; Xu, T.; Elhoseiny, M.; Huang, X.L.; Zhang, S.T.; Elgammal, A. SPDA-CNN: Unifying Semantic Part Detection and Abstraction for Fine-Grained Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [\[CrossRef\]](#)
18. Wei, X.S.; Xie, C.W.; Wu, J.; Shen, C. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognit.* **2018**, *76*, 704–714. [\[CrossRef\]](#)
19. Guan, X.; Yang, Y.; Li, J.J.; Zhu, X.F.; Song, J.K.; Shen, H.T. On the imaginary wings: Text-assisted complex-valued fusion network for fine-grained visual classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 5112–5121. [\[CrossRef\]](#)
20. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014. [\[CrossRef\]](#)
21. Simon, M.; Rodner, E. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In Proceedings of the International Conference of Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015. [\[CrossRef\]](#)
22. Zhang, X.P.; Xiong, H.K.; Zhou, W.G.; Lin, W.Y.; Tian, Q. Picking deep filter responses for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [\[CrossRef\]](#)
23. Zhang, Y.; Wei, X.S.; Wu, J.X.; Cai, J.F.; Lu, J.B.; Nguyen, V.A.; Do, M.N. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Trans. Image Process. (TIP)* **2016**, *25*, 1713–1725. [\[CrossRef\]](#) [\[PubMed\]](#)
24. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [\[CrossRef\]](#)
25. Pu, S.Y.; Zhang, F.; Shu, Y.X.; Fu, W.Q. Microscopic image recognition of diatoms based on deep learning. *J. Phycol.* **2023**, *59*, 1166–1178. [\[CrossRef\]](#)

26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
27. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**. [[CrossRef](#)]
28. Szegedy, C.; Liu, W.; Jia, Y.Q.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [[CrossRef](#)]
29. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009. [[CrossRef](#)]
30. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the ECCV 2014, Zurich, Switzerland, 6–12 September 2014. [[CrossRef](#)]
31. Yan, P.; Shao, W. A Comparative Study on RepVGG and ResNet for Monocular 3D Object Detection. In Proceedings of the 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 5–7 November 2022. [[CrossRef](#)]
32. Haque, M.F.; Lim, H.Y.; Kang, D.S. Object Detection Based on VGG with ResNet Network. In Proceedings of the 2019 International Conference on Electronics, Information, and Communication (ICEIC), Auckland, New Zealand, 22–25 January 2019. [[CrossRef](#)]
33. Liu, Z.W.; Chen, B.S.; Zhang, A. Building segmentation from satellite imagery using U-Net with ResNet encoder. In Proceedings of the 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 25–27 December 2020. [[CrossRef](#)]
34. Hernanda, A.K.; Purnama, I.K.E.; Yuniarno, E.M.; Nugroho, J. Semantic Segmentation of Venous on Deep Vein Thrombosis (DVT) Case using UNet-ResNet. In Proceedings of the 2022 10th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, 2–3 August 2022. [[CrossRef](#)]
35. Chen, Y.; Bai, Y.; Zhang, W.; Mei, T. Destruction and construction learning for fine-grained image recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [[CrossRef](#)]
36. Ding, Y.; Zhou, Y.; Zhu, Y.; Ye, Q.; Jiao, J. Selective sparse sampling for fine-grained image recognition. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019. [[CrossRef](#)]
37. Zhuang, P.; Wang, Y.; Qiao, Y. Learning Attentive Pairwise Interaction for Fine-Grained Classification. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 7–12 February 2020. [[CrossRef](#)]
38. Song, J.; Yang, R. Feature boosting, suppression, and diversification for finegrained visual classification. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021. [[CrossRef](#)]
39. Yu, D.; Fang, Z.; Jiang, Y. Alleviating category confusion in fine-grained visual classification. *Vis. Comput.* **2025**, 1–16. [[CrossRef](#)]
40. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. [[CrossRef](#)]
41. Lu, L.; Cai, Y.C.; Huang, H.; Wang, P. An efficient fine-grained vehicle recognition method based on part-level feature optimization. *Neurocomputing* **2023**, *536*, 40–49. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.