fishes

MDPI

Article

Assessment Method for Feeding Intensity of Fish Schools Using MobileViT-CoordAtt

Shikun Liu, Xingguo Liu * and Haisheng Zou

Fishery Machinery and Instrument Research Institute, Chinese Academy of Fishery Sciences, Shanghai 200092, China; liushikun@fmiri.ac.cn (S.L.)
* Correspondence: liuxingguo@fmiri.ac.cn; Tel.: +86-021-65977260

Abstract: Assessment of fish feeding intensity is crucial for achieving precise feeding and for enhancing aquaculture efficiency. However, complex backgrounds in real-world aquaculture scenarios—such as water surface reflections, wave disturbances, and the stochastic movement of fish schools—pose significant challenges to the precise extraction of feeding-related features. To address this issue, this study proposes a fish feeding intensity assessment method based on MobileViT-CoordAtt. The method employs a lightweight MobileViT backbone network, integrated with a Coordinate Attention (CoordAtt) mechanism and a multi-scale feature fusion strategy. Specifically, the CoordAtt module enhances the model's spatial perception by encoding spatial coordinate information, enabling precise capture of the spatial distribution characteristics of fish schools. The multi-scale feature fusion strategy adopts a three-level feature integration approach (input features, local features, and global features) to further strengthen the model's representational capacity, ensuring robust extraction of key feeding-related features across diverse scales and hierarchical levels. Experimental results demonstrate that the MobileViT-CoordAtt model, trained with transfer learning, achieves an accuracy of 97.18% on the test set, with a compact parameter size of 4.09 MB. These findings indicate that the proposed method can effectively evaluate fish feeding intensity in practical aquaculture environments, providing critical support for formulating dynamic feeding strategies.

Keywords: fish schools; feeding intensity; deep learning; aquaculture; transformer; MobileViT

Key Contribution: This study proposes a method to assess the feeding intensity of fish schools and designs a dynamic feeding strategy to achieve precision feeding.

check for updates

Academic Editor: Michael E. Barnes

Received: 14 April 2025
Revised: 22 May 2025
Accepted: 26 May 2025
Published: 28 May 2025

Citation: Liu, S.; Liu, X.; Zou, H. Assessment Method for Feeding Intensity of Fish Schools Using MobileViT-CoordAtt. *Fishes* **2025**, *10*, 253. https://doi.org/10.3390/fishes10060253

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/).

1. Introduction

Aquaculture, as a globally significant protein source industry, incurs feed costs accounting for 45–65% of total production expenses. Rational control of feeding volumes is critical for reducing costs and improving efficiency [1]. However, current mainstream feeding practices still rely on empirical human judgment, lacking a quantitative regulatory mechanism based on dynamic fish feeding behaviors. Underfeeding prolongs growth cycles and intensifies intra-species competition [2], while overfeeding leads to feed waste, water pollution, and elevated ammonia concentrations from residual feed decomposition, posing severe risks to aquatic health [3]. Achieving timely and precise feeding remains a pressing challenge in aquaculture management.

Existing studies indicate that fish feeding activities generate distinctive collective behavioral patterns and hydrodynamic signals, providing a multimodal data foundation for

non-contact feeding intensity assessment [4]. Computer vision, as a non-invasive and non-contact technology, has advanced feeding behavior analysis. For instance, Sadoul et al. [5] developed a video analysis system using the group dispersion index (Ptot) and swimming activity index (Adiff) to quantify high-density fish behavior, enabling real-time welfare assessment. Liu et al. [6] analyzed the sum of differential frame intensities, defined overlap coefficients to calibrate errors, and filtered surface reflections to establish a computer vision-based feeding activity index. Zhou et al. [7] integrated near-infrared imaging with image enhancement, background subtraction, support vector machine-based reflection removal, and Delaunay triangulation to calculate fish aggregation indices, effectively quantifying feeding dynamics. Zhao et al. [8] improved a complex network model to classify Nile tilapia appetite levels into five tiers by analyzing dispersion, cohesion, and hydrodynamic disturbances. However, these methods rely on handcrafted features with limited robustness in complex scenarios (e.g., surface reflections, ripple noise), hindering practical engineering applications.

Advances in deep learning have significantly enhanced behavior analysis accuracy in complex environments [9]. Chen et al. [10] developed an MEA-BP neural network to automatically capture nonlinear relationships between environmental parameters (e.g., water temperature, dissolved oxygen) and feeding volumes, providing theoretical support for precision feeding systems. Adegboye et al. [11] fused multimodal sensor data and extracted spatiotemporal features via 8-direction chain coding and Fourier transforms to build an effective neural network for feeding behavior recognition. For visual feature modeling, Måløy et al. [12] proposed a dual-stream recurrent network (DSRN) that achieved 80% accuracy in salmon feeding behavior recognition through spatiotemporal fusion. Zhou et al. [13] employed data augmentation to construct a robust CNN model for tilapia feeding intensity classification, attaining 90% accuracy. Yang et al. [14] introduced a dual-attention network based on EfficientNet-B2 for fine-grained short-term feeding analysis, achieving 89.56% precision. Recent efforts focus on model lightweighting and accuracy optimization. Zhang et al. [15] designed a MobileNetV2-SENet hybrid network with feature weighting to improve recognition performance, while Feng et al. [16] developed a 3D ResNet-Glore model that achieved 92.68% accuracy in the four-tier feeding intensity classification, demonstrating the potential of deep learning in spatiotemporal feature extraction.

Despite these advancements, deep learning models face challenges in fish feeding analysis. Complex backgrounds and the small relative size of individual fish in images often lead to feature interference, reducing classification accuracy. While mainstream models improve performance by increasing network depth, this approach escalates parameter sizes, limiting practical deployment in resource-constrained aquaculture settings.

To address these limitations, this study proposes a MobileViT-CoordAtt-based method for fish feeding intensity assessment. Leveraging a lightweight MobileViT backbone, the model integrates a Coordinate Attention (CoordAtt) mechanism and a multi-scale feature fusion strategy to enhance key feature extraction across spatial and hierarchical scales while maintaining computational efficiency. Additionally, a dynamic feeding strategy is designed based on biomass and feeding intensity metrics.

The remainder of this paper is organized as follows: Section 2 details the materials and methods, Section 3 presents experimental results and analysis, and Sections 4 and 5 provide discussion and conclusions, respectively.

2. Materials and Methods

2.1. Overall Process

Accurately assessing the feeding intensity of fish schools and rationally controlling feed input are critical for ensuring healthy fish growth and enhancing the economic ef-

ficiency of aquaculture. Therefore, this study proposes a novel assessment method for feeding intensity in fish schools using MobileViT-CoordAtt and designs a dynamic feeding strategy to achieve precise feed input control. The overall process is outlined in Figure 1.

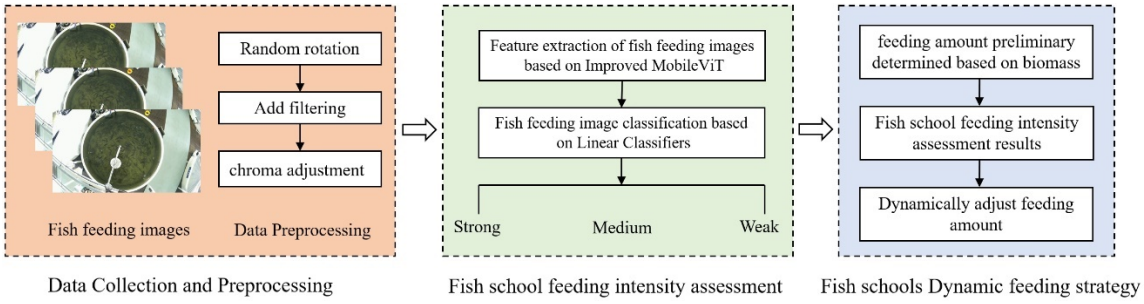


Figure 1. Overall process.

(1) Data Acquisition and Preprocessing. Feeding image data of fish schools were collected, and some preprocessing operations such as size adjustment, random cropping, random horizontal flip, random rotation, color jitter and gaussian blur, to enhance sample diversity and improve model generalization.
(2) Assessment of Feeding Intensity in Fish Schools. A feeding intensity assessment model for fish schools was developed based on MobileViT-CoordAtt. The model was trained using annotated fish school feeding images to extract pre-trained weights encoding feeding intensity features. The pre-trained weights were subsequently employed to assess the feeding intensity of the fish schools.
(3) Dynamic Feeding Strategy. A dynamic feeding strategy was developed by integrating biomass and fish feeding intensity. Before feeding, the total feed ration is determined based on the biomass within the culture area. During feeding, the feed amount is dynamically adjusted according to the assessed feeding intensity of the fish.

2.2. Data Acquisition and Preprocessing

Fish feeding intensity assessment experiments were conducted in the factory-scale aquaculture workshop of the Shanghai Chongming Aquaculture Base. The dataset was collected from rearing tanks with a diameter of 4 m and a water depth of 1.2 m, stocked with approximately 250 largemouth bass. A Hikvision DS-2CD3T86FWDV2-I3S (Hangzhou Hikvision Digital Technology Co., Ltd., Hangzhou, China) camera was mounted directly above each tank, about 4 m above the water surface, with its lens oriented vertically downward to record fish-feeding videos.

Video frames were extracted from the collected fish-feeding videos to construct the dataset. Because the videos were recorded at 20 fps, one frame was sampled every five frames to avoid redundancy, yielding RGB JPG images at 3840 × 2160 pixels. To improve image quality, only the rearing-tank region was cropped from each frame, producing 2160 × 2160 pixel fish-feeding images; the image-processing workflow is shown in Figure 2. Based on long-term on-site observations by experienced aquaculture practitioners and experimenters, the images were classified into three feeding intensity categories: strong feeding, characterized by densely aggregated fish vigorously competing for feed and generating substantial surface splashes; medium feeding, with partially aggregated fish actively feeding and producing moderate splashes; and weak feeding, where fish are dispersed and are minimally consuming surrounding pellets on a calm water surface [17,18]. Sample images of each category are shown in Figure 3.

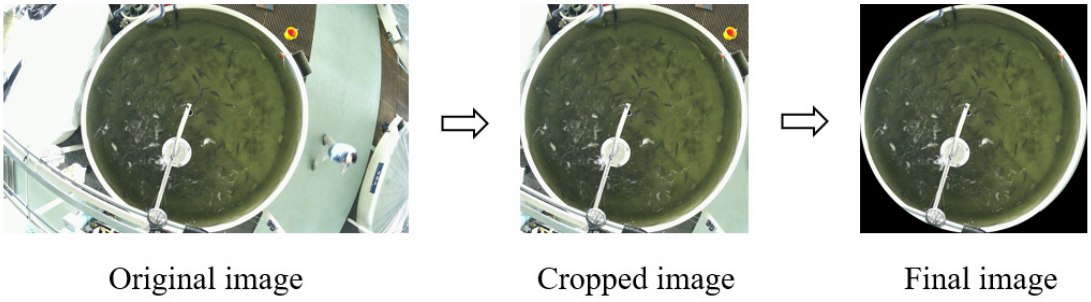


Figure 2. Image-processing workflow.

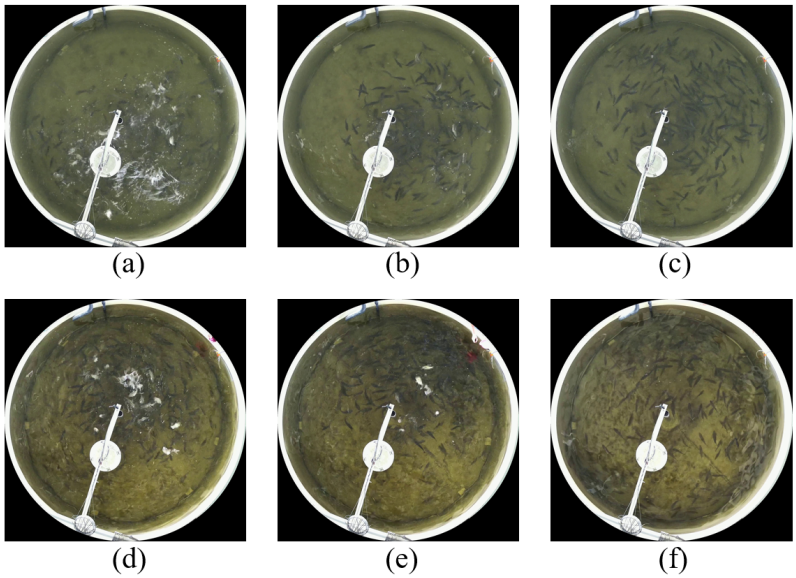


Figure 3. Fish-feeding images: (**a**,**d**) exhibit a strong feeding state; (**b**,**e**) exhibit a medium feeding state; and (**c**,**f**) exhibit a weak feeding state.

The fish-feeding image dataset comprises 4365 images, which were partitioned into training, validation, and test sets. The training and validation sets together account for 70% of the total, while the test set comprises the remaining 30%. Within the combined training and validation sets, images are further split in a 9:1 ratio. Table 1 summarizes the dataset partitioning.

Table 1. Dataset partitioning description.

Category	Training Set	Validation Set	Testing Set	Total
Strong	887	98	423	1408
Medium	879	97	420	1396
Weak	983	109	469	1561
Total	2749	304	1312	4365

The collected images of fish feeding behavior exhibit complex and variable background characteristics due to factors such as water surface reflections, motion blur, and illumination variations, posing significant challenges to accurate feeding intensity assessment. To address this, a multi-dimensional preprocessing and augmentation strategy was applied to the raw images. First, random cropping and rotation were employed to simulate diverse camera perspectives, with all images resized to a uniform 224 × 224 pixel resolution to align with model input requirements. Subsequently, random adjustments to brightness, contrast, saturation, and hue were implemented to mimic natural illumination fluctuations. Finally,

Gaussian blur was added to reduce noise interference. These operations were applied in randomized combinations to enhance image diversity, thereby improving the model's generalization capability and robustness in complex scenarios. The preprocessing workflow is illustrated in Figure 4, with detailed parameter configurations provided in Table 2.

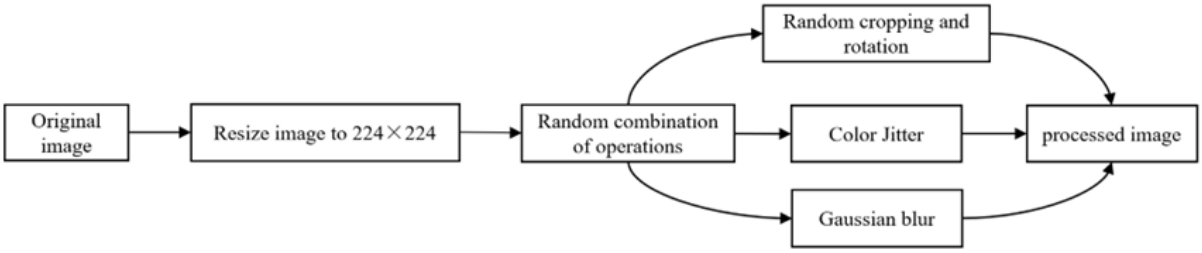


Figure 4. Flow chart of image preprocessing work.

Table 2. Parameter settings of preprocessing operations.

Operation	Parameter Setting	Parameter Description
Size Adjustment	size = 224	Images were resized to 224 × 224
Random Cropping	scale = (0.8, 1.0)	Random cropping was applied to the images, with the cropped area ranging from 80% to 100% of the original image area
Random Horizontal Flip	prob = 0.5	Images were horizontally flipped with a 50% probability
Random Rotation	deg = 15	Images were randomly rotated with rotation angles ranging from −15° to 15°
Color Jitter	params = [0.2, 0.2, 0.2]	Brightness, contrast, and saturation were adjusted by ±20% from their original values
Gaussian Blur	radius = 3	A 3 × 3 Gaussian kernel was applied to blur the images

2.3. Assessment of Feeding Intensity in Fish Schools

2.3.1. Image Feature Extraction Based on MobileViT-CoordAtt

Feature extraction from fish school feeding images involves deriving representative visual characteristics to comprehensively encode image information, serving as the computational foundation for feeding intensity recognition. MobileViT, a hybrid architecture integrating CNN and Transformer components proposed by Rastegari [19], demonstrates superior recognition accuracy while maintaining a compact model size suitable for edge deployment.

In practical aquaculture environments, fish-feeding images are often contaminated by substantial noise originating from complex backgrounds, uneven illumination, and dynamic surface ripples. Such noise progressively accumulates during deep feature extraction, ultimately degrading the model's discriminative capability. To address this challenge, this study proposes a fish-feeding image feature extraction network based on MobileViT-CoordAtt through structural improvement of the original MobileViT module. The network architecture is presented in Figure 5.

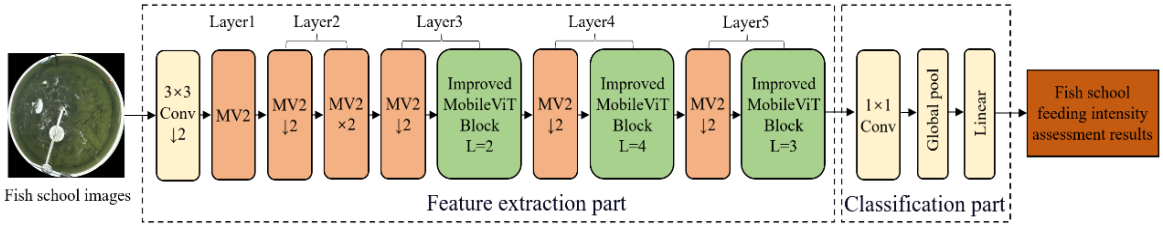


Figure 5. Network architecture of MobileViT-CoordAtt.

2.3.2. MobileNetV2 Block

The structure of the MobileNetV2 block [20], also known as the inverted residual block, is shown in Figure 6. First, a 1 × 1 convolution is used to expand the low-dimensional features to high-dimensional ones. Then, depthwise convolution is performed in the high-dimensional space to extract features. Finally, the features are compressed back to the low-dimensional space through a 1 × 1 convolution. It is a process of "Expansion Phase–Depthwise Convolution Phase–Projection Phase". When the sizes of the input and output feature maps are the same (stride s = 1), a skip connection is added. When the sizes of the input and output feature maps are different (stride s = 2), the skip connection is skipped.

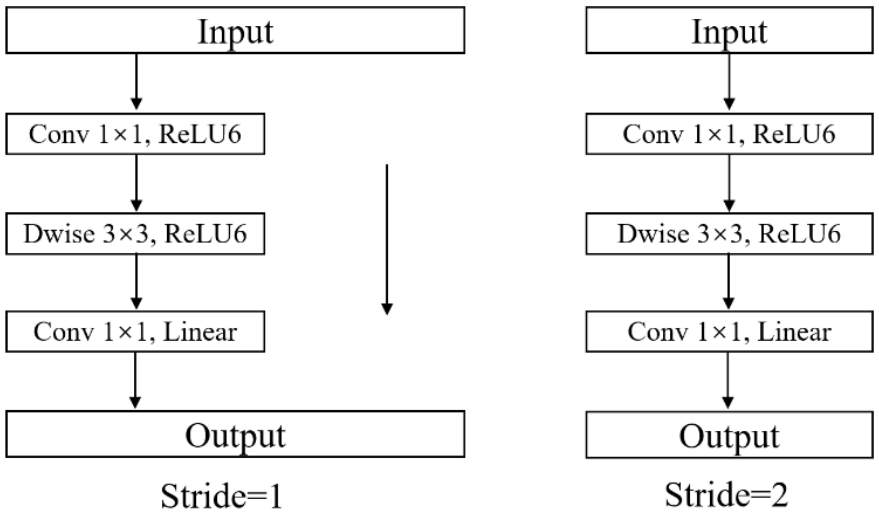


Figure 6. Structure of the MobileNetV2 block.

2.3.3. Improvement of the MobileViT Block

The MobileViT block comprises three core components: local representations for extracting fine-grained spatial features, global representations for capturing long-range dependencies, and a fusion module for integrating multi-scale information. To enhance feature discriminability while preserving the lightweight architecture, we strategically integrate a CoordAtt module between the global encoding and folding operations in the global representations branch. This integration enhances spatial position sensitivity during global context modeling. Furthermore, we redesign the fusion strategy by concatenating the input features, local representations, and enhanced global representations, thereby strengthening the composite feature expression capacity. Figure 7 shows the structure of the improved block, where L represents the number of transformers, Unfold represents the unfolding operation, and Fold represents the folding operation.

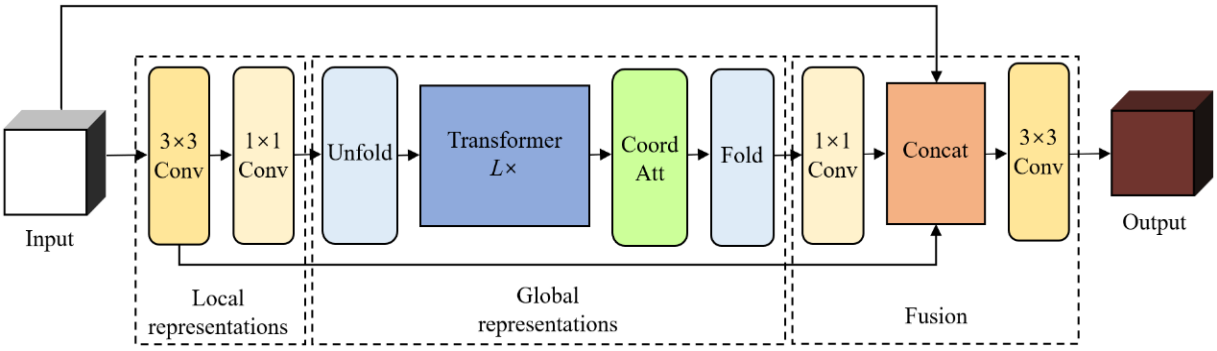


Figure 7. Structure of the improved MobileViT block.

(1) CoordAtt

CoordAtt (Coordinate Attention) [21] is an efficient attention mechanism that enhances model performance while preserving computational efficiency through decomposition of spatial dimensions into horizontal and vertical orientations for separate attention computation. This mechanism operates through two sequential phases: coordinate information embedding and coordinate attention generation, as shown in Figure 8.

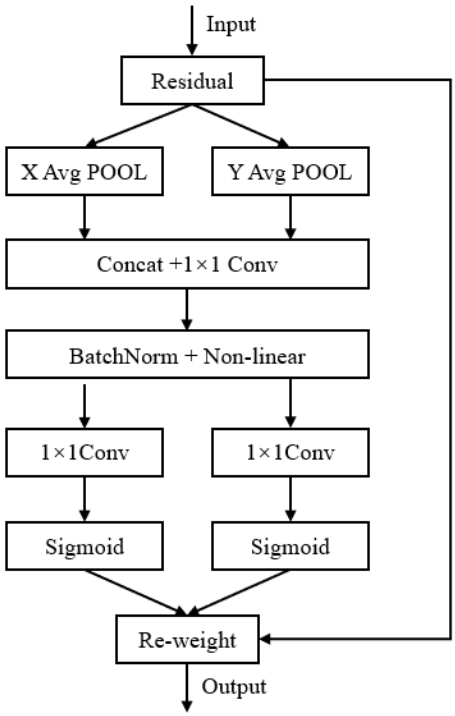


Figure 8. Schematic Architecture of CoordAtt mechanism.

In the coordinate information embedding operation, pooling kernels of size $(H, 1)$ and $(1, W)$ are used to encode each channel along the horizontal and vertical directions for the input feature map. The output of the c-th channel with height H is expressed as follows:

$$z_c^h(h) = \frac{1}{H} \sum_{0 \le i < W} X_C(J, W) \tag{1}$$

Similarly, the output of the c-th channel with width w is expressed as follows:

$$z_c^w(h) = \frac{1}{H} \sum_{0 \le j < H} X_C(J, W) \tag{2}$$

In the coordinate attention generation operation, the two feature maps produced in the horizontal and vertical directions from the previous step are first combined. Then, a shared 1×1 convolution is applied for dimensionality reduction to obtain an intermediate feature representation. Next, a nonlinear activation function is employed to split this intermediate feature into horizontal and vertical branches. Each branch is then processed with a 1×1 convolution followed by a Sigmoid function to generate the horizontal and vertical attention maps, respectively. Finally, these attention maps are multiplied element-wise with the input feature map, resulting in an enhanced feature map.

(2) Fusion Strategy Improvement

The original fusion strategy concatenated only the input feature and the global feature representation, neglecting local feature information. Therefore, the fusion strategy was improved by concatenating the input feature, local feature representation, and global feature representation to enhance the model's feature extraction capability.

Assuming the input feature map X of the MobileViT block has a size of $H \times W \times C$, where H, W, and C denote the height, width, and number of channels of the feature map, the input feature map is first processed by the 3×3 convolution in the Local Representations part to extract local features, yielding a feature map X_L with the size H $\times$ W $\times$ C. Then, X_L is passed through the 1×1 convolution in the Global Representations part to obtain a feature map X_T of size H $\times$ W $\times$ d, where d represents a higher-dimensional feature space. After that, X_T is fed into a 1×1 convolution to generate a feature map X_G with the size H $\times$ W $\times$ C. Finally, X, X_L, and X_G are concatenated to form the fused feature map X_F. The fusion formula is as follows:

$$X_F = Concat(X, X_L, X_G) \tag{3}$$

2.3.4. Image Classification Based on Linear Classifier

The fish-feeding image features based on MobileViT-CoordAtt are fed into a linear classifier to achieve the assessment of feeding intensity in fish-feeding images. The linear classifier makes classification decisions through a linear combination of features, consisting primarily of two components: a score function and a loss function. The score function converts the original fish-feeding images into scores for each category, providing a probability estimation of the image belonging to that category. The formula is as follows:

$$f(x_i, W, b) = Wx_i + b \tag{4}$$

where x_i is the image pixel matrix, W is the weight, and b is the bias vector.

The loss function measures the difference between the predicted labels and the true labels of the images. This paper uses the cross-entropy loss function, whose formula is as follows:

$$L_{cross}(V, T) = -\frac{1}{n}\sum_{i=1}^{n} y^i \ln\left(p\left(v^i\right)\right) \tag{5}$$

where V is the set of all samples in the training set, T is the set of true labels for all samples in the training set, n is the total number of samples in the dataset, and $p(v^i)$ is the output obtained from the i-th sample in the training set via the network.

2.3.5. Dynamic Feeding Strategy

In aquaculture management, reasonably determining the feeding amount is crucial to ensure the healthy growth of fish and farming efficiency. The feeding amount is influenced by multiple factors, such as biomass and the feeding status of fish schools. This study

proposes a dynamic feeding strategy for fish based on biomass and the assessment of the feeding intensity in the fish schools.

Biomass refers to the total weight of fish in the aquaculture area, and the feeding amount determined based on biomass reflects the physiological needs of all fish in the aquaculture area. Therefore, the biomass of the aquaculture area is first obtained, the initial feed quantity is determined by combining the feeding rate of the farmed fish, and this value is set as the maximum feeding amount. The formula is as follows [22]:

$$F = Biomass \times rate \tag{6}$$

where F is the maximum feeding amount, *Biomass* is the total weight of fish in the aquaculture area, and *rate* is the feeding rate.

The feeding intensity of fish reflects their level of feeding demand. When the physiological conditions of fish and environmental factors change, their feeding demand will vary, and their feeding intensity will also change accordingly. Therefore, during the feeding process, real-time analysis of fish feeding intensity is conducted, and feeding strategies are dynamically adjusted based on the analysis results.

During the entire feeding process, a multi-round feeding approach in a single session is adopted. First, the maximum feeding amount for the session is determined based on biomass. The feeding amount for the first round is set as 1/3 of the maximum amount. Starting from the 5th second after each round of feeding, images are collected every 2 s, with a total of 10 feeding images captured. These 10 images are evaluated using the fish feeding intensity assessment method proposed in this study, and the feeding amount for the next round is decided based on the evaluation results. In the evaluation results of the 10 feeding images: If the proportion of "strong" feeding intensity assessments is at least 60%, the next round's feeding amount is 1/3 of the maximum amount; if the proportion of "moderate" assessments is at least 60%, the next round's feeding amount is 1/6 of the maximum amount; if the proportion of "weak" assessments is at least 60%, or if the cumulative feeding amount in the session reaches the maximum amount, then feeding is stopped. The workflow of the dynamic feeding strategy is shown in Figure 9.

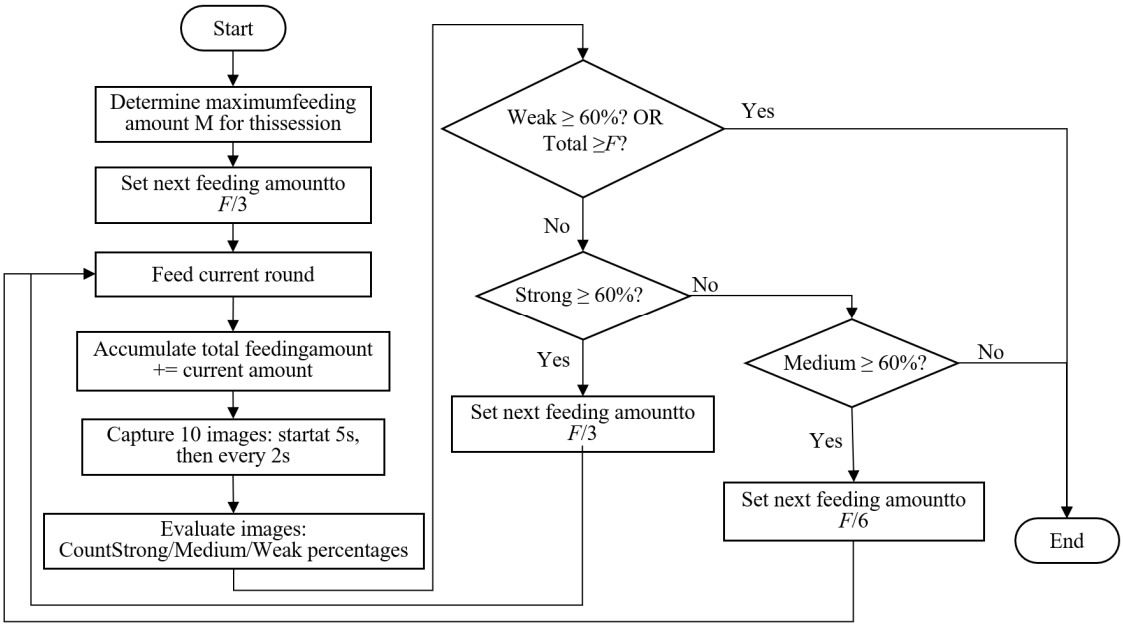


Figure 9. Dynamic feeding strategy workflow diagram.

2.4. Model Training and Evaluation Metrics

2.4.1. Model Training

To enhance the model's generalization capacity for fish school feeding intensity assessment in complex aquaculture environments, this study employs the transfer learning training method [23]. The specific approach is fine-tuning, which includes two stages: Freeze Training and Unfreeze Training. First, based on a model pre-trained on the ImageNet dataset, all network parameters except the classifier are frozen, with only the classification head undergoing gradient updates to achieve task-adaptive training in the initial stage; subsequently, the gradient update permissions for non-classifier parameters are gradually released, enabling the network to deeply optimize for the fine-grained features of fish feeding behavior while retaining pre-trained knowledge.

In terms of training parameter settings, the experiment adopts a cosine annealing learning rate decay strategy. The initial learning rate is set to 0.001, and the minimum learning rate is 0.00001. The Stochastic Gradient Descent (SGD) optimization algorithm is selected as the model optimizer. During the training process, the batch size is set to 64, and a total of 200 iterations of training are carried out. Among them, the first 50 iterations are freeze training, and the subsequent 150 iterations are unfreeze training.

The experimental setup was conducted on a high-performance computing platform equipped with an NVIDIA GeForce RTX 4080 GPU coupled with an Intel Core i9-14900K CPU and 64GB DDR5 RAM, running on an Ubuntu 20.04 LTS operating system. The software stack utilized the PyTorch 2.1.0 deep learning framework accelerated by CUDA 11.8, with all implementations developed in the Python 3.12 programming environment.

2.4.2. Evaluation Metrics

To validate the effectiveness of the proposed model in assessing fish school feeding intensity, the following metrics are employed: Accuracy, Recall, Precision, Specificity, and F1 score. The computational formulas for these metrics are as follows:

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

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

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

where TP, FP, FN, and TN represent true positives, false positives, false negatives, and true negatives, respectively.

3. Results

3.1. Model Training Results

The fish feeding intensity assessment model based on MobileViT-CoordAtt is trained using a transfer learning training method. The loss and accuracy curves on the training and validation sets are shown in Figure 10.

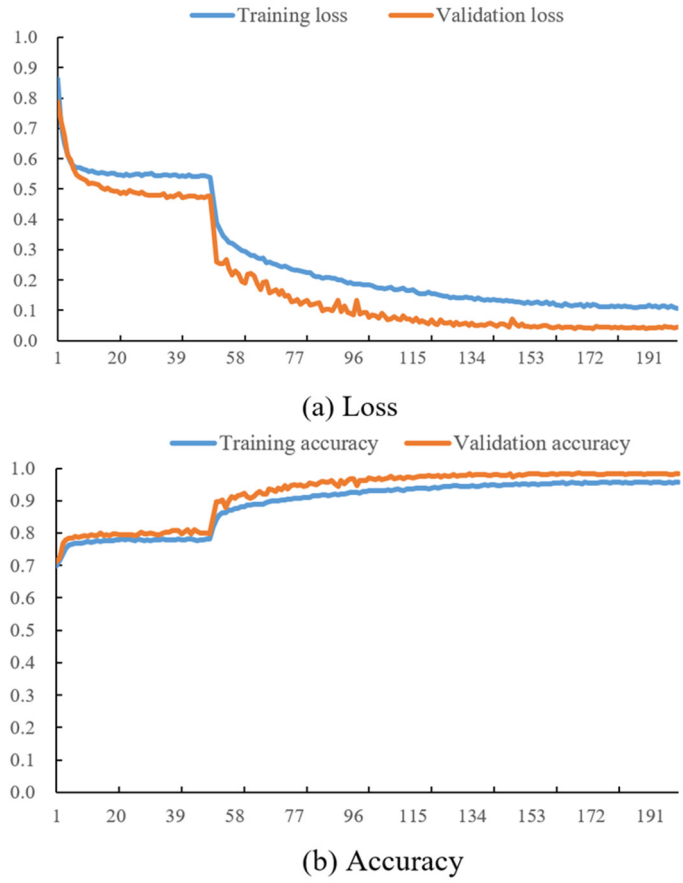


Figure 10. Loss and accuracy on the training and validation sets.

According to the loss rate curve in Figure 10a, during the freeze training phase, the loss rates on both the training and validation sets exhibited an exponential decline and converged to a stable state. After entering the unfreeze training phase, the training set loss rate continued to decrease at a relatively gradual pace, eventually stabilizing at a low level close to 0.1. In contrast, the validation set loss rate exhibited phased fluctuations, with reduced volatility in the later stage, ultimately converging to around 0.05 in synchronization—overall lower than the training set loss rate. As shown in the accuracy curve of Figure 10b, the accuracy on both the training and validation sets demonstrated an upward trend in both the freeze and unfreeze phases, increasing steadily throughout the process.

3.2. Model Evaluation Results

Using the optimal MobileViT-CoordAtt fish feeding intensity assessment model trained in this study, a total of 1312 images in the test set were evaluated, achieving a recognition accuracy of 97.18%. The evaluation metrics for each class are presented in Table 3, and the confusion matrix is shown in Figure 11. The results indicate that among 423 images with the actual class "strong," 417 were correctly identified as the strong class, 5 were misclassified as the moderate class, and 1 was misclassified as the weak class. Among 420 images with the actual class "moderate", 15 were misclassified as the strong class, 401 were correctly identified as the moderate class, and 4 were misclassified as the weak class. Among 469 images with the actual class "weak", 12 were misclassified as the moderate class, and 457 were correctly identified as the weak class. The model's precision, recall, specificity, and F1-score for the strong, moderate, and weak classes all exceed 95.48%, confirming the effectiveness of the proposed method.

Table 3. Evaluation results of model.

Category	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)
Strong	96.53	98.58	98.31	97.54
Medium	95.93	95.48	98.09	95.70
Weak	98.92	97.44	99.41	98.17

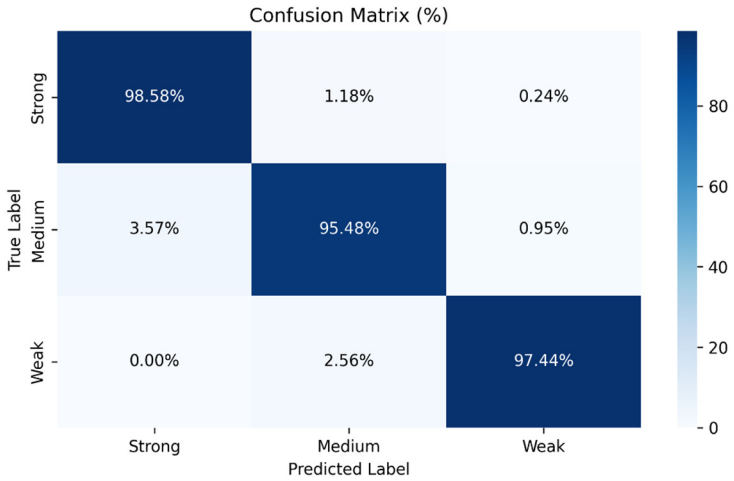


Figure 11. Confusion matrix of model on the test set.

The MobileViT-CoordAtt models trained for 60, 80, 120, and 200 epochs were used to assess 5 fish school feeding images in the test set, with assessment results shown in Figure 12. The numbers of correctly assessed images under different training epochs are 2, 3, 5, and 5. The results indicate that as the number of epochs increases, the model continuously learns through iterative training, and its assessment performance demonstrates a gradual optimization trend.

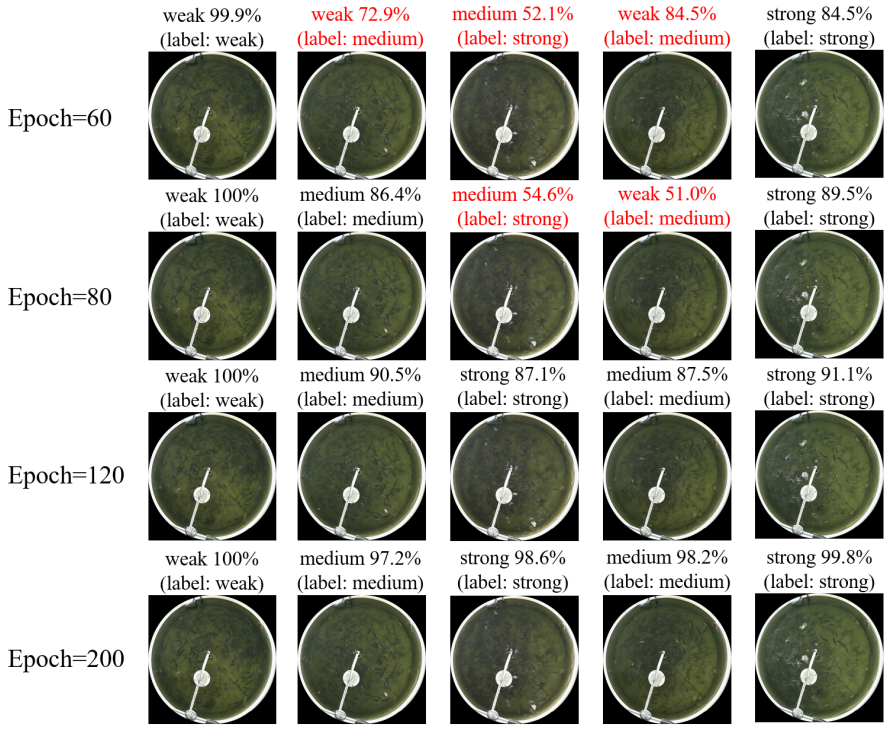


Figure 12. Assessment results of fish school feeding images for different training epochs.

3.3. Results of the Comparison Experiments

3.3.1. Comparison of Preprocessing Operation Effects

To validate the effectiveness of the preprocessing operation, a comparative analysis was conducted on the model's performance using datasets with and without the preprocessing operation applied. Experimental results show that the assessment accuracy of models with and without preprocessing operations are 97.18% and 88.17%. As presented in Figure 13, the evaluation index results for each category indicate that models with preprocessing operations outperform those without in four metrics: Precision, Recall, Specificity, and F1-Score. Among these, the largest difference is observed in the precision of the medium category, reaching 15.48%, while the smallest difference occurs in the specificity of the weak category, at 3.09%. The findings confirm that applying the preprocessing operation to the dataset can effectively enhance sample diversity, thereby improving model performance.

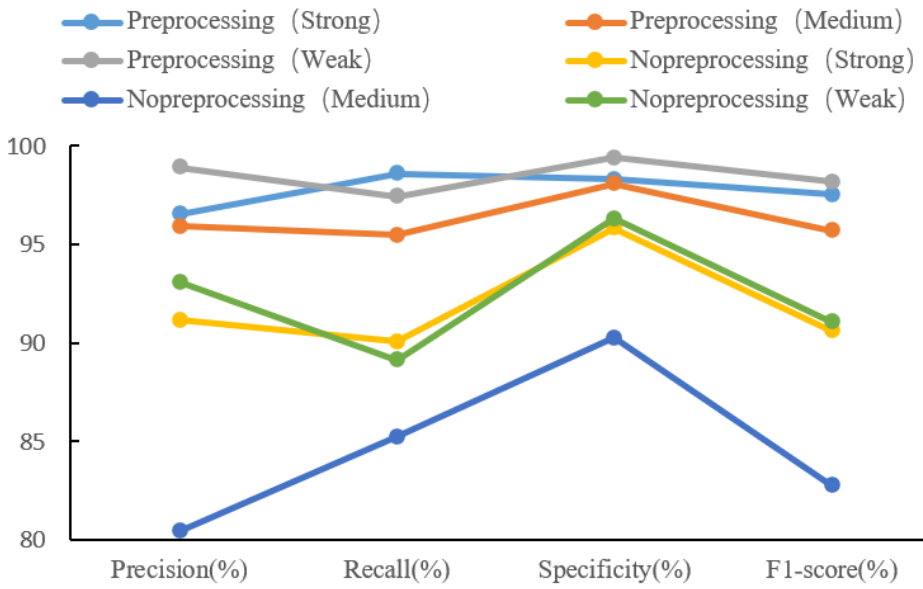


Figure 13. Comparison of evaluation metrics across categories.

3.3.2. Comparison of Transfer Learning Training Methods

To validate the performance of the improved model using transfer learning training, comparative experiments were conducted between transfer learning training and from-scratch learning training. Loss and accuracy curves on the training and validation sets are shown in Figure 14. Experimental results reveal that under transfer learning training, both the training loss and validation loss of the model decreased significantly faster and entered a stable convergence state more rapidly. In contrast, due to the lack of prior knowledge, the convergence process of from-scratch learning was more sluggish, with obvious fluctuations in the early training stage. Analysis of the loss curves shows that both the training loss and validation loss of the transfer learning-trained model were lower than those of the from-scratch learning-trained model, indicating that transfer learning can more efficiently optimize the model and reduce task loss through pre-trained knowledge. In contrast, from-scratch learning requires complete training from the ground up, leading to higher implementation difficulty and relatively weaker performance.

In terms of accuracy, transfer learning rapidly improved accuracy through pre-training and held an early advantage. Although from-scratch learning showed slow early-stage improvement, its final accuracy approached that of transfer learning after sufficient training, suggesting that both methods can achieve high recognition capabilities with long-term training; however, transfer learning is more efficient. On the test set comparison, the model

accuracies of transfer learning and from-scratch learning training were 97.18% and 94.05%, with the former outperforming the latter by 3.13%. Evaluation indices for each category are detailed in Table 4. By adopting transfer learning training methods, the model can not only enhance training efficiency but also effectively improve evaluation accuracy.

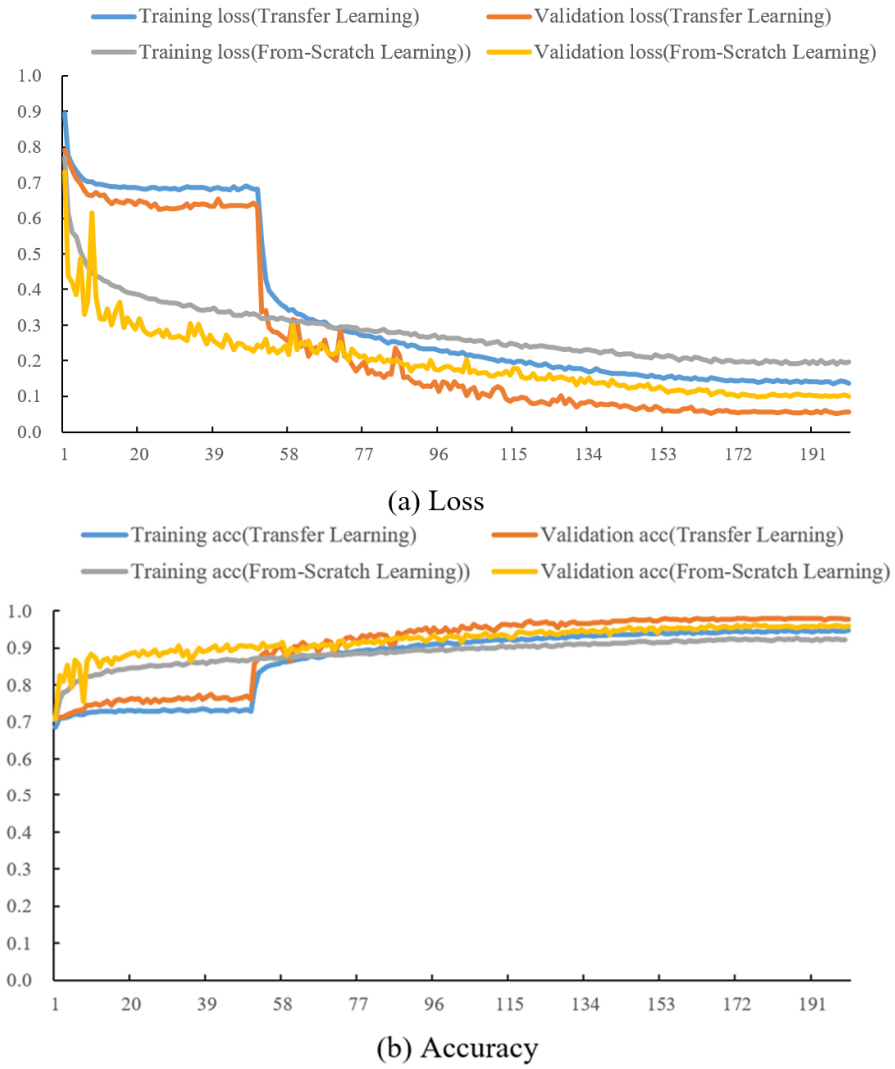


Figure 14. Loss and accuracy curves of different training methods on the training and validation sets.

Table 4. Evaluation metrics by category for different training methods.

Method	Category	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)
Transfer learning	Strong	96.53	98.58	98.31	97.54
	Medium	95.93	95.48	98.09	95.70
	Weak	98.92	97.44	99.41	98.17
From-scratch learning	Strong	94.99	94.09	97.64	94.54
	Medium	91.87	91.43	96.19	91.65
	Weak	95.16	96.38	97.27	95.76

3.3.3. Comparison of the Model Before and After Improvement

To systematically validate the performance of the enhanced MobileViT-CoordAtt model, comparative experiments were conducted across four variants: the baseline MobileViT (Model 1), MobileViT with only the multi-scale feature fusion strategy (Model 2), MobileViT with only the CoordAtt mechanism (Model 3), and the full MobileViT-CoordAtt (Model 4). The results demonstrated significant performance variations: Model 1 achieved

an accuracy of 96.27%; Model 2 showed a limited improvement at 96.65%, with a marginal gain of 0.38% over Model 1; Model 3 achieved 96.95%, marking a more substantial gain of 0.68% compared with Model 1, thereby confirming the critical role of CoordAtt in spatial position modeling for fish schools. Model 4 outperformed all variants with an accuracy of 97.18%, representing a 0.91% improvement over Model 1 (96.49%). This underscores the synergistic benefits of integrating multi-scale fusion and coordinate-aware attention mechanisms. Further validation of Model 4's superiority is provided by its enhanced performance across all evaluation metrics in Table 5.

Table 5. Performance evaluation results of different improved models.

Method	Category	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)
Model 1	Strong	96.88	95.51	98.54	96.19
	Medium	94.95	94.05	97.65	94.50
	Weak	96.87	98.93	98.22	97.89
Model 2	Strong	96.92	96.69	98.54	96.80
	Medium	93.97	96.43	97.09	95.18
	Weak	98.91	96.80	99.41	97.84
Model 3	Strong	98.78	95.51	99.44	97.12
	Medium	93.84	97.86	96.97	95.80
	Weak	98.28	97.44	99.05	97.86
Model 4	Strong	96.53	98.58	98.31	97.54
	Medium	95.93	95.48	98.09	95.70
	Weak	98.92	97.44	99.41	98.17

To visually reveal the role of the CoordAtt mechanism, a heatmap visualization analysis was conducted on the model. Comparisons of heatmaps for fish-feeding images at the Layer3 stage of the MobileViT and MobileViT-CoordAtt models are shown in Figure 15. The original MobileViT model was prone to interference from complex backgrounds during feature extraction; its heatmap focus regions were mixed with substantial background information, making it difficult for the model to accurately focus on fish feeding areas. By contrast, the MobileViT-CoordAtt model, with the CoordAtt mechanism introduced, exhibited significantly reduced complex background interference; its heatmap focus regions concentrated strictly on fish feeding areas, fully demonstrating the module's optimization effect on feature focusing capability.

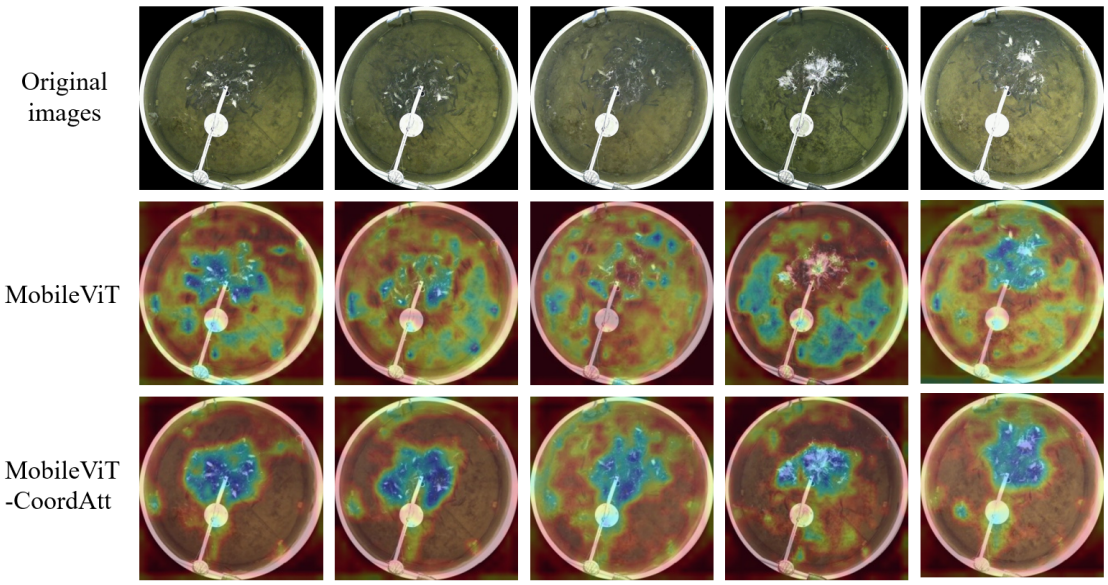


Figure 15. Heatmap at the Layer3 stage.

3.3.4. Comparison of Different Models

To further validate the performance of the MobileViT-CoordAtt model on a self-built dataset, comparative experiments were conducted with five neural network models: EfficientNetV2 [24], MobileNetV3_small [25], ResNet18 [26], Swin-Transformer [27], and Vision-Transformer [28]. The results are shown in Figure 16.

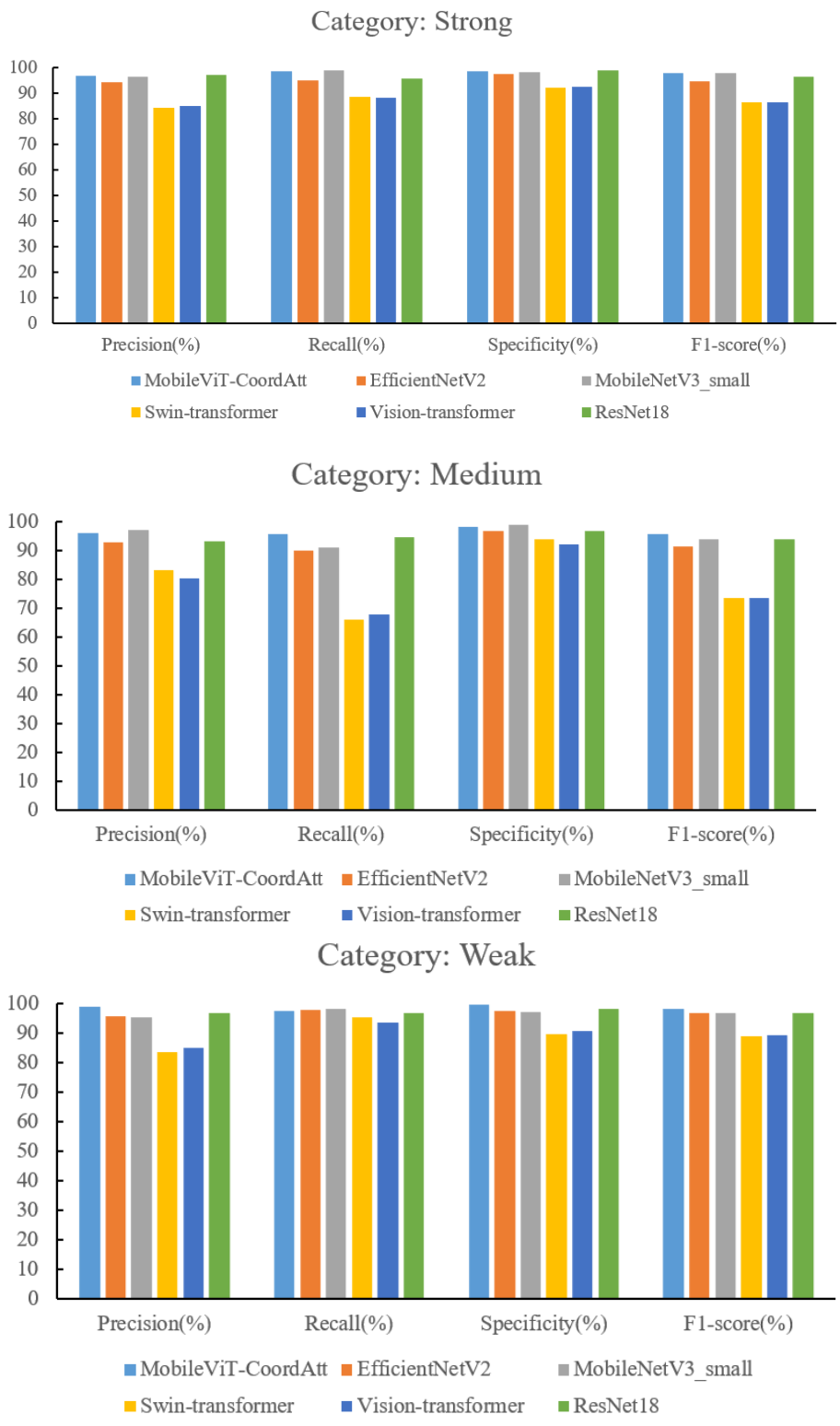


Figure 16. Comparison of evaluation metrics by category for different models.

The specific results of the MobileViT-CoordAtt model and other mainstream models in lightweight performance and multi-class classification performance are shown in Table 6. The MobileViT-CoordAtt model proposed in this study demonstrates remarkable comparative advantages in both memory efficiency and multi-class performance: it achieves a minimal memory footprint of only 4.09 MB, significantly outperforming other models in lightweighting; its overall classification accuracy reaches 97.18%—a notable improvement over competing architectures. Notably, the 97.18% accuracy of MobileViT-CoordAtt represents a 1.14% absolute improvement over the second-best model (MobileNetV3_small) and far surpasses both the Swin-Transformer and the Vision-Transformer. This highlights the model's unique ability to balance extreme lightweighting (4.09 MB) with high-performance classification, offering a more efficient and reliable solution for visual applications in resource-constrained scenarios.

Table 6. Comparison of lightweight and classification performance among models.

Model	Memory Footprint/MB	Category	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)
MobileViT-CoordAtt	4.09	Strong	96.53	98.58	98.31	97.54
		Medium	95.93	95.48	98.09	95.70
		Weak	98.92	97.44	99.41	98.17
EfficientNetV2	81.86	Strong	94.13	94.8	97.19	94.46
		Medium	92.63	89.76	96.64	91.17
		Weak	95.62	97.65	97.51	96.62
MobileNetV3 _small	9.7	Strong	96.31	98.82	98.20	97.55
		Medium	96.95	90.95	98.65	93.86
		Weak	95.04	98.08	97.15	96.54
Swin-Transformer	334.81	Strong	84.23	88.42	92.13	86.27
		Medium	82.93	65.95	93.61	73.47
		Weak	83.52	95.10	89.56	88.93
Vision-Transformer	330.23	Strong	84.77	88.18	92.46	86.44
		Medium	80.28	67.86	92.15	73.55
		Weak	84.91	93.6	90.75	89.05
ResNet18	44.59	Strong	97.12	95.51	98.65	96.31
		Medium	93.19	94.52	96.75	93.85
		Weak	96.60	96.80	98.10	96.70

4. Discussion

4.1. Improvement Efficacy of the Proposed Fish Feeding Intensity Evaluation Methods

The proposed MobileViT-CoordAtt-based method synergizes data augmentation, transfer learning, and model architecture optimization to achieve significant performance improvements. For data augmentation, preprocessing techniques such as Gaussian blur and random cropping enhanced training sample diversity, substantially improving the model's generalization capability in complex scenarios. Experimental results demonstrate that the model trained with augmented data achieved 97.18% accuracy on the test set, representing a 9.01% improvement over the model trained on raw data. In transfer learning, fine-tuning the model using weights pre-trained on the ImageNet dataset accelerated convergence and enhanced training efficiency, with notable advantages in training stability and loss optimization compared with training from scratch. Architecturally, MobileViT-CoordAtt outperformed the baseline MobileViT in classification accuracy and discriminative metrics by integrating the CoordAtt module for spatial feature enhancement and a multi-scale feature fusion mechanism. These results validate the method's effectiveness in optimizing data quality, training efficiency, and feature extraction capability.

4.2. Performance Comparison Across Models

Comparative experiments between the MobileViT-CoordAtt model and five mainstream models—EfficientNetV2, MobileNetV3_small, ResNet18, Swin-Transformer, and Vision-Transformer—demonstrated that MobileViT-CoordAtt achieves an optimal balance between lightweight design and multi-class recognition accuracy. With a parameter size of only 4.09 MB, the model is highly suitable for edge computing devices. In terms of accuracy, it achieved a superior overall classification accuracy of 97.18%, outperforming all counterparts. For fine-grained distinctions, it attained F1-scores of 98.17% and 95.70% for "weak feeding" and "moderate feeding" states, respectively. These results surpass most existing lightweight models, highlighting its potential for real-time, resource-constrained aquaculture applications.

4.3. Dynamic Feeding Strategy Optimization

Traditional aquaculture feeding strategies rely heavily on empirical farmer decisions, often leading to mismatches between feed allocation and actual demand due to inflexibility and low precision. While feeding behavior is influenced by multifaceted factors (e.g., weather, water quality, fish health), collecting and disentangling multimodal data remains challenging. This study introduces a dynamic feeding strategy centered on fish feeding intensity—a direct indicator of physiological status and environmental adaptation. The strategy operates via two mechanisms: (1) estimating a maximum theoretical feed volume based on biomass and feeding rates, and (2) dynamically adjusting subsequent allocations using a multi-phase feeding protocol. Specifically, if fish exhibit high feeding intensity in a given phase, the next allocation is set to one-third of the maximum volume; moderate intensity triggers a reduction to one-sixth, while feeding ceases if intensity further declines or the maximum volume is reached. This real-time feedback mechanism minimizes feed waste while meeting nutritional demands, effectively reducing operational costs.

4.4. Study Limitations

Several limitations warrant further investigation. First, the dataset was collected under specific aquaculture conditions (e.g., limited water quality parameters, weather scenarios, and fish species). Although preprocessing enhanced sample diversity, the model's generalizability to extreme environments (e.g., sudden temperature fluctuations, high turbidity, or hypoxic conditions) remains unverified. Future work should expand data collection to diverse environments and species, particularly addressing interspecific differences in feeding behavior. Second, while feeding intensity was modeled as a function of biomass and real-time behavior, the coupling effects of multifactorial interactions (e.g., fish stress levels, seasonal feeding rhythms, and feed composition) were not explicitly disentangled. For instance, abrupt changes in feed type may alter fish response patterns, which current strategies cannot dynamically adapt to. Integrating multimodal data fusion (e.g., combining visual behavior with acoustic signals of feeding activity or dissolved oxygen sensor data) could improve adaptability to complex scenarios.

5. Conclusions

This study proposes a fish feeding intensity evaluation method based on MobileViT-CoordAtt, which is crucial for improving aquaculture efficiency and promoting healthy fish growth. By introducing the CoordAtt module and improving the fusion strategy on the basis of the original MobileViT model, the method effectively enhances the model's sensitivity to spatial locations while strengthening its overall feature representation capability, leading to a more accurate evaluation of fish feeding intensity.

The method was validated on a real fish image dataset, achieving an accuracy of 97.18%, average precision of 97.17%, average recall of 97.13%, average specificity of 98.6%, and average F1-score of 97.14%. Compared with methods based on EfficientNetV2, MobileNetV3_small, Vision-Transformer, Swin-Transformer, and ResNet18, accuracy improved by 2.97%, 1.14%, 13.57%, 13.57%, and 1.52%; average precision improved by 3.04%, 1.07%, 13.61%, 13.85%, and 1.53%; average recall improved by 3.06%, 1.18%, 13.98%, 13.92%, and 1.52%; average specificity increased by 1.49%, 0.6%, 6.84%, 6.81%, and 0.77%; and the average F1-score increased by 3.06%, 1.16%, 14.25%, 14.13%, and 1.52%. This indicates that the method can accurately evaluate fish feeding intensity.

Furthermore, a dynamic feeding strategy is provided based on the evaluation results of fish feeding intensity. The maximum feeding amount is determined by biomass, and subsequent feeding quantities and feeding endpoints are dynamically adjusted according to the feeding intensity of the fish schools throughout the feeding process. Compared with traditional feeding quantity determination methods, the proposed dynamic feeding strategy effectively controls feeding amounts according to the feeding demands of fish, reduces aquaculture costs, and ensures healthy fish growth.

Author Contributions: Conceptualization, S.L.; methodology, S.L.; project administration, X.L.; resources, S.L. and H.Z.; software, S.L.; investigation, S.L., X.L., and H.Z.; visualization, S.L.; writing—original draft preparation, S.L.; writing—review and editing, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Shanghai Agricultural Science and Technology Innovation Program (Grant No. I2023006), the National Key Research and Development Program of China (No. 2023YFD2400502), Central Public-interest Scientific Institution Basal Research Fund, FMIRI of CAFS (NO. 2024YJS009), and Central Public-interest Scientific Institution Basal Research Fund, CAFS (NO. 2023TD67).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be made available upon request.

Acknowledgments: The authors thank the editor and anonymous reviewers for providing helpful suggestions for improving the quality of this manuscript.

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

References

1. Zhang, L.; Li, B.; Sun, X.; Hong, Q.; Duan, Q. Intelligent fish feeding based on machine vision: A review. *Biosyst. Eng.* **2023**, *231*, 133–164. [CrossRef]
2. Villes, V.S.; Durigon, E.G.; Hermes, L.B.; Uczay, J.; Peixoto, N.C.; Lazzari, R. Feeding rates affect growth, metabolism and oxidative status of Nile tilapia rearing in a biofloc system. *Trop. Anim. Health Prod.* **2024**, *56*, 208. [CrossRef] [PubMed]
3. Wang, M.; Li, F.; Wu, J.; Yang, T.; Xu, C.; Zhao, L.; Liu, Y.; Fang, F.; Feng, J. Response of CH_4 and N_2O emissions to the feeding rates in a pond rice-fish co-culture system. *Environ. Sci. Pollut. Res.* **2024**, *31*, 53437–53446. [CrossRef] [PubMed]
4. Liu, J.; Bienvenido, F.; Yang, X.; Zhao, Z.; Feng, S.; Zhou, C. Nonintrusive and Automatic Quantitative Analysis Methods for Fish Behaviour in Aquaculture. *Aquac. Res.* **2022**, *53*, 2985–3000. [CrossRef]
5. Sadoul, B.; Evouna Mengues, P.; Friggens, N.C.; Prunet, P.; Colson, V. A New Method for Measuring Group Behaviours of Fish Shoals from Recorded Videos Taken in near Aquaculture Conditions. *Aquaculture* **2014**, *430*, 179–187. [CrossRef]
6. Liu, Z.; Li, X.; Fan, L.; Lu, H.; Liu, L.; Liu, Y. Measuring Feeding Activity of Fish in RAS Using Computer Vision. *Aquac. Eng.* **2014**, *60*, 20–27. [CrossRef]
7. Zhou, C.; Zhang, B.; Lin, K.; Xu, D.; Chen, C.; Yang, X.; Sun, C. Near-Infrared Imaging to Quantify the Feeding Behavior of Fish in Aquaculture. *Comput. Electron. Agric.* **2017**, *135*, 233–241. [CrossRef]
8. Zhao, J.; Bao, W.J.; Zhang, F.D.; Ye, Z.Y.; Liu, Y.; Shen, M.W.; Zhu, S.M. Assessing Appetite of the Swimming Fish Based on Spontaneous Collective Behaviors in a Recirculating Aquaculture System. *Aquac. Eng.* **2017**, *78*, 196–204. [CrossRef]

9. Yang, X.; Zhang, S.; Liu, J.; Gao, Q.; Dong, S.; Zhou, C. Deep Learning for Smart Fish Farming: Applications, Opportunities and Challenges. *Rev. Aquac.* **2021**, *13*, 66–90. [CrossRef]
10. Chen, L.; Yang, X.; Sun, C.; Wang, Y.; Xu, D.; Zhou, C. Feed Intake Prediction Model for Group Fish Using the MEA-BP Neural Network in Intensive Aquaculture. *Inf. Process. Agric.* **2020**, *7*, 261–271. [CrossRef]
11. Adegboye, M.A.; Aibinu, A.M.; Kolo, J.G.; Aliyu, I.; Folorunso, T.A.; Lee, S.H. Incorporating Intelligence in Fish Feeding System for Dispensing Feed Based on Fish Feeding Intensity. *IEEE Access* **2020**, *8*, 91948–91960. [CrossRef]
12. Måløy, H.; Aamodt, A.; Misimi, E. A Spatio-Temporal Recurrent Network for Salmon Feeding Action Recognition from Underwater Videos in Aquaculture. *Comput. Electron. Agric.* **2019**, *167*, 105087. [CrossRef]
13. Zhou, C.; Xu, D.; Chen, L.; Zhang, S.; Sun, C.; Yang, X.; Wang, Y. Evaluation of Fish Feeding Intensity in Aquaculture Using a Convolutional Neural Network and Machine Vision. *Aquaculture* **2019**, *507*, 457–465. [CrossRef]
14. Yang, L.; Yu, H.; Cheng, Y.; Mei, S.; Duan, Y.; Li, D.; Chen, Y. A Dual Attention Network Based on EfficientNet-B2 for Short-Term Fish School Feeding Behavior Analysis in Aquaculture. *Comput. Electron. Agric.* **2021**, *187*, 106316. [CrossRef]
15. Zhang, L.; Wang, J.; Li, B.; Liu, Y.; Zhang, H.; Duan, Q. A MobileNetV2-SENet-Based Method for Identifying Fish School Feeding Behavior. *Aquac. Eng.* **2022**, *99*, 102288. [CrossRef]
16. Feng, S.; Yang, X.; Liu, Y.; Zhao, Z.; Liu, J.; Yan, Y.; Zhou, C. Fish Feeding Intensity Quantification Using Machine Vision and a Lightweight 3D ResNet-GloRe Network. *Aquac. Eng.* **2022**, *98*, 102244. [CrossRef]
17. Liu, J.; Becerra, A.T.; Bienvenido-Barcena, J.F.; Yang, X.; Zhao, Z.; Zhou, C. CFFI-Vit: Enhanced Vision Transformer for the Accurate Classification of Fish Feeding Intensity in Aquaculture. *J. Mar. Sci. Eng.* **2024**, *12*, 1132. [CrossRef]
18. Zhang, L.; Liu, Z.; Zheng, Y.; Li, B. Feeding intensity identification method for pond fish school using dual-label and MobileViT-SENet. *Biosyst. Eng.* **2024**, *241*, 113–128. [CrossRef]
19. Mehta, S.; Rastegari, M. MobileViT_Light-weight__General-purpose__and_Mobile-friendly_Vision_Transformer. In Proceedings of the Tenth International Conference on Learning Representations, Virtual, 25 April 2022.
20. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2_Inverted_Residuals_and_Linear_Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
21. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
22. Li, H.; Chen, Y.; Li, W.; Wang, Q.; Duan, Y.; Chen, T. An adaptive method for fish growth prediction with empirical knowledge extraction. *Biosyst. Eng.* **2021**, *212*, 336–346. [CrossRef]
23. Zhu, Z.; Lin, K.; Jain, A.K.; Zhou, J. Transfer Learning in Deep Reinforcement Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13344–13362. [CrossRef] [PubMed]
24. Tan, M.; Le, Q.V. EfficientNetV2 Smaller Models and Faster Training. In Proceedings of the Thirty-Eighth International Conference on Machine Learning, Virtual, 18–24 July 2021.
25. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
26. He, K.; Zhang, X.; Ren, S.; 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, 7–30 June 2019.
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words Transformers for Image Recognition at Scale. In Proceedings of the Ninth International Conference on Learning Representations, Virtual, 3–7 May 2021.

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.