Food Image Recognition Based on Anti-Noise Learning and Covariance Feature Enhancement

Zengzheng Chen †, Hao Chen †, Jianxin Wang *, Yeru Wang *

Posted Date: 16 July 2025

doi: 10.20944/preprints202507.1310.v1

Keywords: food recognition; anti-noise learning; covariance feature enhancement; knowledge distillation; food science and technology

 doi:10.20944/preprints202507.1310.v1

Article

Food Image Recognition Based on Anti-Noise Learning and Covariance Feature Enhancement

Zengzheng Chen [1,†], Hao Chen [1,†], Jianxin Wang [1,*] and Yeru Wang [2,*]

[1] School of Information, Beijing Forestry University, Beijing 100083, China

[2] Risk Assessment Division 1, China National Center for Food Safety Risk Assessment, Beijing 100022, China

* Correspondence: wangjx@bjfu.edu.cn (J.W.); wangyeru@cfsa.net.cn (Y.W.)

† These authors contributed equally to this work.

Abstract

Food image recognition is a key research area in food computing, with applications in dietary assessment, menu analysis, and nutrition monitoring. However, imaging devices and environmental factors introduce noise, limiting classification performance. To address this, we propose a food image recognition method based on anti-noise learning and covariance feature enhancement. Specifically, we design a Noise Adaptive Recognition Module (NARM), which incorporates noisy images during training and treats denoising as an auxiliary task to enhance noise invariance and recognition accuracy. To mitigate the adverse effects of noise and strengthen the representation of small eigenvalues, we introduce Eigenvalue-Enhanced Global Covariance Pooling (EGCP) into NARM. Furthermore, we develop a Weighted Multi-Granularity Fusion (WMF) method to improve feature extraction. Combined with the Progressive Temperature-Aware Feature Distillation (PTAFD) strategy, our approach optimizes model efficiency without adding overhead to the backbone. Experimental results demonstrate that our model achieves state-of-the-art performance on the ETH Food-101 and Vireo Food-172 datasets. Specifically, it reaches a Top-1 accuracy of 92.57% on ETH Food-101, outperforming existing methods, and also delivers strong results in Top-5 on ETH Food-101 and both Top-1 and Top-5 on Vireo Food-172. These findings confirm the effectiveness and robustness of the proposed approach in real-world food image recognition.

Keywords: food recognition; anti-noise learning; covariance feature enhancement; knowledge distillation; food science and technology

1. Introduction

With the rapid development of artificial intelligence technologies, food image recognition has become one of the core technologies in food computing, demonstrating broad application prospects in dietary assessment, menu analysis, and nutritional tracking [1–3]. This technology not only allows people to more conveniently track their dietary intake and analyze nutritional components, but also offers precise support for the health management of patients with chronic diseases [4–7]. Meanwhile, food image recognition is especially significant in the restaurant industry and food delivery platforms, offering robust technical support for features such as menu recommendations, dish classification, and food traceability, effectively enhancing operational efficiency and user experience [8–11]. Furthermore, this technology's application in the food industry has garnered significant attention, particularly in areas such as food classification, quality testing, and labeling, showing substantial practical value and offering strong support for the intelligent modernization of food production and processing workflows [12–14].

However, achieving high-accuracy food image recognition is not straightforward. This technology is a crucial subfield of fine-grained image recognition, which encounters the dual challenges of inter-class similarity and intra-class variability [15,16]. For instance, foods with similar appearances (e.g., salads and vegetable stir-fries) are prone to confusion, while differences within the

 doi:10.20944/preprints202507.1310.v1

same category due to different cooking methods (e.g., steamed fish vs. fried fish) further increase classification difficulty. To address these challenges, the "attention mechanism" is widely used by researchers to improve the effectiveness of food image recognition. The "attention mechanism" enhances classification accuracy by identifying and emphasizing important feature information in the image that is relevant to the classification task [17–19]. For example, some studies developed several attention modules that can be integrated into deep neural networks to improve focus on key features and enhance discriminative power [18,19]. Although the attention mechanism has successfully improved food image classification accuracy, its performance still faces some limitations. For instance, this mechanism is highly dependent on the accuracy of region localization, and its performance may deteriorate in complex backgrounds or noisy environments. Moreover, while many methods perform well in laboratory settings, they often lack the necessary robustness in practical applications, making it challenging to handle the diverse types of noise interference present in real-world scenarios.

Noise interference is one of the core challenges that urgently needs to be addressed in food image recognition. In real-world image capturing, food images are frequently disturbed by noise due to lighting conditions, equipment quality, and background complexity. These noises not only reduce the efficiency of feature extraction but also weaken the deep learning model's ability to precisely capture key features, significantly affecting classification performance. Particularly in fine-grained classification tasks, the presence of noise increases the risk of misjudging subtle features, further complicating the task. While convolutional neural networks (CNNs) are currently the dominant tool in image recognition, they are susceptible to noise. Even minute noise that is almost imperceptible to the human eye can cause significant deviation in CNN classification results. Figure 1 shows that subtle noise interference can directly lead to classification errors, highlighting the severity of this issue. Therefore, reducing the impact of noise interference on food image recognition performance is a problem that urgently needs to be solved.

Based on the above discussion, we proposed a food image recognition method that combines noise-robust learning and covariance feature enhancement. We also designed a Noise Adaptive Recognition Module (NARM) to address the issues of complex backgrounds and noise interference. Specifically, during the training phase, we actively injected Gaussian noise into the images. Then, we input the noisy images into the backbone network, extracted multi-level feature vectors through the convolutional neural network (CNN), and introduced them into the NARM. NARM adopts a multi-task learning framework consisting of two core components: the Adaptive Recognition Unit (ARU) and the Restorative Transformation Unit (RTU). The ARU is designed to boost the model's capability in identifying critical features and enhancing its classification accuracy. At the same time, the RTU is responsible for denoising the image and restoring the disturbed feature information.

In the ARU, we began by using convolutional operations and activation functions for initial feature extraction. Since noise has been actively added to the input image, direct classification may reduce recognition accuracy. To address this, we have innovatively designed an Enhanced Global Covariance Pooling (EGCP) to replace the traditional global max pooling layer. By fully exploiting the potential class-discriminative information in small eigenvalues, EGCP allows the model to retain global features while better capturing category-related details, thus improving the accuracy and robustness of fine-grained classification tasks. EGCP applies regularization, dynamic scaling, and feature enhancement to the covariance matrix, which not only alleviates the negative effects of noise but also significantly enhances the expressiveness of small eigenvalues. In the RTU, we progressively restore the feature vectors to the original image size using convolutional operations and pixel rearrangement. To address potential loss of detailed information during this process, we incorporate a Low-level Feature Enhancement (LFE) module to better preserve fine details. Finally, we guided the model in learning the invariance of noise by calculating the mean squared error between the denoised image and the original image, thereby improving its adaptability to complex noisy environments. With the collaboration of ARU and RTU, this method effectively reduces the impact

 doi:10.20944/preprints202507.1310.v1

of noise on food image recognition, significantly improving the model's performance in real-world scenarios, thereby enabling more robust and efficient recognition.

To further enhance the noise resistance of the recognition model, we propose the Weighted Multi-granularity Fusion (WMF) method. In standard CNN, we introduce multiple NARMs to improve the noise resistance of different network layers at various depths. We also extract more valuable feature representations by weighted fusion of multi-layer features. As CNN progressively abstracts low-level information into high-level features, the performance of deep layers heavily depends on the quality of shallow features. To avoid competition between shallow and deep NARMs, we design a progressive learning strategy to ensure reasonable gradient propagation and improve overall performance. The model excels at handling intra-class variation caused by noise. However, the WMF method introduces a large number of parameters compared to traditional backbone networks (such as ResNet50 [20]). To address this, we propose the Progressive Temperature-Aware Feature Distillation (PTAFD) method, which efficiently transfers the knowledge of the WMF method to traditional CNN through feature distillation. PTAFD aligns the features of a standard backbone CNN with the WMF method during training and ensures that the final prediction distribution closely matches the WMF method's distribution. This method effectively enhances the computational efficiency and practicality of the model while maintaining its performance.

In summary, the key contributions of this paper can be highlighted as follows: First, we introduce a novel food image recognition approach that integrates noise-resilient learning with enhanced covariance features. We designed the NARM and introduced the EGCP to mitigate noise interference in feature extraction. This approach significantly enhances fine-grained classification performance and model robustness. Second, we propose the WMF method combined with the PTAFD method. This approach reduces the number of model parameters while enhancing noise resistance and classification performance in complex scenarios, providing a feasible solution for lightweight deployment. Moreover, the enhanced capabilities of the proposed model are confirmed through evaluations on the ETH Food-101 [21] and Vireo Food-172 [22] food datasets. The experimental outcomes demonstrate that our approach achieves better classification accuracy compared to current leading methods.

Dataset	ETH Food-101		Vireo Food-172	
Original image				
Noisy image				
Original label	Fried rice	Clam chowder	Chopped chicken	Stir-fried eggs with green peppers
Predicted label	Risotto	Lobster bisque	Mouth-watering chicken	Stir-fried eggs with chives

Figure 1. Convolutional neural networks (CNNs) can be affected by noise when performing food image recognition tasks. We used the ResNet50[20] model to train on the ETH Food-101[21] and Vireo Food-172[22] data sets, respectively, and used the trained network to identify the original test set and the test images with a standard deviation of 0.01 Gaussian noise. Even though the human eye may not detect significant differences between the original and perturbed images, CNNs tend to misclassify some of these perturbed images.

2. Related Work

 doi:10.20944/preprints202507.1310.v1

In recent years, substantial advancements have been achieved in the recognition of food images and noise-resilient technologies, making it an increasingly prominent research focus within the domain of image recognition. Food image recognition focuses on improving the model's classification ability under complex backgrounds and diverse food categories, while noise-robust technology focuses on enhancing the robustness and stability of recognition systems in noisy environments. This section will review the latest research achievements in this field, analyze the advantages and disadvantages of existing methods, and provide a theoretical foundation for the proposed approach in this paper.

2.1. Food Image Recognition

Food image recognition has emerged as a key research area in the field of computer vision and has garnered growing attention in recent years. Traditional approaches typically depend on manually designed feature extraction [23–29], including color descriptors (SCD, DCD) for capturing image color information [23], texture descriptors (FDE, EBD) for describing image texture features [24], and local feature methods (e.g., SIFT, multi-scale dense SIFT) for extracting local features. However, these methods are limited in their ability to handle complex images. For example, classification accuracy often decreases when food images have complex backgrounds or numerous categories. Furthermore, handcrafted feature extraction designs are rigid, making it difficult to adapt to dynamic application scenarios, which limits their practical performance.

As deep learning, especially convolutional neural networks (CNNs), has advanced rapidly, substantial progress has been made in the field of food image recognition technology. CNNs can automatically learn high-level image features and demonstrate outstanding performance on large-scale datasets. In recent years, a number of novel approaches have been introduced. Shah et al. [30] proposed a depth-restricted convolutional neural network (DRCNN), which achieved substantial enhancements in classification accuracy by incorporating batch normalization techniques and optimizing model parameters. Ying-Chieh Liu et al. [31] developed a multi-dish food recognition model based on EfficientDet, which performed exceptionally well on a dataset of 87 Taiwanese local dishes, demonstrating its potential for integration into mobile and cloud applications. Moreover, Chenglin Wang et al. [32] examined the utilization of CNNs in fresh fruit cultivation, highlighting the significant potential of deep learning techniques in agriculture. Certain studies have developed attention modules that integrate seamlessly into deep neural networks to strengthen attention to critical features and improve discriminative abilities. For example, Wang et al. [17] proposed a novel deep learning framework that combines elements of EfficientNet, the Swin Transformer, and the Feature Pyramid Network (FPN). Utilizing attention mechanisms to effectively capture long-range dependencies within images has notably enhanced the accuracy and efficiency of recognizing food nutrition. Alahmari et al. [18] also proposed a method using attention mechanisms, integrating attention masks into deep learning models to enable efficient and accurate recognition of food types and states. Additionally, Sreedharan et al. [19] developed a convolutional neural network model called NutriFoodNet, aimed at automating food image recognition and nutrient estimation while utilizing attention mechanisms to enhance model performance.

Although food image recognition research has made remarkable progress, specialized research on denoising food images remains insufficient. The influence of noise on feature extraction is especially evident in settings involving complex backgrounds or low-light environments. Therefore, applying noise-robust techniques in food image recognition holds significant value. To address this, we proposed a method that combines noise-robust learning with covariance feature enhancement, and we designed a NARM to improve robustness against noise. Furthermore, we proposed a Weighted Multi-Granularity Fusion (WMF) approach to merge features across various scales. Combined with Progressive Temperature-Aware Feature Distillation (PTAFD), this method improves recognition performance and model practicality.

2.2. Application of Anti-Noise Technology in Image Recognition

 doi:10.20944/preprints202507.1310.v1

In recent years, significant progress has been made in the research of anti-noise techniques in image recognition, providing diverse solutions to enhance recognition performance and system robustness [33–38]. The noise issue in image recognition often severely interferes with feature extraction and representation, especially in complex scenes or constrained hardware environments. Therefore, anti-noise techniques have become an essential means of ensuring the reliability of recognition systems. Deep learning-based anti-noise methods have attracted considerable attention and shown remarkable performance advantages. For instance, the Swin Transformer for Image Restoration (SwinIR) proposed by Zhang et al. [35] enhances anti-noise performance by introducing adaptive modeling of global and local features, excelling in detail retention and adapting to various noise distributions. The Denoising Diffusion Probabilistic Model (DDPM) developed by Xie et al. [36], based on diffusion models, achieved a breakthrough in anti-noise technology by gradually restoring the noise data distribution from a generative perspective. These methods, leveraging the nonlinear feature extraction capability of deep learning, perform excellently in image restoration tasks under complex noise environments.

At the same time, optimizing traditional anti-noise algorithms has opened new directions for image recognition. Liu et al. [37] proposed Fast Block-Matching and 3D Collaborative Filtering (Fast BM3D) based on the classic BM3D algorithm, which significantly reduces computational complexity while improving processing efficiency by enhancing block-matching and filtering processes. Li et al. [38] combined traditional Non-local means (NLM) with CNN, proposing the NLM-CNN method, which integrates the advantages of non-local similarity and deep feature representation, successfully alleviating the performance degradation of traditional methods in high-noise image scenarios.

Inspired by the above studies, we innovatively proposed the EGCP to address the weakness in feature representation in existing anti-noise techniques. This method enhances the influence of small eigenvalues, improving feature representation capability, thus effectively capturing weak yet discriminative detail information in food images while ensuring the integrity of overall feature representation, providing a novel technical approach for the anti-noise performance of image recognition.

3. Approach

This research introduces several innovative approaches to tackle the noise problems in food image classification and restoration, including the Noise Adaptive Recognition Module (NARM), the Weighted Multi-Granularity Fusion (WMF) method, and the Progressive Temperature-Aware Feature Distillation (PTAFD) method. These methods are designed to address noise processing, feature fusion, and model efficiency optimization, with detailed explanations provided in the following sections.

3.1. Noise Adaptive Recognition Module

Figure 2 shows the design of NARM. Initially, Gaussian noise was injected into the original image to generate a Noisy Image. The Noisy Image is subsequently fed into the CNN backbone network. In this paper, the backbone network uses the ResNet50 architecture. The output feature vector from the backbone network is input into both the ARU and the RTU for image recognition and denoising.

 doi:10.20944/preprints202507.1310.v1

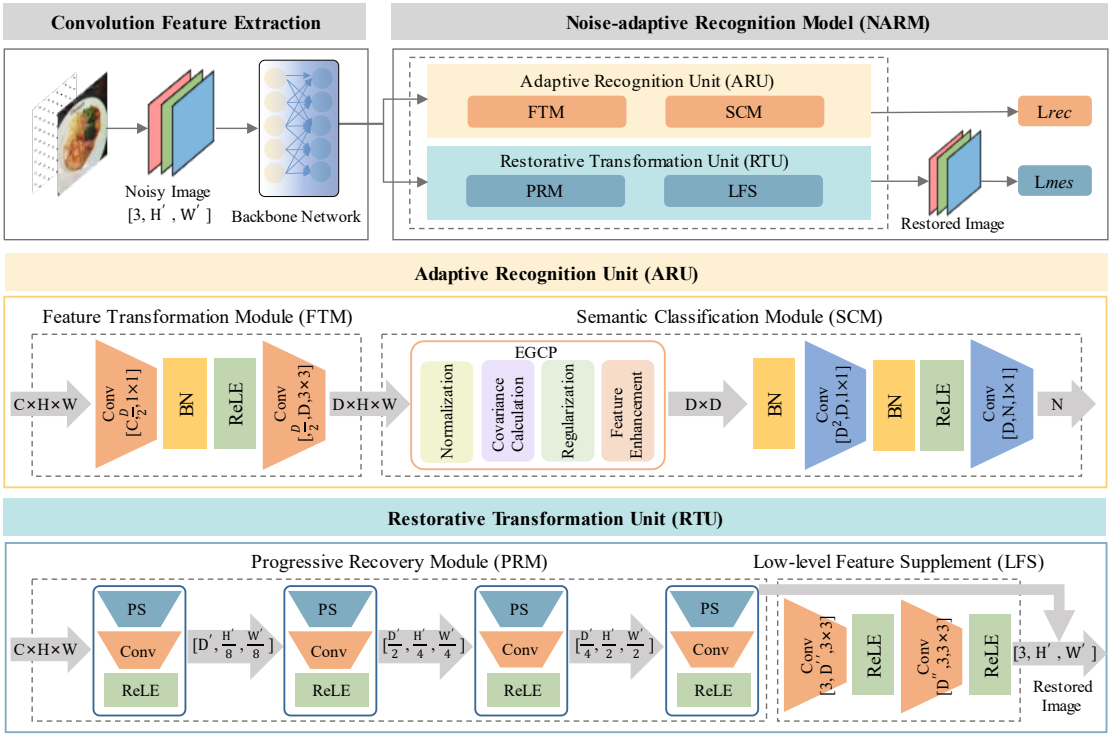

Figure 2. The architecture of the NARM is shown, which includes the Adaptive Recognition Unit (ARU) and the Recovery Transformation Unit (RTU). The Adaptive Recognition Unit (ARU) consists of the Feature Transformation Module (FTM) and the Semantic Classification Module (SCM). The Recovery Transformation Unit (RTU) includes the Progressive Recovery Module (PRM) and the Low-level Feature Supplement (LFS) module. D, D', and D'' serve as hyperparameters that regulate the number of channels within both the convolutional and fully connected layers. $H' \times W'$ is the image size input into the backbone network.

Adaptive Recognition Unit (ARU): ARU consists of the Feature Transformation Module (FTM) and the Semantic Classification Module (SCM). For the feature vector $x \in \mathbb{R}^{C \times H \times W}$ from the backbone network, it is first input into the FTM module for preliminary feature extraction. Then the extracted results are passed to the SCM module for covariance feature enhancement and final classification. The FTM consists of multiple convolutional layers, followed by Batch Normalization and ReLU activation functions. Specifically, the formula is given as:

$$d_1 = \mathrm{Conv}_{1\times1}(x), \quad d_1 \in \mathbb{R}^{\frac{D}{2}\times H\times W}, \tag{1}$$

$$d_1^{'} = \mathrm{ReLU}(\mathrm{BN}(d_1)), \quad d_1^{'} \in \mathbb{R}^{\frac{D}{2}\times H\times W}, \tag{2}$$

$$d_2 = \mathrm{Conv}_{3\times3}(d_1^{'}), \quad d_2 \in \mathbb{R}^{D\times H\times W}. \tag{3}$$

In Equation 1, the size of the convolutional kernel filter is denoted as $[C, \frac{D}{2}, 1\times1]$. In Equation 3, the size of the convolutional kernel filter is denoted as $[\frac{D}{2}, D, 3\times3]$. The resulting preliminary feature extraction output is denoted as $d_2 \in \mathbb{R}^{D\times H\times W}$.

In the SCM, for $d_2 \in \mathbb{R}^{D\times H\times W}$, the feature value enhancement is first performed using EGCP, and then the final classification task is completed through a series of convolutions, batch normalization, and ReLU operations. The EGCP process consists of four steps: feature normalization, weighted covariance matrix computation, covariance matrix regularization, and feature

 doi:10.20944/preprints202507.1310.v1

enhancement mechanism. Specifically, the introduction of feature normalization helps balance the feature distribution, thereby effectively reducing the impact of noise on covariance matrix computation:

$$d_{2,d}^{\text{norm}}(h,w)=\frac{d_2(d,h,w)-\mu_d}{\sqrt{\sigma_d^2+\epsilon}}, \tag{4}$$

where μ_d represents the mean of the d-th channel, σ_d represents the standard deviation of the d-th channel, and ϵ is a positive number to prevent division by zero. Through the above process, the normalized feature vector $d_{2,d}^{\text{norm}}$ is obtained. Subsequently, we compute the weighted covariance matrix to adjust the importance of feature channels dynamically:

$$w_d=\frac{1}{\sigma_d^2+\epsilon}, \tag{5}$$

$$P=\sum_{d=1}^{D} w_d\left(d_{2,d}^{\text{norm}}-\mu\right)\left(d_{2,d}^{\text{norm}}-\mu\right)^T, \tag{6}$$

where w_d represents the weight, and the larger the variance of a channel, the smaller its corresponding weight, effectively reducing the noise impact that high-variance channels may introduce. The resulting $P\in\mathbb{R}^{D\times D}$ expresses the second-order correlation between feature channels. To improve the numerical stability of the covariance matrix and avoid the occurrence of ill-conditioned matrices, we regularize P, and perform eigenvalue decomposition on the regularized covariance matrix:

$$P_{\text{reg}}=P+\epsilon I, \tag{7}$$

$$P_{\text{reg}}=U\Lambda U^T, \tag{8}$$

where, I is the identity matrix, and U is the eigenvector matrix, which results in the feature matrix $\Lambda=\text{diag}(\lambda_1,\lambda_2,\ldots,\lambda_D)$. To avoid numerical instability, we perform a lower bound correction on the eigenvalues to ensure the stability of the computation:

$$\lambda_i^{\text{reg}}=\max(\lambda_i,\epsilon). \tag{9}$$

The core mechanism of the EGCP module is to enhance the small eigenvalues in the covariance matrix, thereby improving feature discriminability in classification tasks. Specifically, the EGCP module first uses a logarithmic function for nonlinear transformation of the eigenvalues, adjusting the eigenvalues of the regularized covariance matrix:

$$\lambda_i^{\text{en}}=\log(1+\lambda_i). \tag{10}$$

Then, the enhanced eigenvalues are used to construct the improved covariance matrix:

$$P_{\text{en}}=U\Lambda_{\text{en}}U^T. \tag{11}$$

This results in the enhanced diagonal eigenvalue matrix Λ_{en}. However, in practical applications of food images, due to the significant differences in feature distributions across different foods, relying solely on the enhancement mechanism may lead to uneven feature distributions, which in turn affects classification performance. To alleviate this issue, we design a dynamic scaling factor mechanism for the EGCP module, which adaptively adjusts the enhancement strength to make the

 doi:10.20944/preprints202507.1310.v1

enhanced feature distribution more uniform and stable, thereby improving classification performance. First, construct the dynamic scaling matrix:

$$S = \exp(-P_{\text{en}}) = U \exp(-\Lambda_{\text{en}})U^{T}. \tag{12}$$

The differences between eigenvalues are effectively balanced by taking the inverse exponential matrix of the enhanced covariance matrix. Then, the cross-covariance matrix is computed, and the dynamic scaling factor is defined based on the Frobenius norm of the matrix:

$$Q_{\text{cross}} = P_{\text{en}}^{1/2} \cdot S, \tag{13}$$

$$\text{SF} = \| Q_{\text{cross}} \|_{F} = \sqrt{\sum_{i=1}^{D} \left(\lambda_i^{1/2} e^{-\lambda_i} \right)^2}, \tag{14}$$

where D is the feature dimension, and λ_i is the enhanced eigenvalue. By introducing the dynamic scaling factor, the EGCP module can further adjust the feature distribution to accommodate the complex feature distributions of different food types. Ultimately, the EGCP module and the dynamic scaling factor further optimize the enhanced covariance matrix, producing the final feature representation. Specifically, the final feature matrix is represented as:

$$A = (\text{SF} + 1) \cdot P_{\text{en}}^{1/2}, \tag{15}$$

and the dimension of the feature matrix A is $D \times D$,which can be flattened into $\mathbb{R}^{D^2}$. In the final stage of the ARU, we complete the final classification task through a series of convolution operations, Batch Normalization, and ReLU activation functions, as expressed by the formulas:

$$f_1^{'} = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}(A))), \quad f_1^{'} \in \mathbb{R}^{D\times H\times W}, \tag{16}$$

$$f_2 = \text{Conv}_{1\times1}(d_1^{'}), \quad f_2 \in \mathbb{R}^{D\times H\times W}, \tag{17}$$

where in Equation 16, the convolution kernel filter size is $[D^2, D, 1\times1]$. In contrast, in Equation 17, the convolution kernel filter size is $[D, N, 3\times3]$. Here, $p \in \mathbb{R}^{N}$ denotes the predicted score, where N denotes the overall number of food classifications.

Restorative Transformation Unit (RTU): RTU consists of the Progressive Recovery Module (PRM) and the Low-level Feature Supplement (LFS) module. The former is responsible for extracting image information from the intermediate features of the backbone network. At the same time, the latter is used to supplement detailed information from noisy inputs, achieving comprehensive recovery from noisy images to clear images. PRM mainly recovers the clean image through progressive upsampling operations from the feature vector output by the backbone network. PRM is composed of multiple progressively connected upsampling modules, and the definition of each module is given by:

$$M_k^{\text{up}}(x) = \text{ReLU}(\text{Conv}(\text{PixelShuffle}(x))). \tag{18}$$

PRM is composed of four such modules, and the specific output is given by:

$$I_{\text{PRM}} = \text{PRM}(x) = M_4^{\text{up}}\left(M_3^{\text{up}}\left(M_2^{\text{up}}\left(M_1^{\text{up}}(x)\right)\right)\right). \tag{19}$$

Through the progressive recovery process, PRM ultimately changes the resolution from $H \times W$ to the target size $H' \times W'$, and the output of the recovered image is $I_{PRM} \in \mathbb{R}^{3\times H'\times W'}$.

Since intermediate features of the backbone network may lose low-level detail information during the extraction process, LFS directly extracts low-level features from the input noisy image to supplement this crucial information. LFS is implemented through two shallow convolution operations, and the specific formula is as follows:

 doi:10.20944/preprints202507.1310.v1

$$I_{\text{LFS}} = \text{LFS}(I_{\text{noisy}}) = \text{Conv}_2(\text{Conv}_1(I_{\text{noisy}})), \tag{20}$$

where $I_{\text{noisy}} \in \mathbb{R}^{3\times H'\times W'}$ is the input noisy image. The outputs of PRM and LFS were fused by pixel-wise addition to generate the final recovered image:

$$I_{\text{restored}} = I_{\text{PRM}} + I_{\text{LFS}}, \tag{21}$$

where I_{restored} is the final clean image generated, combining the global details recovered by PRM with the low-level features supplemented by LFS.

The initial component involves the Softmax loss calculated between the classification output $p \in \mathbb{R}^N$ and the target label. The second part is the pixel-wise Mean Square Error (MSE) between the recovered image I_{restored} and the clean image I_{clear} without noise injection. The calculation method is as follows:

$$L_{\text{rec}} = \text{Softmax}(p, g), \tag{22}$$

$$L_{mse} = \text{MSE}(I_{\text{restored}}, I_{\text{clean}}), \tag{23}$$

$$L_{NARM} = \alpha L_{rec} + \beta L_{mse}, \tag{24}$$

where α and β are balancing parameters, producing the final loss function L_{NARM}.

3.2. Weighted Multi-Granularity Fusion

To significantly improve the representational capacity of features across various stages of the backbone network, we propose the Weighted Multi-Granularity Fusion (WMF), which improves recognition task performance by integrating features from different depth layers.

In the backbone network, layers at varying depths capture information at different levels of abstraction. Early layers in the network tend to capture fine-grained and specific features, whereas deeper layers emphasize more abstract and semantically rich information. For recognition tasks, the training process is typically dominated by deep-layer features. In comparison, for low-level vision tasks like image restoration, shallow-layer features generally have a more substantial impact unless an advanced mechanism is used to balance deep and shallow features.

This paper emphasizes recognition as the primary objective while incorporating image restoration as a secondary supportive task. This approach ensures that image restoration does not overly focus on superficial features. To address this, we adopt a stepwise training approach, first training the shallow layers, then gradually progressing to deeper layers, and ultimately completing the training of the deepest layer. During the training process, by minimizing each loss gradually rather than optimizing them simultaneously, the model can more effectively learn the features of each layer.

In the final stage, we apply the WMF to combine the network layer features from each stage with weighting. This strategy helps the model improve performance in recognition tasks by synthesizing multi-level information, while moderately considering the auxiliary role of shallow features in image restoration, thereby achieving a balance across stages and improving overall performance.

Specifically, as shown in Algorithm 1, we insert S NARM modules into the backbone network, dividing the entire training process into S+1 steps. In the first S steps, these S NARM modules receive feature vectors from the backbone network and process them to optimize the loss functions L_{rec} and L_{mse}. This phase concentrates on extracting and optimizing features via each NARM module, allowing the model to efficiently capture and represent essential information and characteristics from the input data.

 doi:10.20944/preprints202507.1310.v1

In step S+1, we use the weighted fusion method to integrate the feature maps from all previous stages. First, the feature maps from all stages are weighted and fused to generate a multi-granularity fused feature map:

$$x^* = \sum_{i=1}^{s} \beta_i x_i, \tag{25}$$

where β_i is a hyperparameter used to adjust the contribution of each stage's feature map to the final fused feature map. By appropriately setting this hyperparameter, we can flexibly control the significance of feature maps from different stages. Finally, for the resulting fused feature map x^*, we use the ARU in EGCP to compute the loss function.

Algorithm 1 Weighted Multi-Granularity Fusion

Require: Given a dataset $\mathcal{D}=\{(\text{input}^i,\text{target}^i)\}_{i=1}^{I}$ (where I represents the total number of batches in $\mathcal{D}$)

for epoch = 1 to num_of_epochs **do**
  for (input, target) in $\mathcal{D}$ **do**
    for n = 1 to S **do**
      $x_i = \{x_1, x_2, \ldots, x_s\}$
      # NARM
      $L_{\text{NARM}}^{(i)} = \alpha L_{\text{rec}}^{(i)} + \beta L_{\text{mse}}^{(i)}$
      BACKWARD($L_{\text{NARM}}^{(i)}$)
    end for
    # WMF
    $x^* = \sum_{i=1}^{s} \beta_i x_i$
    $L_{\text{cls}} = \text{ARU}(x^*)$
    BACKWARD(L_{cls})
  end for
end for

3.3. Progressive Temperature-Aware Feature Distillation

As discussed earlier, multiple NARM modules are introduced in WMF to enhance the model's feature extraction capability, and a multi-step training strategy is adopted. However, this method inevitably adds extra computational costs. At the same time, with the introduction of new modules, the computational complexity in the backpropagation process also increases significantly, which impacts the overall training efficiency. In practical applications, it is crucial to recognize that the efficiency during the inference stage frequently holds greater importance compared to the training stage. This necessitates a careful balance between the complexity involved in training and the efficiency achieved during inference.

To tackle this challenge, we introduce a Progressive Temperature-Aware Feature Distillation (PTAFD) approach that integrates temperature regulation with a step-by-step learning framework for knowledge distillation. The core concept of PTAFD is to introduce a temperature control mechanism into the knowledge distillation process, and use a progressive learning strategy to gradually optimize the student model's absorption of knowledge from the teacher model, thus improving the student model's classification performance.

We adjust the degree of attention the student model pays to labels in a staged manner, gradually improving its learning efficiency. As illustrated in Figure 3, a network comprising multiple NARM modules and WAF serves as the teacher model, while a conventional CNN backbone network acts as

 doi:10.20944/preprints202507.1310.v1

the student model. It is crucial to observe that both the teacher and student models employ the same backbone network structure, ensuring they share an equal number of stages.

Specifically, during training, the teacher model generates soft labels using the Softmax function:

$$\hat{y}_{\text{teacher}} = \text{softmax}\left(\frac{x_{\text{teacher}}}{T}\right), \tag{26}$$

where x_{teacher} is the final feature used by the teacher model for classification, and T is the temperature parameter. When $T<1$, the output distribution becomes sharper; when $T>1$, the output distribution becomes smoother. During training, the student model is divided into two branches: the soft prediction branch and the complex prediction branch, which are used to calculate the distillation loss and the classification loss of the student network:

$$L_{\text{distillation}} = -\sum_{i} \hat{y}_{\text{teacher},i} \log(\hat{y}_{\text{student},i}), \tag{27}$$

$$L_{\text{student}} = -\sum_{j} y_j \log(\hat{y}_{\text{student},j}), \tag{28}$$

where $\hat{y}_{\text{student}}$ represents the soft labels of the student model, y represents the actual labels, and L_{student} is the output of the student model when $T=1$. The variable i is used to identify the current category being evaluated, while j is used to determine the index of the true class in the complex labels. By combining the two losses above, weighted by the coefficient α, the total loss function is formed:

$$L = \alpha \cdot L_{\text{distillation}} + (1-\alpha) \cdot L_{\text{student}}. \tag{29}$$

The choice of temperature and the implementation of a progressive strategy are essential throughout various phases of the training process. In the initial stage, a lower temperature T ($T<1$) is maintained to focus on the high-probability labels of the teacher model, thereby reducing the interference from negative labels. In later stages, the temperature is gradually increased (from 1 to 2), which enhances the student model's attention to negative labels and facilitates more comprehensive learning and knowledge absorption. By dynamically adjusting the temperature and weighting coefficient, PTAFD can improve the student model's performance in complex classification tasks and effectively utilize information from negative labels.

In the PTAFD approach, the training procedure is split into two phases. During the initial phase, referred to as the distillation phase, the primary objective is to progressively refine the feature representation of the student network by leveraging feature distillation techniques. In this stage, we extract features from each corresponding stage of the teacher and student networks. We use a distillation loss function to align their feature distributions for efficient knowledge transfer from the teacher network. Specifically, we introduce a temperature-aware mechanism, making the distillation process progressive: in the early stages, high temperature is used to align coarse-grained feature representations, helping the student network capture global features; in the later stages, the temperature is reduced to focus on fine-grained feature alignment, encouraging the student network to more precisely learn high-quality feature representations. Once the distillation phase is completed, the student model no longer relies on the teacher model's guidance. At this stage, the student model is trained utilizing a loss function that is determined by the classification results.

 doi:10.20944/preprints202507.1310.v1

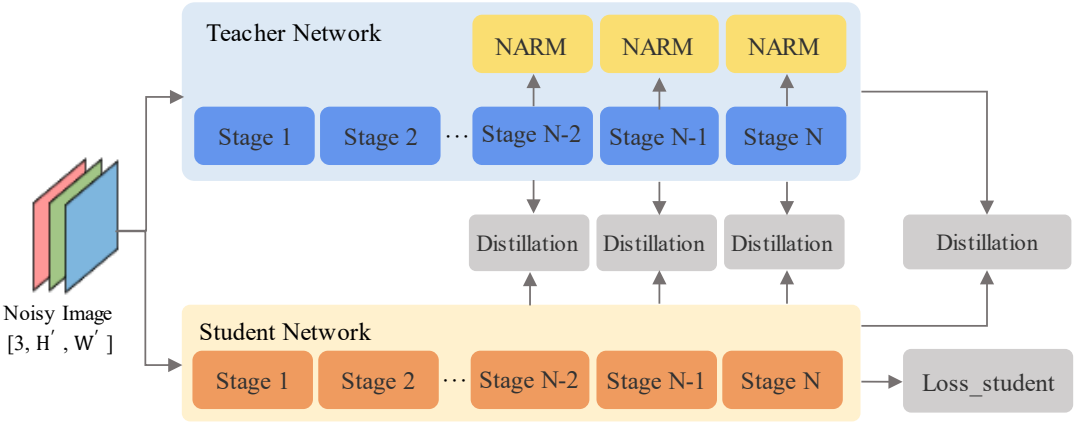

Figure 3. Schematic diagram of Progressive Temperature-Aware Feature Distillation (PTAFD). The teacher network contains multiple NARM modules, with each network layer corresponding to the respective layer in the student network.

4. Experiment

4.1. Dataset

To validate the effectiveness of the proposed method, we evaluated our model on two popular food datasets, ETH Food 101[20] and Vireo Food-172[21].

ETH Food-101 is a classic dataset containing 101 categories of Western foods with 101,000 images. Each category includes 1,000 images, divided into 750 training and 250 test images at a 3:1 ratio. The images in this dataset come from different angles, lighting conditions, and backgrounds, offering high diversity and challenges. ETH Food-101 is unprocessed; researchers can crop, scale, and normalize the images. As a widely used benchmark for image classification, feature extraction, and transfer learning, it provides a reliable evaluation basis for fine-grained food image classification tasks.

Vireo Food-172 is a dataset focused on Chinese cuisine, comprising 172 classes of Asian dishes with a total of 110,241 images. The dataset is split into training, validation, and test sets at a ratio of 60%, 10%, and 30%, respectively. The design of Vireo Food-172 aims to advance the development of food image classification tasks in computer vision. The number of images per category in this dataset is imbalanced, reflecting the data distribution in real-world applications. It offers a wealth of varied visual data, featuring multiple camera angles and backgrounds. This makes it highly suitable for both training and evaluating deep learning models designed for food recognition tasks.

4.2. Experimental Settings

In the experiment, ResNet50[20] was used as the backbone network, which consists of five stages. In its last three stages, we added three NARM modules, so in Algorithm 1, the parameter S is set to 3. To highlight the importance of deep layers, the feature fusion weights β_i of these three network layers are set to 0.2, 0.35, and 0.45, respectively. In the PTAFD framework, we also use these three intermediate feature maps to transfer knowledge, with the weighting coefficient α for the student network's classification loss set to 0.5.

This study assesses the model's classification accuracy and its resilience to noise using the ETH Food 101 and Vireo Food-172 datasets. For data preparation, images are first resized to a dimension of 550×550, followed by center cropping to 448×448, and then normalized. During the training process, the model undergoes 200 epochs of training, utilizing a batch size set to 32. The chosen optimization algorithm is stochastic gradient descent (SGD), initialized with a learning rate of 0.002 and employing a cosine annealing schedule for the learning rate adjustment. The hyperparameters

 doi:10.20944/preprints202507.1310.v1

D, D', and D'' are used to control the number of channels in the convolutional and fully connected layers, set to 1024, 256, and 64, respectively.

The evaluation metrics include Top-1 classification accuracy (Top-1 Acc.) and Top-5 classification accuracy (Top-5 Acc.). All neural network models are built utilizing the PyTorch framework.

4.3. Comparison with Baselines

To verify the effectiveness of our method in food category recognition tasks, we compare our model with state-of-the-art food recognition methods on ETH Food-101 and Vireo Food-172. We also compare the backbone networks of different techniques for a fair comparison. Table 1 presents the performance outcomes on the ETH Food-101 and Vireo Food-172 datasets. For the ETH Food-101 dataset, our approach achieved remarkable results in both Top-1 and Top-5 classification accuracy under identical experimental settings. It is evident that our model surpasses the ResNet50 backbone network by 5.15% in Top-1 accuracy and by 1.30% in Top-5 accuracy. Additionally, compared to PRENet, which employs the same backbone architecture, our method demonstrates improvements of 2.66% in Top-1 accuracy and 0.66% in Top-5 accuracy.

On the ETH Food-101 dataset, our method performs best in Top-1 accuracy, while in Top-5 accuracy, it is second only to SICL with CBiAFormer-B as the backbone, primarily due to the advantages of CBiAFormer-B's network architecture. The network uses the more advanced Swin-B architecture, providing more substantial feature extraction and modeling capabilities. Additionally, CBiAFormer-B integrates a dual-branch adaptive attention mechanism and a category-component cross-task interaction module (SICL), which can fully model the complex interaction between components and categories, thus effectively capturing diverse features and subtle semantic relationships in images, achieving excellent performance in Top-5 recognition tasks. However, CBiAFormer-B does not adopt lightweight techniques, so its model complexity is relatively high, which may make it less convenient for deployment in resource-constrained environments. In contrast, our method introduces distillation techniques, effectively reducing model complexity while maintaining good performance, demonstrating the advantages of our approach in flexibility and adaptability.

We can also observe that although some fine-grained methods perform better than the baseline network, their results on food datasets are not as strong as on standard fine-grained datasets. For example, DCL performs worse on ETH Food-101 compared to other fine-grained datasets, possibly because it fails to fully consider texture information from shallow networks and the differences in feature distributions within the same category. Furthermore, certain fine-grained approaches, like PMG, exhibit inferior performance compared to the baseline network. This is attributed to PMG's inadequate emphasis on local feature interactions and its limited capability to facilitate effective multi-level feature learning. This causes the model to focus on standard semantic features. However, the non-rigid structure of many food categories and the lack of fixed semantic information make it difficult for these methods to perform well. The experimental findings indicate that simply utilizing existing fine-grained approaches does not ensure optimal performance in food recognition. This highlights the adaptability and effectiveness of our proposed model for fine-grained food recognition.

On the Vireo Food-172 dataset, our approach achieves higher Top-1 and Top-5 classification accuracy compared to the majority of food recognition methods, placing second only to IVRDRM. Despite IVRDRM utilizing the stronger ResNet-101 backbone and achieving outstanding performance with Top-1 accuracy of 93.33% and Top-5 accuracy of 99.15% through multi-task learning and component-region discovery combined with graph relationship modeling, our method also performs well with the ResNet-50 backbone, reaching 92.37% Top-1 accuracy and 98.55% Top-5 accuracy. Despite using a less complex backbone, this indicates that our model significantly improves classification robustness and accuracy in complex scenarios by combining anti-noise learning modules and multi-granularity feature fusion. The higher accuracy of IVRDRM is mainly due to its

 doi:10.20944/preprints202507.1310.v1

strong backbone network and deep modeling of component relationships, whereas our model emphasizes improving stability and robustness through anti-noise learning and feature enhancement modules. These results demonstrate that our method effectively controls model complexity while improving classification performance.

Furthermore, the Vireo Food-172 dataset suffers from class imbalance. When there are fewer samples in specific categories, the model struggles to learn sufficient features, resulting in poor recognition performance for those categories. This data imbalance significantly affects the model's accuracy on minority classes, thereby reducing overall classification performance. Nonetheless, our method demonstrates strong robustness in addressing the data imbalance problem, and its overall recognition performance surpasses most existing methods.

Table 1. and VIREO Food-172 Datasets (%).

Method	Backbone	ETH Food-101		Vireo Food-172	
		Top-1 acc	Top-5 acc	Top-1 acc	Top-5 acc
ResNet152+SVM-RBF [39]	ResNet152	64.98	-	-	-
FS_UAMS[40]	Inceptionv3	-	-	89.26	-
ResNet50[20]	ResNet50	87.42	97.40	-	-
DenseNet161[41]	DenseNet161	-	-	86.98	97.31
SENet-154[42]	ResNeXt-50	88.68	97.62	88.78	97.76
PAR-Net[43]	ResNet101	89.30	-	89.60	-
DCL[44]	ResNet50	88.90	97.82	-	-
PMG[45]	ResNet50	86.93	97.21	-	-
WS-DAN[46]	Inceptionv3	88.90	98.11	-	-
NTS-NET[47]	ResNet50	89.40	97.80	-	-
PRENet[48]	ResNet50	89.91	98.04	-	-
PRENet[48]	SENet154	90.74	98.48	-	-
SGLANet[49]	SENet154	89.69	98.01	90.30	98.03
Swin-B[50]	Transformer	89.78	97.98	89.15	98.02
DAT[51]	Transformer	90.04	98.12	89.25	98.12
EHFR-Net[16]	Transformer	90.70	-	90.30	-
IVRDRM[52]	ResNet-101	92.36	98.68	**93.33**	**99.15**
SICL(CBiAFormer-T)[53]	Swin-T	91.11	98.63	90.70	98.05
SICL(CBiAFormer-B)[53]	Swin-B	92.40	**98.87**	91.58	98.75
Our method	ResNet50	**92.57**	98.70	92.37	98.55

4.4. Ablation Study

In this section, we conduct a thorough evaluation of the model's various modules and design choices using a set of ablation studies. The goal is to examine how different components influence the overall performance. Specifically, we first explore the role of core modules (such as NARM and WMF) in improving classification accuracy and noise robustness. Next, we analyze the impact of key factors, such as Gaussian noise intensity, network layer design, and loss function weights, on model performance. In addition, we further validate the model's effectiveness and robustness for specific tasks through distillation strategies and noise invariance experiments. These experiments reveal the logic of model optimization and the key factors driving performance improvement, offering a theoretical foundation for future enhancements.

NARM and WMF Contribution Analysis Experiment: In this module contribution analysis experiment, we aim to investigate NARM and WMF's contribution to the model's overall performance. To validate the effectiveness of the EGCP module in NARM, we designed an experiment where the model uses the NARM architecture but excludes the EGCP module. We also compared the basic ResNet50 model, which only uses the last three feature layers, with other settings consistent with the final model design, called "Simple Feature Fusion" (SFF). The experimental

 doi:10.20944/preprints202507.1310.v1

results are shown in Table 2, and we can observe that following: (1) Compared to SFF, our final model shows a significant improvement in Top-1 classification accuracy on both datasets; (2) After introducing the NARM module on top of SFF, the model with the EGCP module achieved Top-1 classification accuracy improvements of 1.88% and 2.16% on the two datasets, compared to the network without the EGCP module. This indicates that EGCP, by providing global feature covariance, effectively enhances the model's feature extraction and noise robustness, thereby improving classification performance; (3) With the introduction of the EGCP module, the classification performance was further improved compared to the model that uses the WMF method with multi-step training. This shows that WMF fully utilizes global information and local details by weighted fusion of features from multiple layers, further enhancing the model's classification ability and improving its adaptability to complex samples.

Table 2. Ablation Study on the Contributions of NARM and WMF (%).

	ETH Food-101				Vireo Food-172			
	p_1	p_2	p_3	Top-1	p_1	p_2	p_3	Top-1
SFF	86.43	87.23	86.79	87.86	82.87	86.12	85.72	86.63
SFF+NARM(no EGCP)	87.73	88.36	88.97	90.31	86.58	87.65	88.67	89.72
SFF+NARM	89.25	90.80	91.23	92.19	88.23	89.82	91.10	91.88
SFF+NARM+WMF	89.77	91.27	92.03	**92.57**	88.69	90.03	91.59	**92.37**

Experiment on the Effect of Different Gaussian Noise Intensities: Gaussian noise is injected into the original images as input in the proposed method. In this experiment, Gaussian noise with different standard deviations (σ) is injected into the NARM, with specific values of [0.01, 0.05, 0.10, 0.15, 0.20]. By comparing the model's classification accuracy under different noise levels, we analyze the effect of noise intensity on noise-robust learning and feature enhancement. The experimental results are shown in Figure 4. As the noise intensity increases, the model's classification accuracy first rises and then declines. When the noise intensity is σ=0.10, the model exhibits the best noise robustness and classification performance, with a significant increase in accuracy. However, when the noise intensity further increases ($\sigma > 0.15$), the interference from noise on the model's feature representation becomes dominant, leading to a significant drop in classification performance. Considering both noise-robust performance and classification accuracy, we ultimately choose σ=0.10 as the standard deviation for injecting Gaussian noise. We will use this parameter in subsequent experiments for further model optimization and validation.

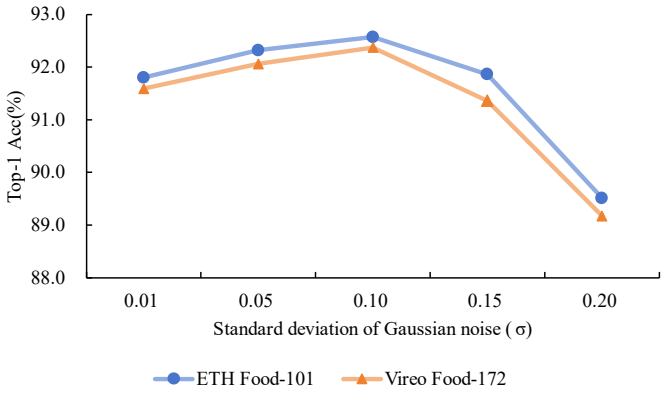

Figure 4. Impact of Gaussian Noise Intensity on Classification Accuracy for Different Datasets (Highest accuracy achieved at σ=0.10).

RTU Noise Invariance Validation Experiment: To verify the effectiveness of the proposed Residual Transformation Unit (RTU) in achieving noise invariance, we initially evaluate the

 doi:10.20944/preprints202507.1310.v1

performance of the three integrated NARM modules for image denoising tasks within the backbone network. As shown in Figure 5, processing through the three NARM modules generates a relatively smooth denoised image for a noisy image. This result indicates that our model can effectively capture noise invariance. Furthermore, to further validate RTU's advantages, we compared the NARM model integrated with RTU (NARM-ARU and RTU) with the control model that contains only ARU (NARM-ARU). For the ETH Food-101 dataset, we injected different noise levels into the test data and recorded the classification accuracy of the two models under different noise intensities. The results are shown in Figure 6. As the noise intensity increases, the model with RTU experiences a slower decline in accuracy, while the model without RTU shows a sharp decrease in accuracy. These experimental results effectively demonstrate the significant advantages of RTU in noise handling.

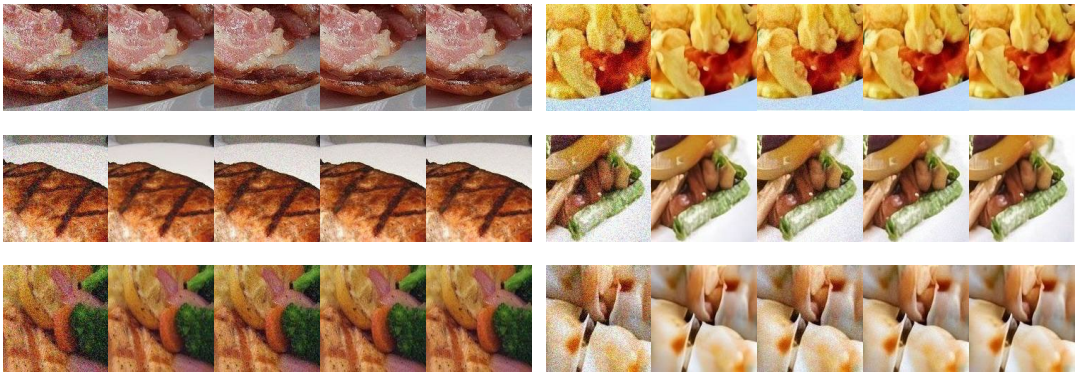

Figure 5. Visual analysis of noise reduction effect. The three sets of images on the left are from the ETH Food-101 dataset, and those on the right are from the Vireo Food-172 dataset. In each collection of five images, the progression from left to right is shown: starting with the image that has added Gaussian noise, followed by the original image, and concluding with the results obtained after three rounds of NARM denoising. All images are uniformly cropped to 150×150 size for display.

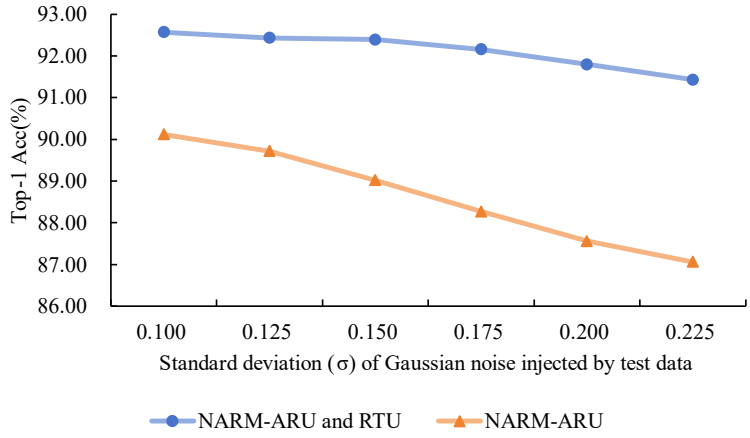

Figure 6. The accuracy performance of the NARM model with RTU and the NARM model with ARU is compared when different intensity noise is injected into the test data.

Distillation Strategy Comparison Experiment: The PTAFD method is designed based on the learning process of Algorithm 1. This experiment demonstrates its superiority as a teacher network compared to other advanced knowledge distillation (KD) methods [54–60]. We adopted the network outlined in Algorithm 1 as the teacher model and used ResNet50 as the student model to assess the impact of different knowledge distillation (KD) techniques on the accuracy of the student model throughout the knowledge transfer procedure. KD methods can generally be divided into two categories: the first focuses on aligning point-to-point features of the student model with the teacher model, typically by guiding the student to learn intermediate features from the teacher or selecting

 doi:10.20944/preprints202507.1310.v1

important features for distillation to improve training performance [57–60], and PTAFD belongs to this category; the second approach aims to align the feature distribution or correlations of the student model with those of the teacher model. This is usually achieved by minimizing the discrepancies between the student and teacher models using feature relationship metrics, such as distance or angular loss [54–56]. Figure 7 compares the distillation performance between various distillation methods and PTAFD on the ETH Food-101 dataset. We trained using different methods five times and presented the results in the form of error bar plots, where the bar represents the average of each method, and the top and bottom of the error lines represent the maximum and minimum values, respectively. On the ETH Food-101 dataset, the student model trained with PTAFD achieved an average accuracy 0.63%-1.74% higher than the other methods, indicating that when PTAFD is used as the teacher model, it can transfer knowledge more effectively than other KD methods.

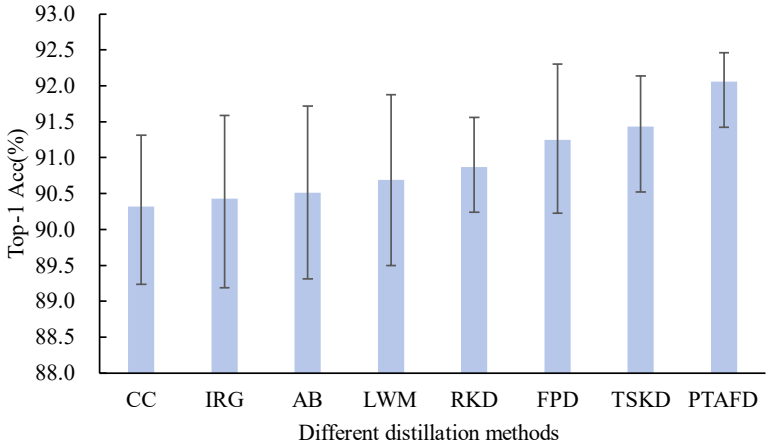

Figure 7. Comparison of distillation performance of different distillation methods and PTAFD on dataset ETH Food-101. The bar chart represents the average value of each method, and the top of the upper error line and the bottom of the lower error line represent each method's maximum and minimum values, respectively. The comparison includes feature alignment-based methods such as CC[55], IRG[56], AB[57], LWM[58], and FPD[59], and feature distribution alignment methods like RKD[54], TSKD[60], alongside PTAFD.

Network Layer Design Optimization Experiment: To evaluate the model's effectiveness, we compared the impact of inserting NARM modules at different stages on performance. The experiment used ResNet50 as the base network, with NARM modules inserted at the last one, two, three, four, and all five stages. The experimental outcomes, as illustrated in Figure 8, demonstrate that the model incorporating NARM modules in the final three stages attained superior performance on both the ETH Food-101 and Vireo Food-172 datasets. In comparison, the model with NARM modules inserted only at the last stage exhibited lower accuracy due to the absence of multi-level feature support. Although inserting NARM modules in the previous two stages showed some improvement, it failed to utilize features from additional layers fully. When inserted up to the last four stages, the accuracy improved to 92.32% and 91.85% on the Food-101 and Vireo Food-172 datasets, respectively, but the gain leveled off. Lastly, when NARM modules were inserted at all five stages, classification performance was slightly improved, but the training time and computational resource consumption increased. In summary, inserting NARM modules in the last three stages provides the best balance between performance and computational complexity, effectively utilizing multi-level features.

 doi:10.20944/preprints202507.1310.v1

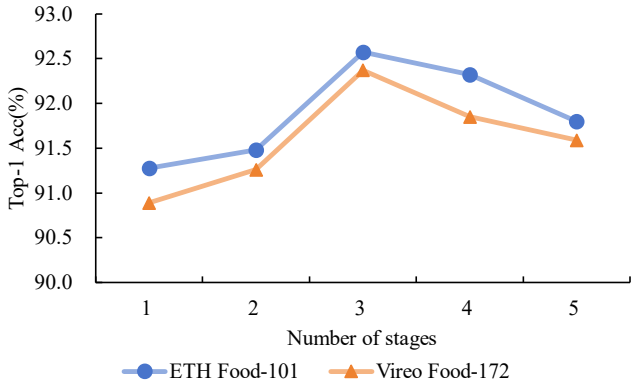

Figure 8. Comparison of the effects of different stages in backbone network selection reveals that features from the final three stages yield the best results.

Loss Function Parameter Sensitivity Experiment: The model proposed in this paper uses a weighted loss function composed of classification loss L_{rec} and recovery loss L_{mse} in the backward propagation of the last three layers to optimize classification performance and denoising recovery ability. The experiment tested the impact of different weight settings on model performance. The results, shown in Figure 9, indicate that when (α, β) = (0.6, 0.4), the model performed best on both datasets, validating the necessity of appropriate weight allocation between classification and recovery performance. Specifically, the weight settings reflect the dominant role of the recognition task, with classification being the core function and requiring a higher weight ($\alpha > \beta$), while denoising, as a secondary task, holds relatively less importance. Further analysis revealed that when β=0 (only optimizing classification loss), the model neglects noise recovery, resulting in uncorrected biases that affect classification performance. When α=0 (only optimizing recovery loss), while low-level visual features are well restored, the lack of high-level semantic enhancement limits the classifier's effectiveness. This indicates that classification loss and recovery loss optimize the model from high-level semantics and low-level recovery perspectives, complementing each other. Optimizing a single loss leads to biased feature representation, thus affecting overall performance. Through weighted combination, the model achieves a balance between classification performance and noise robustness, significantly improving performance in complex noise scenarios.

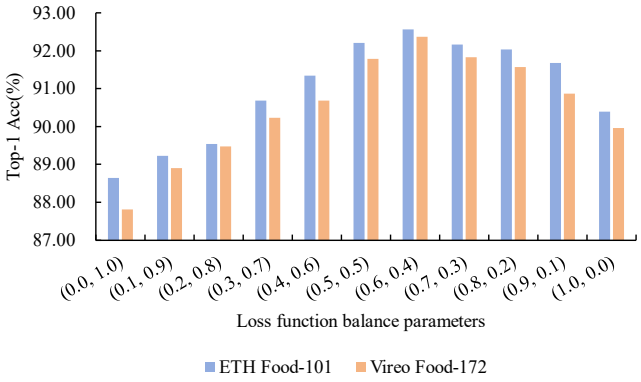

Figure 9. The analysis of how different balancing parameters affect experimental outcomes reveals that relying solely on either loss function L_{rec} or L_{mse} results in suboptimal performance. The highest recognition accuracy is attained when the balancing parameters are configured to (0.6, 0.4).

 doi:10.20944/preprints202507.1310.v1

5. Conclusions and Discussion

This paper proposes a food image recognition method that combines noise-robust learning with covariance feature enhancement. It addresses the noise interference problem in food image recognition by designing the Noise Adaptive Recognition Module (NARM) and introducing Eigenvalue-Enhanced Global Covariance Pooling (EGCP). NARM utilizes a multi-task learning framework to simultaneously optimize feature extraction and noise reduction, while EGCP boosts smaller eigenvalues within the covariance matrix to enhance both performance and robustness of the model in fine-grained classification tasks. In addition, this paper proposes the Weighted Multi-Granularity Fusion (WMF) method to integrate multi-level features and enhance the model's classification ability. It balances the model's lightweight design and efficient performance with Progressive Temperature-Aware Feature Distillation (PTAFD). Experimental results show that this method outperforms existing noise robustness and classification accuracy methods. Through the innovative design of noise-robust modules and feature enhancement strategies, this paper provides a more robust technical solution for food image recognition, suitable for food classification tasks in complex scenarios. It offers novel insights for food computing and fine-grained image recognition research.

Author Contributions: Conceptualization, Z.C. and H.C.; methodology, J.W. and Y.W.; software, Z.C.; validation, H.C., J.W., and Y.W.; formal analysis, Z.C.; investigation, H.C.; resources, J.W. and Y.W.; data curation, H.C.; writing—original draft preparation, Z.C. and H.C.; writing—review and editing, J.W. and Y.W.; visualization, H.C.; supervision, J.W. and Y.W.; project administration, J.W.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China, Grant Number: 32302241; National Key Research and Development Program of China, Grant Number: 2024YFF1106705.

Institutional Review Board Statement: There are only computational experiments in this work, and they were all established according to the ethical guideline of the Helsinki Declaration.

Informed Consent Statement: The data used in this work are from public source and the participants' personal information were desensitized. The manuscript is approved by all authors for publication.

Data Availability Statement: The experimental data used in this study are sourced from the following publicly available datasets: ETH Food-101 (accessible at: https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/) and Vireo Food-172 (accessible at: https://fvl.fudan.edu.cn/dataset/vireofood172/list.htm). We appreciate the provision of these resources, which have significantly supported the conduct of this research.

Acknowledgments: The authors would like to thank Chen Zhou and Dongbo Liu in Beijing Forestry University for their technical support for the computational experiments.

Conflicts of Interest: No conflict of interest exists in the submission of this manuscript.

References

1. Z. Zhao, R. Wang, M. Liu, L. Bai, Y. Sun, Application of machine vision in food computing: A review. Food Chemistry 463, 141238 (2025).
2. Y. Zhang et al., Deep learning in food category recognition. Information Fusion 98, 101859 (2023).
3. W. Min, S. Jiang, L. Liu, Y. Rui, R. Jain, A Survey on Food Computing. ACM Comput. Surv. 52, Article 92 (2019).
4. D. Allegra, S. Battiato, A. Ortis, S. Urso, R. Polosa, A review on food recognition technology for health applications. Health Psychol Res 8, 9297 (2020).
5. A. Rostami, N. Nagesh, A. Rahmani, R. Jain, paper presented at the Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management, Lisboa, Portugal, 2022.
6. A. Rostami, V. Pandey, N. Nag, V. Wang, R. Jain, paper presented at the Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 2020.

 doi:10.20944/preprints202507.1310.v1

7. A. Ishino, Y. Yamakata, H. Karasawa, K. Aizawa, RecipeLog: Recipe Authoring App for Accurate Food Recording. Proceedings of the 29th ACM International Conference on Multimedia, (2021).
8. W. Wang et al., A review on vision-based analysis for automatic dietary assessment. Trends in Food Science & Technology 122, 223-237 (2022).
9. M. F. Vasiloglou et al., Multimedia Data-Based Mobile Applications for Dietary Assessment. J Diabetes Sci Technol 17, 1056-1065 (2023).
10. Y. Yamakata, A. Ishino, A. Sunto, S. Amano, K. Aizawa, paper presented at the Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 2022.
11. K. Nakamoto, S. Amano, H. Karasawa, Y. Yamakata, K. Aizawa, paper presented at the Proceedings of the 1st International Workshop on Multimedia for Cooking, Eating, and related APPlications, Lisboa, Portugal, 2022.
12. Y. Zhu, X. Zhao, C. Zhao, J. Wang, H. Lu, Food det: Detecting foods in refrigerator with supervised transformer network. Neurocomputing 379, 162-171 (2020).
13. I. Mohammad, M. S. I. Mazumder, E. K. Saha, S. T. Razzaque, S. Chowdhury, paper presented at the Proceedings of the International Conference on Computing Advancements, Dhaka, Bangladesh, 2020.
14. E. Aguilar, B. Remeseiro, M. Bolaños, P. Radeva, Grab, Pay, and Eat: Semantic Food Detection for Smart Restaurants. IEEE Transactions on Multimedia 20, 3266-3275 (2018).
15. D. Peng et al., Defects recognition of pine nuts using hyperspectral imaging and deep learning approaches. Microchemical Journal 201, 110521 (2024).
16. G. Sheng et al., A Lightweight Hybrid Model with Location-Preserving ViT for Efficient Food Recognition. Nutrients. 2024 (10.3390/nu16020200).
17. H. Wang et al., Nutritional composition analysis in food images: an innovative Swin Transformer approach. Front Nutr 11, 1454466 (2024).
18. S. S. Alahmari, M. R. Gardner, T. Salem, Attention guided approach for food type and state recognition. Food and Bioproducts Processing 145, 1-10 (2024).
19. S. E. Sreedharan, G. N. Sundar, D. Narmadha, NutriFoodNet: A High-Accuracy Convolutional Neural Network for Automated Food Image Recognition and Nutrient Estimation. Traitement du Signal 41, (2024).
20. Z. Wu, C. Shen, A. Van Den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition. Pattern recognition 90, 119-133 (2019).
21. L. Bossard, M. Guillaumin, L. Van Gool, in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars, Eds. (Springer International Publishing, Cham, 2014), pp. 446-461.
22. J. Chen, C.-w. Ngo, paper presented at the Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 2016.
23. K. M. Wong, L. M. Po, K. W. Cheung, in 2007 IEEE International Conference on Image Processing. (2007), vol. 6, pp. VI - 365-VI - 368.
24. M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, E. J. Delp, in 2011 19th European Signal Processing Conference. (2011), pp. 764-768.
25. Y. He, C. Xu, N. Khanna, C. J. Boushey, E. J. Delp, in 2014 IEEE International Conference on Image Processing (ICIP). (2014), pp. 2744-2748.
26. D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60, 91-110 (2004).
27. P. F. Felzenszwalb, Representation and detection of deformable shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 208-220 (2005).
28. S. Yang, M. Chen, D. Pomerleau, R. Sukthankar, in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2010), pp. 2249-2256.
29. H. Hoashi, T. Joutou, K. Yanai, in 2010 IEEE International Symposium on Multimedia. (2010), pp. 296-301.
30. B. Shah, H. Bhavsar, Depth-restricted convolutional neural network—a model for Gujarati food image classification. The Visual Computer 40, 1931-1946 (2024).
31. Y.-C. Liu, D. D. Onthoni, S. Mohapatra, D. Irianti, P. K. Sahoo, Deep-Learning-Assisted Multi-Dish Food Recognition Application for Dietary Intake Reporting. Electronics. 2022 (10.3390/electronics11101626).

 doi:10.20944/preprints202507.1310.v1

32. C. Wang et al., Application of Convolutional Neural Network-Based Detection Methods in Fresh Fruit Production: A Comprehensive Review. Front Plant Sci 13, 868745 (2022).
33. K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian, Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. IEEE Transactions on Image Processing 16, 2080-2095 (2007).
34. L. Zhang, W. Dong, D. Zhang, G. Shi, Two-stage image denoising by principal component analysis with local pixel grouping. Pattern Recognition 43, 1531-1549 (2010).
35. J. Liang et al., in 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). (2021), pp. 1833-1844.
36. J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models. Advances in neural information processing systems 33, 6840-6851 (2020).
37. G. Xu et al., ASQ-FastBM3D: An Adaptive Denoising Framework for Defending Adversarial Attacks in Machine Learning Enabled Systems. IEEE Transactions on Reliability 72, 317-328 (2023).
38. R. Kundu, A. Chakrabarti, P. Lenka, A Novel Technique for Image Denoising using Non-local Means and Genetic Algorithm. National Academy Science Letters 45, 61-67 (2022).
39. P. McAllister, H. Zheng, R. Bond, A. Moorhead, Combining deep residual neural network features with supervised machine learning algorithms to classify diverse food image datasets. Computers in Biology and Medicine 95, 217-233 (2018).
40. W. Zhang, J. Wu, Y. Yang, Wi-HSNN: A subnetwork-based encoding structure for dimension reduction and food classification via harnessing multi-CNN model high-level features. Neurocomputing 414, 57-66 (2020).
41. G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017), pp. 2261-2269.
42. J. Hu, L. Shen, G. Sun, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018), pp. 7132-7141.
43. J. Qiu, F. P.-W. Lo, Y. Sun, S. Wang, B. P. L. Lo, in British Machine Vision Conference. (2019).
44. Y. Chen, Y. Bai, W. Zhang, T. Mei, in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2019), pp. 5152-5161.
45. R. Du et al., in Computer Vision – ECCV 2020, A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm, Eds. (Springer International Publishing, Cham, 2020), pp. 153-168.
46. T. Hu, H. Qi, Q. Huang, Y. Lu, See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:1901.09891, (2019).
47. Z. Yang et al., in Computer Vision – ECCV 2018, V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss, Eds. (Springer International Publishing, Cham, 2018), pp. 438-454.
48. W. Min et al., Large Scale Visual Food Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 9932-9949 (2023).
49. W. Min et al., paper presented at the Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 2020.
50. Z. Liu et al., in Proceedings of the IEEE/CVF international conference on computer vision. (2021), pp. 10012-10022.
51. Z. Xia, X. Pan, S. Song, L. E. Li, G. Huang, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2022), pp. 4794-4803.
52. Z. Wang et al., Ingredient-Guided Region Discovery and Relationship Modeling for Food Category-Ingredient Prediction. IEEE Transactions on Image Processing 31, 5214-5226 (2022).
53. Y. Liu, W. Min, S. Jiang, Y. Rui, Convolution-Enhanced Bi-Branch Adaptive Transformer With Cross-Task Interaction for Food Category and Ingredient Recognition. IEEE Transactions on Image Processing 33, 2572-2586 (2024).
54. W. Park, D. Kim, Y. Lu, M. Cho, in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2019), pp. 3962-3971.
55. B. Peng et al., Correlation Congruence for Knowledge Distillation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 5006-5015 (2019).

 doi:10.20944/preprints202507.1310.v1

56. Y. Liu et al., in 2019 IEEE/CV activation boundaries formed F Conference on Computer Vision and Pattern Recognition (CVPR). (2019), pp. 7089-7097.
57. B. Heo, M. Lee, S. Yun, J. Y. Choi, in Proceedings of the AAAI conference on artificial intelligence. (2019), vol. 33, pp. 3779-3787.
58. P. Dhar, R. V. Singh, K. C. Peng, Z. Wu, R. Chellappa, in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2019), pp. 5133-5141.
59. Q. Wang et al., paper presented at the Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part I, New Delhi, India, 2023.
60. C. Xu et al., Teacher-student collaborative knowledge distillation for image classification. Applied Intelligence 53, 1997-2009 (2023).