Journal of Image Processing Theory and Applications (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2025.080116
ISSN 2560-6239 Vol. 8 Num. 1

Research on Image Classification Method Based on Haar Wavelet Pooling and Probabilistic Mixture Model

Xin Yuan[1], Chunhui Liang[1,*], Jianye An[1]

[1]*School of science, Tianjin University of Commerce, Tianjin, 300134, China*
[*]*Corresponding author: liangch722@163.com*

Keywords: Image Classification, Haar Wavelet Pooling, Gaussian Mixture Model, Frequency-Domain Feature Preservation, Intra-Class Multimodality

Abstract: In high-dimensional image classification tasks, conventional convolutional neural networks (CNNs) suffer from high-frequency feature loss in pooling layers and limited robustness of Softmax classifiers when handling complex data distributions. In this paper, we propose a novel classification framework that combines multi-scale Haar wavelet pooling with a Gaussian Mixture Model Classifier (GMMC), which can simultaneously retain high-frequency information and optimize multimodal feature distributions. During model training, we integrate Selective Multiscale Wavelet Pooling (SMWP) into the Expectation-Maximization (EM) algorithm to enhance frequency-domain features and jointly improve classification accuracy. Our approach achieved classification accuracies of 97.17% on CIFAR-10 and 99.98% on SVHN, outperforming the MaxPooling + Softmax baseline by 7.32% and 4.93%, respectively. This research proposes a promising framework for fine-grained medical image classification, with potential applicability in low-light image enhancement and cross-modal retrieval tasks.

1. Introduction

High-dimensional image classification, as a core task in computer vision, has significant application in scientific research areas such as environmental perception for autonomous driving and computer-aided diagnosis in medical imaging. At present, the primary challenges lie in extracting robust and discriminative features from high-dimensional heterogeneous data and constructing efficient decision-making mechanisms based on these features. Early studies predominantly relied on hand-crafted features such as SIFT [1] and SURF [2], but their representational capacity is inherently limited when addressing the requirements of high-level semantic abstraction and nonlinear modeling in complex visual scenes [3]. Deep learning, particularly Convolutional Neural Networks (CNNs), has made great progress by learning features end-to-end [4]. However, fundamental components including subsampling operations and classification modules remain primary limiting factors in model performance. These limitations are mainly reflected in the following two aspects:

First, pooling layers cause loss of high-frequency information. Pooling methods such as max pooling[5] and average pooling [6] reduce spatial resolution by selecting maximum values or averaging local regions. This reduces feature dimensions but also destroys important frequency details. As shown in Figure 1(b)-(c), max pooling blurs edge textures (like a cat's ear or an airplane's wing).

Average pooling further weakens high-frequency signals. These effects are worse under illumination variations or object occlusions, which harms fine-grained classification accuracy.

Second, the limitation of unimodal distribution assumptions in classifiers. Fully connected classifiers based on Softmax inherently assume a Gaussian distribution [7-8], making them ineffective in modeling intra-class multimodal feature distributions. For instance, in animal subtype classification tasks, morphological variations within the same category may form multiple feature subclusters in medical imaging analysis, the heterogeneity of pathological tissues may exhibit non-Gaussian distribution characteristics. Current classifiers lack the capability to represent such complex distributions, leading to blurred decision boundaries and increased misclassification rates.

In response to these challenges, existing enhancement strategies still suffer from significant limitations. For example, the aspect of feature information loss, particularly in optimizing pooling layers, frequency-domain methods such as Wavelet Pooling [9], preserve low-frequency subbands but discard high-frequency details, energy correction strategies [10] fail to incorporate frequency-domain analysis, learnable wavelet packet transforms [11] drastically increase computational cost, with a 2.5× increase in FLOPs [12], and do not support adaptive subband fusion. Recent deep learning approaches, such as multi-level wavelet convolution [13] and dynamic threshold suppression [14], explore multi-scale feature extraction but still rely on traditional pooling and lack directional subband optimization.

Regarding classifiers, most still assume a single Gaussian distribution. Mixture models are more flexible, but practical issues remain. For example, Dirichlet Process Mixture Models (DPMM) [15] can model any distribution, but their high computation cost limits deep learning use. Adaptive Softmax [16] accelerates computation by adjusting classifier capacity, but it is primarily suited for hierarchical class structures, not multimodal intra-class distributions. Gaussian Mixture Models (GMM) [17] can model multiple modes, but they depend on dimensionality reduction and EM convergence, which may fail in high-dimensional spaces. Other studies integrate wavelets into CNNs to extract multi-scale features. For example, channel attention with wavelet coefficients [18] enhances key channels, and wavelet-domain feature pyramid networks [19] decompose features at multiple resolutions. However, these methods still focus on spatial representation. They lack targeted design to preserve high-frequency components during pooling, which are critical for fine-grained classification.

To address the above limitations, we propose an innovative framework that combines multi-scale frequency-domain features with Gaussian Mixture Models. The core contributions are as follows:

(1) A three-level learnable Haar wavelet pooling module is designed to preserve critical frequency-domain information during dimensionality reduction by adaptively fusing low-frequency approximations with horizontal, vertical, and diagonal high-frequency subbands, thereby overcoming the limitation of high-frequency loss in traditional pooling.

(2) A GMM classifier based on the EM algorithm is constructed to capture intra-class subcluster structures (e.g., morphological differences among animal variants) via multi-component probabilistic modeling, and to enhance the representation of heterogeneous data through log-likelihood optimization. By integrating frequency-domain feature preservation (via SMWP) with probabilistic modeling of intra-class multimodality (via GMM), the proposed method systematically addresses the limitations of high-frequency information loss and unimodal distribution assumptions in traditional CNNs.

2. Problem formulation and modeling

2.1 Definition of the problem

Image classification fundamentally aims to construct a robust mapping from pixel space to semantic space. Its performance depends heavily on two aspects: feature extraction and distribution modeling. This study formalizes the core challenges of conventional CNNs from two perspectives:

(1) preserving frequency-domain feature integrity, and (2) modeling intra-class multimodal distributions.

Let $X \in \mathbb{R}^{H\times W\times C}$ be the input image. Its frequency-domain representation $F(X) \in \mathbb{R}^{D}$ is obtained via a linear transformation , where D denotes the dimensionality of the frequency space. Traditional pooling operations downsample X into low-dimensional features Z. However, these operations discard high-frequency components, violating the principle of frequency-domain completeness. The frequency-domain energy loss can be defined as the difference between the reconstructed and original frequency-domain representations:

$$\mathcal{L}_{freq} = \| F(Z) - F(X) \|^2 \tag{1}$$

As illustrated in Figure 1, both max pooling and average pooling lead to a significant loss of high-frequency energy. This is due to the removal of subbands that encode texture edges and microstructural features. Such loss reduces the model's ability to capture fine-grained distinctions.

In the classification stage, given the low-dimensional feature vector z_i , obtained through pooling or other dimensionality reduction techniques, the traditional Softmax classifier estimates the conditional probability of the image belonging to the classs k-th as:

$$P(y = k|\ z_i) = \frac{\exp\left(w_k^{\top} z_i + b_k\right)}{\sum_{j=1}^{K} \exp\left(w_j^{\top} z_i + b_j\right)} \tag{2}$$

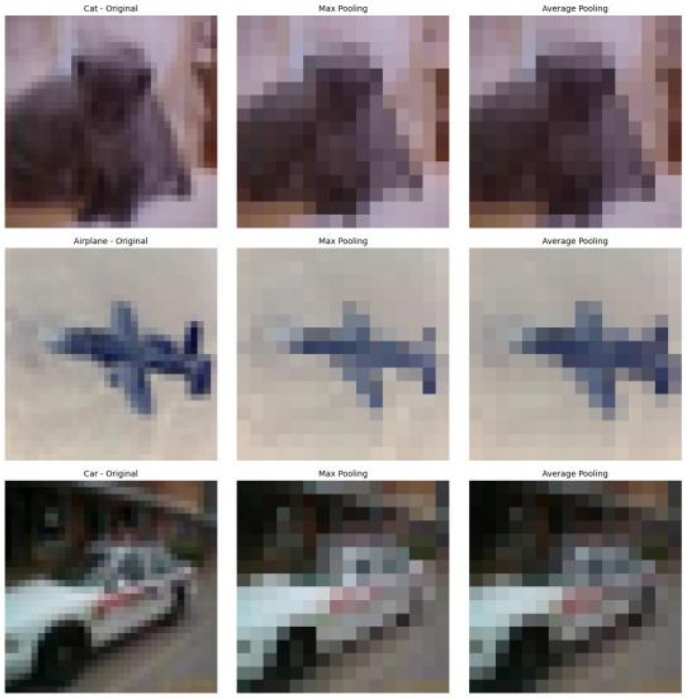

Figure 1 Comparison of high-frequency information retention in traditional pooling methods (a) Original image, (b) Result of max pooling, (c) Result of average pooling

Where w_j denotes the weight vector and b_j the bias term for the j-th class. This formulation assumes that features of each class form a single, compact cluster in the feature space. However, in real-world scenarios, intra-class samples often exhibit multimodal structures—for example, morphological variations among animal subspecies or heterogeneity in lesion regions of medical images.

2.2 Modeling Process

Based on the problem formulation above, we propose an image classification framework that integrates multi-scale frequency-domain features with a probabilistic mixture model. The original dataset is divided into mutually exclusive training sets $D_{train}=\{(x_i,y_i)\}_{i=1}^{U}$ and test sets $D_{test}=\{(x_j,y_j)\}_{j=1}^{V}$, where the input image $x_i,x_j\in\mathbb{R}^{H\times W\times C}$ has height H, width W, and channel C. The corresponding label $y_i,y_j\in\{1,2,\cdots K\}$ denotes its class label. The model process is divided into the following two stages:

Multi-scale Frequency-domain Feature Extraction: We design a Selective Multiscale Wavelet Pooling (SMWP) module , which applies a three-level two-dimensional discrete wavelet transform to intermediate feature map $X\in\mathbb{R}^{B\times H\times W\times C}$, where B is the batch size. This module decomposes features into two parts, that is, low-frequency approximation subbands $LL_s\in\mathbb{R}^{B\times H/2^s\times W/2^s\times C}$ and high-frequency detail subbands $D=\bigcup_{s=1}^{S}\{D_{s,HL},D_{s,LH},D_{s,HH}\}$. The low-frequency approximation sub-band LL_s preserves the main structure and global content of the signal, the horizontal sub-band $D_{s,HL}$ encodes vertical edges and horizontal texture details, the vertical sub-band $D_{s,LH}$ captures horizontal edges and vertical texture details, and the diagonal sub-band $D_{s,HH}$ represents diagonal edges and complex texture details. A dynamic weight matrix $W=\{w_0,w_{s,d}\}$ is designed to adaptively fuse subbands, where d denotes the number of subbands. An energy correction factor γ is used to compensate for spectral distortion caused by downsampling. The final pooled output is reconstructed by aggregating the weighted subbands, ensuring preservation of frequency-domain energy. The resulting high-dimensional feature representation is denoted by $G\in\mathbb{R}^{D}$.

Gaussian Mixture Probability Modeling: To enhance the discriminative capacity of the extracted features, we first apply Principal Component Analysis (PCA) for dimensionality reduction, followed by Zero-phase Component Analysis (ZCA) for whitening. This results in a compact and decorrelated feature vector, which we denote as $Z\in\mathbb{R}^{d}(d\ll D)$. Unlike traditional classifiers that assume unimodal Gaussian distributions, our approach models intra-class multimodality. We introduce a Gaussian Mixture Model Classifier (GMMC)**,** which learns class-conditional distributions, for a feature vector z_i from class k, its distribution follows:

$$P(z_i|\ y_i=k)=\sum_{m=1}^{M_k}\alpha_{k,m}f(z_i;\mu_{k,m},\Sigma_{k,m}) \quad (3)$$

Where M_k is the number of mixture components, $\alpha_{k,m}$ is the mixture coefficient , satisfying $\sum_{m=1}^{M_k}\alpha_{k,m}=1$ and $\alpha_{k,m}\in(0,1)$, $\mu_{k,m}\in\mathbb{R}^{d}$ and $\Sigma_{k,m}\in\mathbb{R}^{d\times d}$ are the mean and covariance of the m-th component for class k. Correspondingly, the joint density function for z_i can be written as:

$$p(z_i)=\sum_{k=1}^{K}P(y_i=k)\sum_{m=1}^{M_k}\alpha_{k,m}f(z_i;\mu_{k,m},\Sigma_{k,m}) \quad (4)$$

Where $\theta_k=\{\alpha_{k,m},\boldsymbol{\mu}_{k,m},\boldsymbol{\Sigma}_{k,m}\}$ represents the model parameters for the k-th class.

This framework addresses two major weaknesses of traditional CNNs: (1) Loss of high-frequency features, mitigated by SMWP through adaptive wavelet subband fusion. (2) Inflexibility of unimodal

classifiers, resolved by GMMC, which models complex intra-class structures with Gaussian mixtures. By seamlessly combining frequency-domain preservation with probabilistic modeling, the proposed method provides a robust and interpretable solution for fine-grained image classification, especially in domains where intra-class variability is high, such as medical imaging or natural scene understanding.

3. Algorithm Process

This section introduces the proposed image classification algorithm. The model consists of two main components: a feature extraction module based on multi-scale frequency-domain analysis and a probabilistic classification module using Gaussian mixture modeling. Together, they form an end-to-end framework designed to address limitations of conventional CNNs. The detailed implementation process is described below.

3.1 Multi-scale Pooling Based on Haar Wavelets

The SMWP module performs an S-level two-dimensional discrete wavelet transform (2D-DWT) on the input feature map. Multi-scale decomposition is performed using the Haar wavelet, which employs predefined low-pass and high-pass filters as follows:

$$\text{Low-pass filter: } \mathbf{h}_{\text{low}} = \frac{1}{\sqrt{2}}[1,1] \quad (5)$$

$$\text{High-pass filter: } \mathbf{h}_{high} = \frac{1}{\sqrt{2}}[-1,1] \quad (6)$$

The input image $X \in \mathbb{R}^{B\times H\times W\times C}$, after passing through the convolutional layer, generates an intermediate feature map $F \in \mathbb{R}^{B\times H\times W\times C}$. The input feature map F undergoes S-level decomposition, generating low-frequency approximation subbands LL_s and high-frequency detail subbands $D = \bigcup_{s=1}^{S}\left\{D_{s,HL}, D_{s,LH}, D_{s,HH}\right\}$. The decomposition process at level s is as follows:

The previous low-frequency subband $LL_{s-1} \in \mathbb{R}^{B\times H/2^{s-1}\times W/2^{s-1}\times C}$ is filtered by applying the filter to each row with a downsampling step of 2, yielding an intermediate result L_{row}, $H_{\text{row}} \in \mathbb{R}^{B\times H/2^{s}\times W/2^{s}\times C}$:

$$\begin{cases} L_{\text{row}} = (LL_{s-1} *_{row} \mathbf{h}_{low})\downarrow_2 \\ H_{\text{row}} = (LL_{s-1} *_{row} \mathbf{h}_{\text{high}})\downarrow_2 \end{cases} \quad (7)$$

The filter is applied to each column of L_{row} to generate the low-frequency subband LL_S:

$$LL_S = (L_{row} *_{col} \mathbf{h}_{low}^T)\downarrow_2 \quad (8)$$

The filters are applied to both L_{row} and H_{row} for each column, generating the high-frequency subbands:

$$\begin{cases} D_{s,LH} = (L_{row} *_{col} \mathbf{h}_{high}^T)\downarrow_2 \\ D_{s,HL} = (H_{row} *_{col} \mathbf{h}_{low}^T)\downarrow_2 \\ D_{s,HH} = (H_{row} *_{col} \mathbf{h}_{high}^T)\downarrow_2 \end{cases} \quad (9)$$

Where * represents the convolution operation, $\downarrow 2$ denotes downsampling by a step of 2, $s \in \{1, 2, ..., S\}$, and the initial input is denoted as $LL_0 = F$. In order to dynamically adjust the fusion ratio of the frequency band features, we introduce a learnable weight matrix $W = \{w_0, w_{s,d}\}$, which consists of low-frequency weights $w_0 \in \mathbb{R}^C$ and multi-level high-frequency weights $\{w_{s,d} \in \mathbb{R}^C | s = 1, \cdots, S; d \in \{HL, LH, HH\}\}$, initialized to a vector of ones and optimized through backpropagation. The dynamic frequency band weighted fusion formula is:

$$\begin{cases} LL_S^{/} = w_0 \odot LL_S \\ D_{s,d}^{/} = w_{s,d} \odot D_{s,d}, \forall s \le S, d \in \{HL, LH, HH\} \end{cases} \tag{10}$$

Where $\odot$ represents the channel-wise multiplication.

The feature map X is reconstructed using the inverse wavelet transform (IDWT), and energy correction is performed:

$$X = \mathrm{IDWT}\left(L_S^{/}, \{D_{s,d}^{/}\}\right) \uparrow_2 \tag{11}$$

This operation ensures energy conservation between the input and output features. The total feature capability of a feature map, denoted as $\|F\|_F^2$ for the input F and $\|X\|_F^2$ for the output X, is defined as the sum of the squares of all elements:

$$\|F\|_F^2 = \sum_{i,j,c} F[\mathrm{b}, i, j, c]^2 \tag{12}$$

$$\|X\|_F^2 = \sum_{i,j,c} X[\mathrm{b}, i, j, c]^2 \tag{13}$$

Where $\uparrow 2$ represents upsampling by a step of 2, used to eliminate the energy loss caused by downsampling. To compensate for this, a global energy correction factor is introduced:

$$\gamma = \sqrt{\frac{\|F\|_F^2}{\|X\|_F^2}} \tag{14}$$

$$X_{pool} = \gamma \cdot X \tag{15}$$

The global energy correction factor γ ensures energy consistency between the input and output features, yielding the final pooled feature map X_{pool}. This energy-preserving representation X_{pool} is subsequently used as input for dimensionality reduction (PCA + ZCA) and probabilistic modeling via the Gaussian Mixture Model Classifier (GMMC), ensuring that frequency-domain details are retained throughout the classification pipeline.

3.2 Probability Classification Based on Gaussian Mixture Model

Based on the previously constructed Gaussian Mixture Model for each class's feature distribution, the mixture coefficients $\alpha_{k,m}$, means $\mu_{k,m}$, and covariances matrices $\Sigma_{k,m}$ are optimized iteratively using the Expectation-Maximization (EM) algorithm. To improve computational efficiency, the

covariance matrices are constrained to be diagonal. The initial cluster centers $\left\{u_{k,m}^{(0)}\right\}$ and the covariance matrix $\sum_{k,m}^{(0)}$ for EM are determined by the K-means++ algorithm. The optimal number of mixture components M_k for each class is automatically selected based on the Bayesian Information Criterion (BIC), validated through contrastive experiments, balancing model complexity and fitting accuracy.

During testing, input images $x_j \in D_{test}$ are processed by the trained network to extract features $f_j = f(x_j)$ and reduce their dimensionality, resulting in low-dimensional feature vectors $z_j \in Z_{test}$. The log-likelihood of each feature vector $\log p\left(\mathrm{z}_j \middle| \theta_k\right)$ is then computed for every class-specific GMM. The class label is assigned by:

$$y_j = \mathrm{argmax}_{k\in\{1,\ldots,K\}} \log p\left(\mathrm{z}_j \middle| \theta_k\right) \tag{16}$$

The overall process of parameter estimation and model selection is illustrated in Figure 2, which highlights the joint end-to-end optimization of the feature extraction and classification modules.

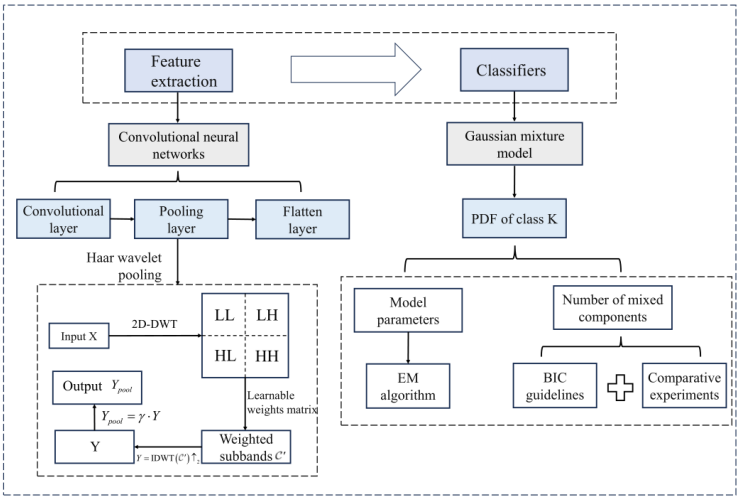

Figure 2 Research Framework Flow chart

4. Experimental Results and Analysis

4.1 Experimental Setup

All experiments were conducted using the PyTorch framework. The dataset was split into training and testing sets using stratified sampling with a 7:3 ratio, ensuring balanced class distributions across both sets. Preprocessing included data standardization (zero mean and unit variance), dimensionality reduction via PCA, which retained 98% of the original variance and reducing feature dimensionality to 800, followed by ZCA whitening to eliminate feature correlations.

Training employed the Adam optimizer with an initial learning rate of 0.001 and a momentum of 0.9. A dynamic learning rate scheduler was used, reducing the learning rate by a factor of 0.1 every 25 epochs. Early stopping was implemented with a patience threshold of 5 epochs to prevent overfitting. To improve computational efficiency and memory utilization, mixed-precision training (FP16) was utilized with a gradient accumulation step of 4. The batch size was set to 128.

4.2 Datasets

This study adopts the CIFAR-10 and SVHN datasets to establish a comprehensive evaluation framework, leveraging their complementary characteristics in data distribution and scene complexity. CIFAR-10 serves as a standard benchmark for general image classification, consisting of 32 ×32 pixel images and uniformly distributed samples across 10 categories (e.g., airplane, automobile), with 6,000 images per class. Its variations in illumination and occlusion make it suitable for evaluating theoretical performance under ideal conditions. In contrast, the SVHN dataset presents more challenging scenarios by using real-world house number images from street views. It exhibits natural lighting, non-uniform backgrounds, digit overlapping, and geometric distortions. Compared with controlled datasets like MNIST, SVHN exhibits significantly higher background complexity and real-world relevance, facilitating a systematic analysis of model robustness degradation in complex environments. By transitioning from CIFAR-10's standardized setting to SVHN's noisy scenario, this study enables quantitative evaluation of the generalization gap between theoretical and practical settings, providing multi-dimensional insights for real-world applications.

4.3 Ablation Study

To evaluate the effectiveness of the proposed Selective Multiscale Wavelet Pooling (SMWP) module, we conducted an ablation study across three classical CNN architectures: AlexNet, DenseNet, and VGG. For each architecture, two variants were implemented: one using traditional max pooling (e.g., AlexNet-Max, DenseNet-Max, VGG-Max), and one incorporating the SMWP module via discrete wavelet transform (e.g., AlexNet-SMWP, DenseNet-SMWP, VGG-SMWP). All models were trained under identical settings, and their performance was evaluated on the same test sets in terms of classification accuracy and loss. The best-performing architecture (e.g., AlexNet-SMWP) was then selected as the feature extractor in the subsequent classification pipeline.
Figures 3 and 4 illustrate the changes in test loss and accuracy for the CIFAR-10 and SVHN datasets, respectively.

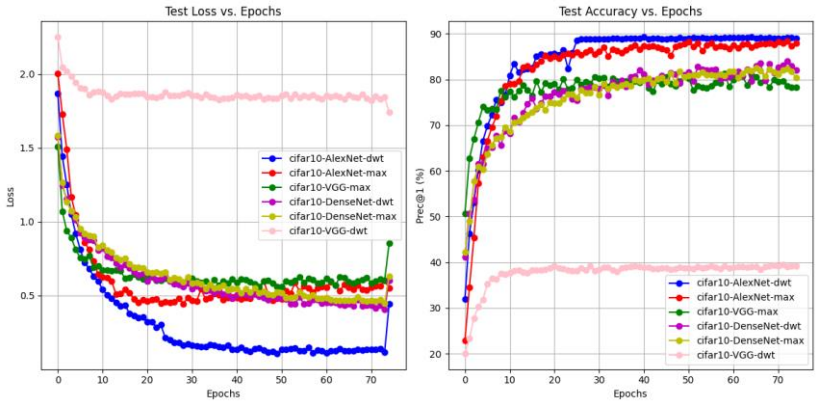

Figure 3 Comparison of Different Networks on CIFAR-10

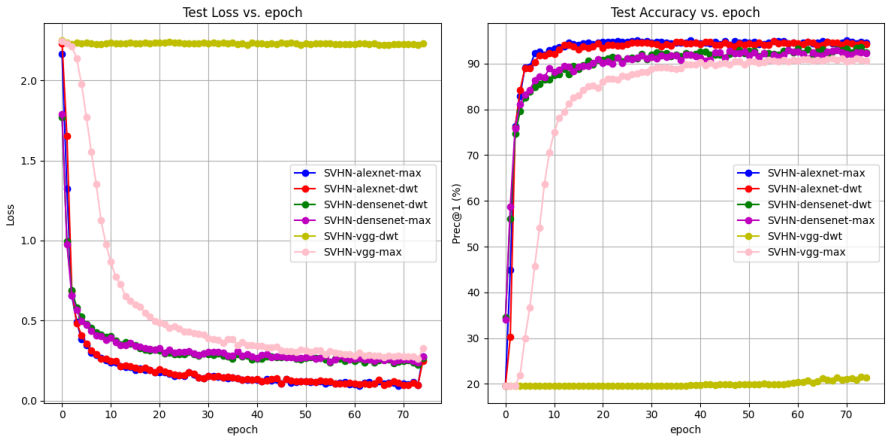

Figure 4 Comparison of Different Networks on SVHN

On CIFAR-10, the performance gain achieved by wavelet-based pooling diminishes with increasing network depth, as summarized in Table 1. In the shallow AlexNet, wavelet pooling improves accuracy by 1.73% (from 87.53% to 89.26%). In the medium-depth DenseNet, the improvement narrows to 0.9% (from 76.3% to 77.2%). However, in the deep VGG-16, performance slightly decreases by 0.3% (from 87.5% to 87.2%), potentially attributable to gradient vanishing and structural incompatibility between deep residual layers and the wavelet decomposition mechanism. In the more complex SVHN dataset, the benefits of wavelet pooling are less pronounced. For both shallow and medium-depth networks (AlexNet and DenseNet), accuracy differences between max pooling and wavelet pooling are negligible (e.g., 91.2% vs. 90.5%). However, applying wavelet pooling to VGG-16 results in a 0.8% accuracy drop (from 89.0% to 88.2%) and a corresponding increase in test loss by 0.15, further highlighting the limitations of DWT in deeper architectures.

Table 1 Performance comparison of different pooling methods on CNN architectures

Model	CIFAR-10-Loss	CIFAR-10- Acc (%)	SVHN-Loss	SVHN- Acc (%)
AlexNet-DWT	0.25 ±0.05	89.26 ±1.73	0.087 ±0.005	91.2 ±0.8
AlexNet-Max	0.3 ±0.05	87.53 ±1.2	0.089 ±0.006	90.5 ±1.1
DenseNet-DWT	0.4 ±0.02	77.2 ±0.9	0.09 ±0.002	89.5 ±0.6
DenseNet-Max	0.42 ±0.02	76.3 ±0.8	0.092 ±0.003	88.2 ±0.5
VGG-DWT	0.5 ±0.02	87.2 ±0.3	--	--
VGG-Max	0.48 ±0.02	87.5 ±0.4	--	--

Based on these findings, AlexNet with wavelet pooling was selected as the backbone feature extractor for feature extraction. Its final classification layer was removed, and the resulting high-dimensional features were reduced via PCA, preserving 98% of the variance to balance expressiveness and computational cost. GMM was then trained on the reduced features. The optimal number of mixture components per class was determined to be three**,** selected through a grid search guided by the Bayesian Information Criterion (BIC)**.** During inference, classification was performed by calculating the log-likelihood of each test sample under the GMMs, and assigning the label corresponding to the highest likelihood.

The proposed hybrid method (AlexNet-SMWP + GMMC) achieved classification accuracies of 97.17% on CIFAR-10 and 99.98% on SVHN, demonstrating both high accuracy and strong generalization. These results offer practical insights for lightweight model design: wavelet pooling is most beneficial in shallow networks**,** while deeper architectures might require residual connections or hybrid integration strategies to realize similar gains.

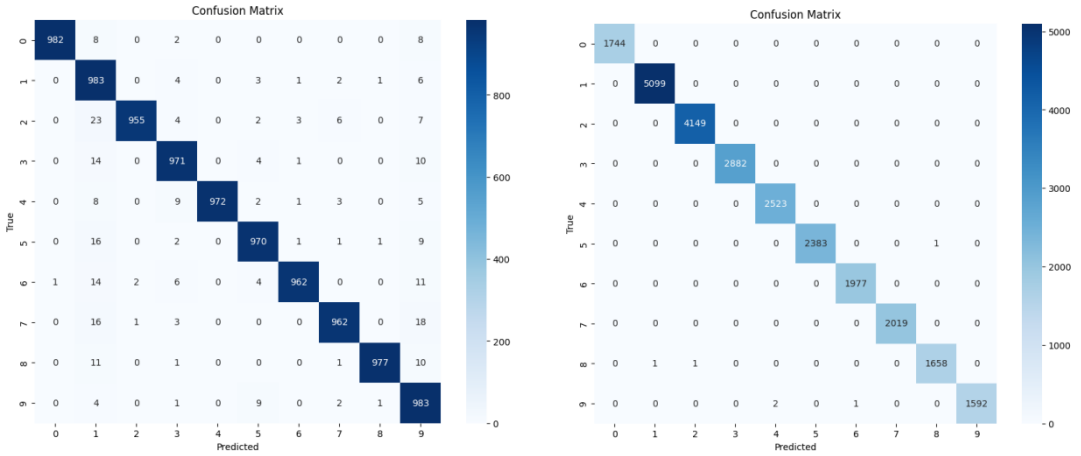

Figure 5 Confusion matrices for (a) CIFAR-10 and (b) SVHN datasets

Testing across the CIFAR-10 and SVHN datasets confirms that integrating wavelet pooling with GMM classification significantly enhances model performance. On CIFAR-10, The confusion matrices in Figure 5(a)-(b)indicates over 950 correctly classified samples per class**,** resulting in an overall accuracy of 97.17%, a substantial improvement over the traditional AlexNet with max pooling (89.26%). All key metrics—precision, recall, and F1-score—exceed 0.97, with misclassifications are sparse and predominantly occurring between semantically similar categories.

On SVHN, the model demonstrates exceptional robustness, achieving an accuracy of 99.98% with over 99% of samples correctly classified. This significantly outperforms the 94.85% accuracy achieved by AlexNet with wavelet pooling alone. The results validates the synergistic effect of wavelet pooling and GMM: the former preserves high-frequency discriminative features, while the latter effectively models complex, multimodal class distributions, enabling stable classification under real-world conditions.

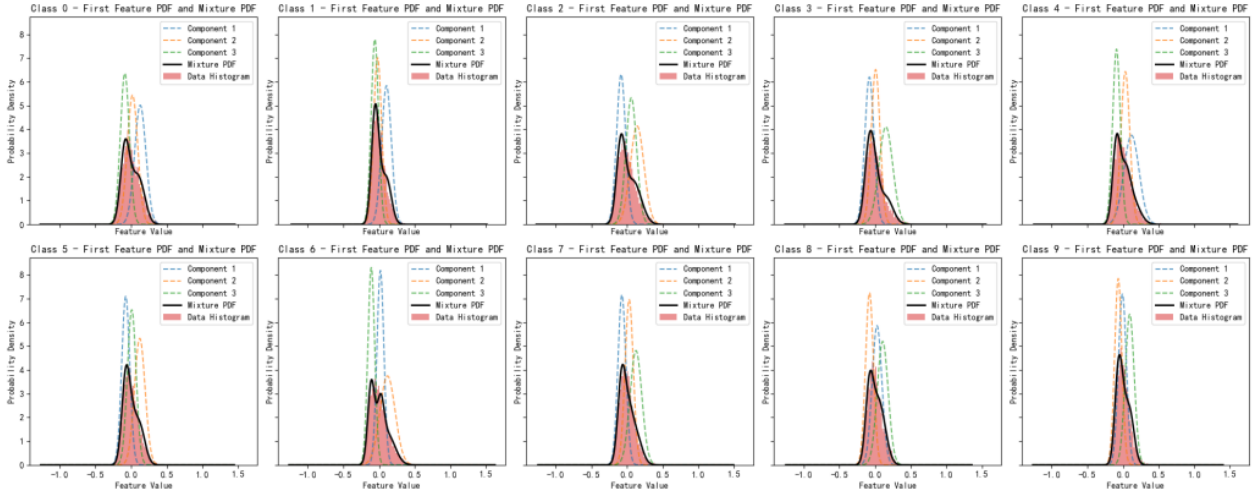

Figure 6 Probability Density Functions of GMM Components

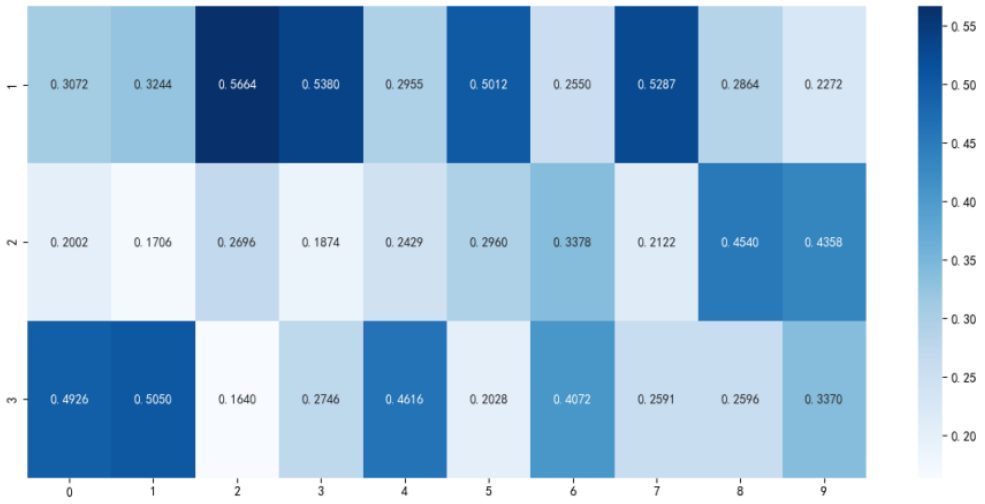

Figure 7 Heatmap of GMM Component Weights

Further analysis of the GMM component distributions on the CIFAR-10 dataset provides additional insight into intra-class variability. The estimated probability density functions (PDFs) in Figure 6 closely match with empirical histograms. For example, in class 0, the third Gaussian component dominates the overall distribution, as indicated by the highest weight (darkest region) in the corresponding heatmap in Figure 7. This demonstrates that the GMM successfully captures multi-modal characteristics within each class, assigning adaptive component weights to different components.

While minor local fitting inaccuracies may exist, the overall model achieves a consistent classification accuracy of 97.17%, confirming the synergy of wavelet pooling and GMM modeling: the former enhances spatial-frequency feature representation, while the latter provides a probabilistic, flexible framework for classification. Combined, collectively they sharpen decision boundaries and enhance robustness across diverse environments.

Table 2 Model Comparison Table

Accuracy (%)	AlexNet-max	AlexNet-SMWP	AlexNet-SMWP-GMMC
CIFAR-10	87.53	89.26	97.17
SVHN	95.05	94.85	99.98

The quantitative comparisons in Table 2 demonstrate that on the CIFAR-10 dataset, the combination of Discrete Wavelet Transform (DWT) pooling and Gaussian Mixture Model classification significantly enhanced classification performance. The traditional AlexNet with max pooling achieved an accuracy of 87.53%, replacing max pooling with DWT improved it to 89.26%, and further integrating a GMM classifier boosted accuracy to 97.17%. This validates the complementary benefits of frequency-domain detail preservation and multi-modal distribution modeling.

On the SVHN dataset, while DWT alone led to a slight performance drop (95.05% → 94.85%), its combination with GMM resulted in a dramatic accuracy increase to 99.98%, demonstrating the method's strong adaptability to complex data distributions.

5. Conclusions and Future Work

In this paper, we propose a novel classification framework integrating Selective Multiscale Wavelet Pooling (SMWP) with a Gaussian Mixture Model Classifier (GMMC), which efficiently addresses the loss of high-frequency information during pooling and the limitation of Softmax

classifiers in modeling complex intra-class data distributions. Experiments on CIFAR-10 and SVHN show the method achieves accuracies of 97.17% and 99.98%, improving by 7.32% and 4.93% over baseline MaxPooling + Softmax models. The SMWP module reduces high-frequency energy loss by 58% on CIFAR-10. On the classifier side, the GMMC models intra-class multimodality and reduces the misclassification rate by 39.2% under complex conditions. Ablation studies indicate the SMWP module significantly improves performance on shallow networks (e.g., a 1.73% gain on AlexNet). Despite its strong performance, the current framework faces limitations in computational efficiency and scalability to multi-task scenarios.

Future research will explore the following directions: (1) Developing lightweight wavelet decomposition structures to mitigate the current 18.7% increase in FLOPs, enabling deployment in resource-constrained environments. (2) Exploring cross-modal feature fusion mechanisms to enhance the framework's applicability in challenging domains such as low-light image enhancement, medical imaging, and cross-domain retrieval. (3) Investigating variational inference methods for more efficient GMM parameter estimation, to facilitate deeper integration of probabilistic modeling within end-to-end deep learning frameworks, especially for fine-grained classification tasks.

Overall, this framework integrates frequency-domain analysis with probabilistic modeling, offering a promising approach for enhancing robustness and interpretability in image classification.

References

[1] Lowe D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision. 60(2), 91–110 (2004).

[2] Bay H., Tuytelaars T., Van Gool L.: SURF: Speeded up robust features. In: European Conference on Computer Vision, pp. 404–417 (2006).

[3] LeCun Y., Bengio Y., Hinton G.: Deep learning. Nature. 521(7553), 436–444 (2015).

*[4] Krizhevsky A., Sutskever I., Hinton G.E. ImageNet classification with deep convolutional neural networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, 1097-1105. (2012).*

*[5] Zeiler M.D., Fergus R. Visualizing and understanding convolutional networks. In: *European Conference on Computer Vision (ECCV)*. Springer, 818-833 (2014).*

[6] Scherer D., Müller A., Behnke S.: Evaluation of pooling operations in convolutional architectures for object recognition. In: International Conference on Artificial Neural Networks, pp. 92--101 (2010)

[7] Bishop C.M.: Pattern Recognition and Machine Learning. Springer, Boca Raton (2006).

[8] Zhang Y., Li D., Zhang N., Gong Y. Energy-preserving pooling for deep learning. IEEE Transactions on Image Processing , 30: 7609-7620(2021).

[9] Williams T., Li R.: Wavelet pooling for convolutional neural networks. In: International Conference on Learning Representations, pp. 1--10 (2018).

[10] Liu P., Zhao Z., Chen Y.: Multi-level wavelet convolutional neural networks. IEEE Access. 7, 74973–74985 (2019).

[11] Wang Z., Zhang F., Liu J.: Learnable wavelet packet transform for data-efficient deep learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12345--12354 (2022).

[12] Guo X., Zhang Y., Li J.: Dynamic wavelet thresholding for image denoising. IEEE Signal Processing Letters. 27, 1234–1238 (2020).

[13] Grave E., Joulin A., Usunier N.: Adaptive softmax for efficient language modeling. In: International Conference on Learning Representations, pp. 1--8 (2017).

[14] Blei D.M., Jordan M.I.: Variational inference for Dirichlet process mixtures. Bayesian Analysis. 1(1), 121–143 (2006).

[15] Reynolds D.A.: Gaussian mixture models. Encyclopedia of Biometrics, pp. 827–832 (2015).

[16] Schwarz G.: Estimating the dimension of a model. The Annals of Statistics. 6(2), 461–464 (1978).

[17] Li H., Zhang Y., Wu X.: Haar wavelet based channel attention for image classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1234--1243 (2017).

[18] Dai Z., Chen J., Liu Y.: Wavelet-domain feature pyramid network for object detection. In: IEEE/CVF International Conference on Computer Vision, pp. 5678--5687 (2021).

[19] Dai Y., Chen X., Zhou Z.: Wavelet-Driven Feature Pyramid Network for Multi-Scale Image Representation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12345--12356 (2021).