FFENet: frequency-spatial feature enhancement network for clothing classification

Feng Yu [1, 2], **Huiyin Li** [1], **Yankang Shi** [1], **Guangyu Tang** [1], **Zhaoxiang Chen** [1], **Minghua Jiang** [Corresp. 1, 2]

[1] School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, Jiangxia District, China

[2] Engineering Research Center of Hubei Province for Clothing Information, Wuhan, Jiangxia District, China

Corresponding Author: Minghua Jiang
Email address: minghuajiang@wtu.edu.cn

Clothing analysis has been widely concerned by people, and clothing classification, as one of the most basic technologies, plays a very important role in the field of clothing analysis. Due to the complexity of clothing scenes in real scenes, it is of profound significance to study the classification of clothing with small sample data sets in complex scenes. The learning of clothing features in complex scenes is disturbed. Because clothing classification relies on the contour and texture information of clothing, clothing classification in such scenes may lead to poor classification results. Therefore, this paper proposes clothing classification network based on frequency-spatial domain conversion, which combines frequency domain information with spatial information and does not compress channels. It aims to enhance the extraction of clothing features and improve the accuracy of clothing classification. In our work, 1) we use the frequency domain feature enhancement module to realize the preliminary extraction of clothing features, 2) we combine the frequency domain information and spatial information to establish a clothing feature extraction clothing classification network without compressed feature map channels, and 3) we organize a clothing dataset in complex scenes (clothing 8). The effectiveness of our network is verified on this dataset and the public clothing dataset fashion-mnist. Our network achieves 93.4% top-1 model accuracy on clothing 8 dataset and 94.62% top-1 model accuracy on fashion-mnist dataset, which also has a very significant improvement in other evaluation metrics such as recall and precision

FFENet: Frequency-Spatial Feature Enhancement Network for Clothing Classification

Feng Yu[1,2], **Huiyin Li**[1], **Yankang Shi**[1], **Guangyu Tang**[1], **Zhaoxiang Chen**[1], **and Minghua Jiang**[1,2]

[1]**School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China.**
[2]**Engineering Research Center of Hubei Province for Clothing Information, Wuhan 430200, China.**

Corresponding author:
Minghua Jiang[1,2]

Email address: minghuajiang@wtu.edu.cn

ABSTRACT

Clothing analysis has been widely concerned by people, and clothing classification, as one of the most basic technologies, plays a very important role in the field of clothing analysis. Due to the complexity of clothing scenes in real scenes, it is of profound significance to study the classification of clothing with small sample data sets in complex scenes. The learning of clothing features in complex scenes is disturbed. Because clothing classification relies on the contour and texture information of clothing, clothing classification in such scenes may lead to poor classification results. Therefore, this paper proposes clothing classification network based on frequency-spatial domain conversion, which combines frequency domain information with spatial information and does not compress channels. It aims to enhance the extraction of clothing features and improve the accuracy of clothing classification. In our work, 1) we use the frequency domain feature enhancement module to realize the preliminary extraction of clothing features, 2) we combine the frequency domain information and spatial information to establish a clothing feature extraction clothing classification network without compressed feature map channels, and 3) we organize a clothing dataset in complex scenes (clothing 8). The effectiveness of our network is verified on this dataset and the public clothing dataset fashion-mnist. Our network achieves 93.4% top-1 model accuracy on clothing 8 dataset and 94.62% top-1 model accuracy on fashion-mnist dataset, which also has a very significant improvement in other evaluation metrics such as recall and precision

INTRODUCTION

With the rapid popularization of online shopping in the clothing industry, efficient clothing image classification (Shajini and Ramanan, 2022) can not only realize the automatic classification of clothing, but also greatly improve the efficiency of clothing retrieval and virtual try-on. The complexity and variety of clothing scenes lead to the problem of poor clothing classification, which is an urgent problem to be overcome in the application of clothing images in real scenes.

In the field of clothing classification, a lot of researchers have done a lot of research before. Clothing classification is different from other classification tasks in that different clothing categories have certain similarities, and the same category of clothing has its differences such as patterns and colors. At present, the commonly used classification methods can be divided into two main types: 1) traditional machine learning methods (Zhou, 2022; Ölçer et al., 2023), and 2) deep neural network methods (Hassan et al., 2022; Sun et al., 2022; Al Shehri, 2022). In the research of clothing classification based on traditional machine learning methods, some basic classifiers are usually improved. For example, (Zhang et al., 2016) incorporates histogram of oriented gradient (HOG) (Déniz et al., 2011) into the example support vector machine (E-SVM) (Noble, 2006) classifier to achieve robustness to light and improve the accuracy of the E-SVM classifier in clothing classification. Some people also propose to improve the fusion of scale invariant feature transform (SIFT) (Cheung and Hamarneh, 2009) and HOG to realize ethnic clothing classification. Others use texture features and speed up robust features (Bay et al., 2008) obtained by modifying SIFT for clothing classification. Some of the above methodological improvements are just the tip of the iceberg of innovation in the field of clothing classification with traditional machine learning algorithms. More scenes can be adapted in a faster time using traditional machine learning algorithms, but traditional machine learning methods are generally less accurate than deep neural network algorithms. At present, most clothing classification tasks are based on deep neural network methods, which are applied to clothing scenes by improving the mainstream classification models. Some people propose an improved convolutional neural network (CNN) (Kiranyaz et al., 2021; Pan et al., 2023) for clothing classification by adjusting the structure of the original CNN model and increasing the volume of the convolution kernel in the adjusted structure. In the field of clothing, most people improve the effect of clothing classification by improving the structure of neural network and integrating other technologies. For example, (Bai et al., 2019) proposes the introduction of bidirectional convolutional recurrent neural networks , which efficiently handles message-passing to syntactic topology and generates regularized landmark layouts. Two attention mechanisms, both landmark-aware attention and category-driven attention, are designed on the basis of this network to enhance the classification of clothing categories.

These previous works show us that the classification performance of deep neural networks can be improved by adjusting the structure of the neural network and adding feature enhancements according to the clothing scene, and it has also been shown that enhancing information about clothing outlines can improve the accuracy of clothing classification. The following problems still exist in the clothing scene: 1) poor accuracy of clothing classification in complex scenes, 2) it is not enough to improve the accuracy of clothing classification by improving the spatial image features, and 3) different clothing categories have many similar parts, while the texture information of clothing in the same category is variable, which raises the difficulty of clothing classification.

To solve the above problems, we propose a clothing classification network based on frequency-spatial feature enhancement network. The main idea of the framework is as follows: the image input to the network is converted from the spatial domain to the frequency domain using the discrete cosine transform (DCT) (Pang et al., 2019), then the information in different frequency domains is extracted, different frequency domains store different information. The image information is divided into high frequency information and low frequency information, where the high frequency information stores the contour information and detail information, and the low frequency information stores the texture information. Finally, the spatial information and frequency information are used to enhance the objective feature for improving the classification accuracy. Our main contributions are threefold:

- The frequency domain enhancement module is proposed to extract high and low frequency information from the feature maps and transform this information from the frequency domain into a spatial domain image. This transformation does not lose the original information, but increases the number of feature maps, allowing the network to focus on both contour and texture information.
- A novel clothing classification network is proposed to improve the accuracy with frequency information and optimal backbone network, that is, frequecy-spatial feature enhancement network for clothing classfication (FFENet). Our proposed optimal backbone network consists of effective convolutional modules and efficient channel attention (ECA) (Wang et al., 2020) modules. A large number of experiments indicate that our proposed method can achieve the best performance among state of the art methods.
- By collecting some public complex scene clothing images on kaggle websites and shopping websites, combining with a small part of clothing data in the deepfashion dataset (Liu et al., 2016), and manually filtering the collected images, we obtain a dataset of 8 classified clothing styles with 5156 high quality images.

RELATED WORK

The related work consists of two main parts: 1) application of frequency domain in the field of image classification, and 2) mainstream deep neural networks for classification.

Application of Frequency Domain in the Field of Image Classification

Spatial domain images can be classified directly by using trained neural networks, and good classification results can be obtained, but this approach does not fully exploit the information in the image, which is the frequency domain information implies in the image. There are many ways to extract frequency domain information from an image, such as the Fourier transform, discrete cosine transform, wavelet transform, and other methods. Frequency domain information, as an alternative representation of the spatial domain image, may contain information that is not used by the neural network and is useful for classification. Researchers have also conducted research into the use of frequency domain information extracted from images to complement image processing tasks when using deep learning techniques.

First, for DCT, Qin et al. (2021) studies the effect of partially compressed input images using DCT algorithm on the performance of neural networks. DCT algorithm is used to reduce some data redundancy in the network, but there is also a risk of reducing valuable features for network learning. Xu et al. (2020) studies the DCT transformation of the original image and then the use of CNN for classification, and proved through experiments that the DCT features obtained directly from the JPG format can be processed as effectively as the original image data using the same CNN architecture. The neural network architecture with DCT features performs as well as the original image data. Borhanuddin et al. (2019) from a different perspective, this paper rethinks the channel attention mechanism from the perspective of frequency analysis, proves that the regular global average pooling is a special case of frequency domain feature decomposition, and proposes a novel multi-spectral channel attention structure. Liu et al. (2018) proposes a family of methods to compress and accelerate neural network training in the frequency domain by focusing on all weights and their underlying connections. The paper also explores a data-driven approach to remove redundancy in the spatial and frequency domains, which enables the network to discard more useless weights by maintaining similar accuracy. After obtaining the optimal sparse CNN in the frequency domain, they reduce the computational burden of the convolution operation in the CNN by linearly combining the convolutional responses of the DCT basis. Gueguen et al. (2018) proposes and explores a simple idea where they directly used JPG image processing to generate DCT coefficients and modified the Resnet50 (He et al., 2016) network to accommodate DCT coefficients directly as input, evaluated the performance of this model on the ImageNet dataset.

In addition to DCT, as a powerful time evaluation analysis method, wavelet transform can also provide additional frequency domain information for deep learning techniques. Li et al. (2020) uses nonlinear model and average pooling to wavelet transform and proposes wavelet scattering network. The first network layer of this network outputs SIFT-type descriptors, while the next layer provides complementary translation invariant information to improve classification. The network computes a translation invariant image representation that is stable to deformation and preserves high-frequency information for classification. However, the network cannot be easily transferred to other tasks due to strict mathematical assumptions. In order to solve the problem that CNN is prone to noise interruption (that is, small image noise will cause drastic changes in the output), Bruna and Mallat (2013) use discrete wavelet transform to replace max-pooling, step convolution, and average pooling to enhance CNN, and proposes a universal discrete wavelet transform and its inverse transform layer suitable for all kinds of wavelets. And these layers are used to design a wavelet ensemble CNN for image classification.

In addition to the wavelet transform, a number of variants based on the wavelet transform (e.g. the contour wavelet transform) have also been studied accordingly. Liu et al. (2020) proposes a new network architecture called contourlet convolutional neural network, which is designed to learn sparse and effective feature representations of images. The contour wave transform is first applied to obtain spectral features from the image, then the spatial spectral feature fusion method is used to integrate the spectral features into the CNN architecture, followed by statistical feature fusion to integrate the statistical features into the network, and finally the fused features are classified to obtain the results.

Mainstream Deep Neural Networks for Classification

The image classification task is the task of determining which categories in the category space the input image belongs to. There are two types of mainstream classification networks, one based on convolutional neural networks and the other based on Transformer. Each of these two types has its own advantages and disadvantages. Convolutional neural network-based models work better for both small sample datasets and large datasets, and the network inference is faster. The Transformer-based classification model may not work as well on small datasets as the convolutional neural network-based one, which consumes a lot of memory space, but it also performs well on very large datasets and the network is less prone to overfitting.

Convolutional neural network based classification models include GoogleNet (Szegedy et al., 2015), ResNet, DenseNet (Huang et al., 2017), EfficientNet (Tan and Le, 2019), ConvNext (Liu et al., 2022), and EfficientNetV2 (Tan and Le, 2021), among which the EfficientNetV2 model has best accuracy and computing speed compared with many classification networks. The focus of this paper is to improve the performance of the network on a small sample dataset, so in this paper, CNN is used to build our clothing classification network. The transformer (Dong et al., 2022; Hua et al., 2022) based classification models, such as ViT and SwinT (Liu et al., 2021), are initially used in the field of natural processing, where the transformer framework based on the attention mechanism achieved good results. Later, (Vaswani et al., 2017) introduces the transformer to the field of computer vision, which worked well in mega databases, so the improvement of the Transformer based classification models hung a boom.

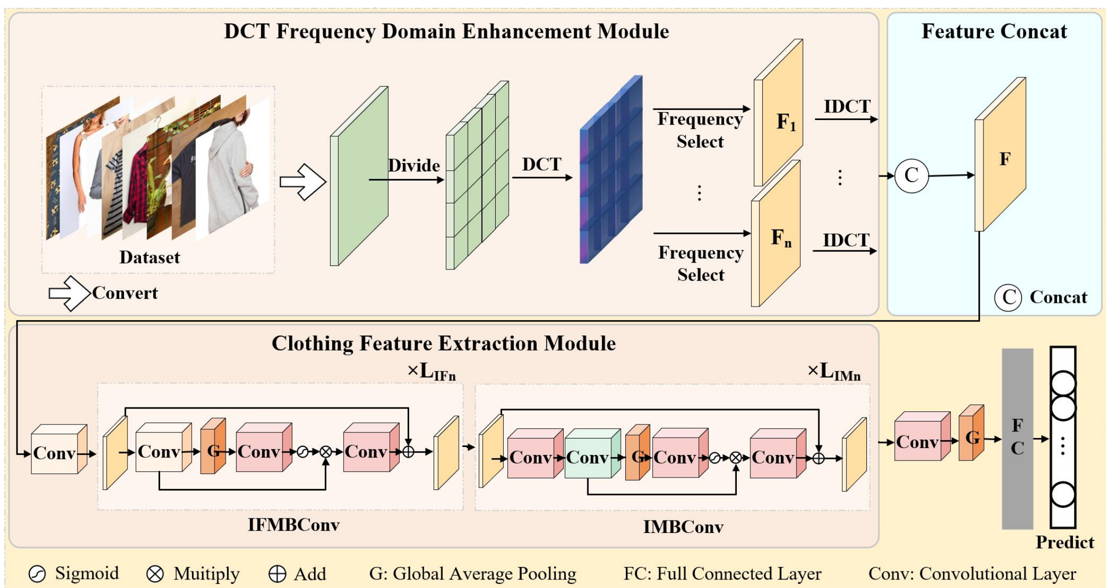

Figure 1. An overview of the proposed network. We use the DCT frequency domain enhancement module to extract the spatial information of different frequency bands of the image, and use the stitching operation to concatenate all the spatial information feature maps to obtain F. The information of the feature map F is further extracted using the clothing feature extraction module, where L_{IFN} and L_{IMN} represent the number of repetitions of the corresponding layer of the IFMBConv block and IMBConv block, respectively. Finally, one 1×1 convolution operation, one global average pooling operation and two full connection operations were carried out in turn to obtain the final clothing classification results.

OUR METHODS

The main task of the approach in this paper is to construct a model for clothing classification on a small sample clothing dataset for complex scenes. The category of clothing depends greatly on the silhouette features and textural characteristics of the clothing. We have summarised these rules, and if we can extract and learn these corresponding features through some techniques, it will be of great help in clothing classification.

Based on the above discussion we propose our approach (FFENet), where from the perspective of frequency domain, texture information and contour information in spatial domain are the information of different frequency bands. So we convert the spatial domain images into frequency domain images, transform them into different spatial feature maps by selecting information from different frequency bands, and put the spatial feature map information into the network we build for learning to improve the accuracy of clothing classification.

Network Overview

Our proposed network structure is shown in Figure 1. When the clothing images are input to the DCT frequency domain enhancement (DCT-FDE) module, the image will be converted into ycbcr format, and then the converted feature map will be divided into blocks, and then the information of each block will be converted from the spatial domain to the frequency domain using DCT, and then the list of spatial domain feature maps will be generated according to the frequency domain information at the corresponding position of each block. Finally, the generated feature maps are stitched to obtain the feature map F. DCT-FDE module obtained our initially feature map information, in order to learn the feature map information more deeply we propose the clothing feature extraction module for further learning of the information in the feature map F. In this feature extraction model, a 3×3 convolution operation is performed first,then the modified fused MBConv block and the modified MBConv block. Finally, we put the feature map output from the clothing feature extraction module into the classification header for classification, first for 1×1 convolutional collation of the channels then for global average pooling. Finally, two fully connected layers are used, the first fully connected layer (FC) is used to obtain the preliminary sequence, and the second FC is used to obtain the final prediction result.

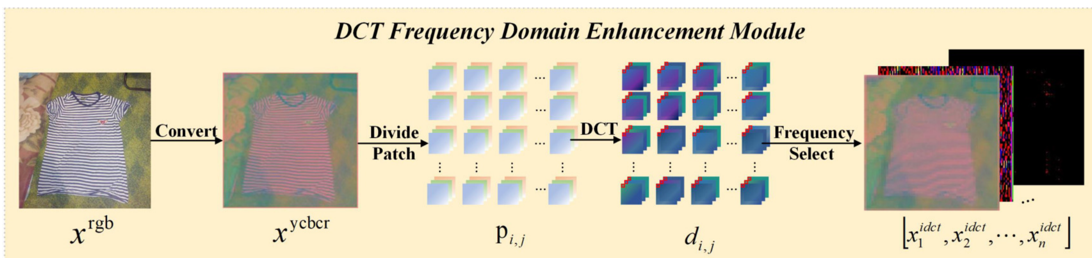

Figure 2. Processing flow of the DCT frequency domain enhencement module.

DCT Frequency Domain Enhancement Module

The transformation process of the DCT frequency domain enhancement module is depicted in Figure 2, which first converts the input RGB image into ycbcr format to obtain the feature map $x_{ycbcr} \in R^{H \times W \times 3}$. Subsequently, x_{ycbcr} is partitioned into a set of 4×4 patches to obtain $\{p_{i,j} \in R^{4 \times 4 \times 3} \mid 1 \leq i \leq H//4, 1 \leq j \leq W//4\}$, where the patches are three channels. A dense DCT transformation is performed on the image window for each one, and each patch is processed in the frequency domain to obtain, where represents the patch corresponding to a particular colour channel in $\{d_{i,j} \in R^{4 \times 4 \times 3} \mid 1 \leq i \leq H//4, 1 \leq j \leq W//4\}$. Here each value in the patch corresponds to the intensity of a particular frequency band. In order to extract the information of different frequency bands separately, we filter the frequency bands for the number of times of chunk size squared, taking the information of only one frequency band and filtering out the information of other frequency bands each time, and perform a DCT inverse transform to convert the filtered time-frequency domain information into spatial domain information after each filtering operation, and finally get a list of feature maps $x_1^{idct} \in R^{H \times W \times 3}, x_2^{idct} \in R^{H \times W \times 3}, \ldots, x_n^{idct} \in R^{H \times W \times 3}$, where the value of n is the square of the block size. We have a block size of 4 here, so we end up with 16 feature maps. As Figure 3 shows an experiment we did, visualising the 16 feature maps obtained after inputting the image to the DCT-FDE module, we can see that the first band of the chunk stores the most colour and texture information, and that the other bands store more shape and detail information, which is what we call low-frequency information and high-frequency information. By this method we do not lose any information in any of the frequency bands, but it allows our subsequent proposed classification network to learn both high-frequency and low-frequency information, which in fact replaces the convolution operation in a sense and has a feature enhancement effect.

Clothing Feature Extraction Module

The MBConv block is a portable module proposed by MobileNetV2 (Sandler et al., 2018), and later Efficientnet is built on top of the MBConv block, and they both achieved good results. Then later EfficientNetV2 suggests that using fused MBConv blocks at the shallow end of the network had better results through neural architecture search techniques. The module we built is initially a combination of

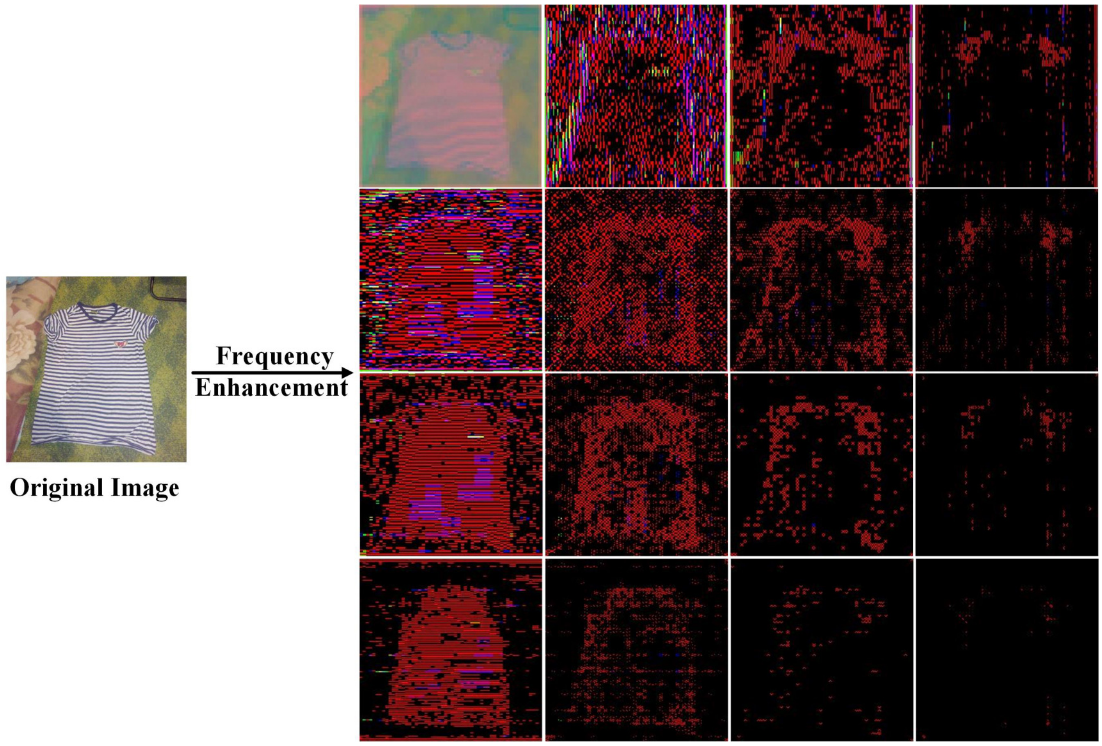

Figure 3. The image after 4×4 DCT transformation and IDCT transformation of a single frequency band.

the fused MBConv block and the MBConv block, but in combination with the previous DCT-FDE module, we guess that there is channel compression in the squeeze-and-excitation (SE) module, which might lead to inadequate learning of our frequency domain information, so our DCT-FDE module improves the structure of the fused MBConv block and the MBConv block by replacing the SE module with the ECA module. Our experimental results prove our conjecture, please see the experimental section for details. We use ECA module to enforce the feature map, which is defined as follows:

$$F_t^{eca} = F_{t-1}(\sigma(Conv^{1\times1}(GAP(F_{t-1})))) \quad (1)$$

where F_{t-1} and F_t^{eca} are the input and output feature maps of the ECA module, respectively. σ denotes the sigmoid function. $Conv^{1\times1}$ denotes the convolution operateion with a filter size of 1×1. GAP denotes for global average pooling operateion. As shown in Figure 4(a) and Figure 4(b), the improved fused

Stage	Operator	Input Size	Stride	Channel	Layers
1	Conv3 × 3	224×224	2	48	4
2	IFMBConv1	112×112	1	48	7
3	IFMBConv2	112×112	2	48	7
4	IFMBConv2	56×56	2	64	10
5	IMBConv1	28×28	2	96	19
6	IMBConv2	14×14	1	192	25
7	IMBConv1	14×14	2	224	7

Table 1. The structure of the CFEM.

MBconv block (IFMBConv) and the improved MBConv block (IMBConv) used in this paper are shown schematically. Module specific parameter information can be found in Table 1, which describes each phase of the DCT-FDE module in detail. The parameter Stride indicates whether the first convolution in the first IFMBConv block or IMBConv block in a phase consisting of IFMBConv or IMBConv compresses the feature map, when the step size is 1, the feature map size is unchanged, and when the step size is 2, the compressed feature map size is one half of the input feature map. The parameter Channel indicates

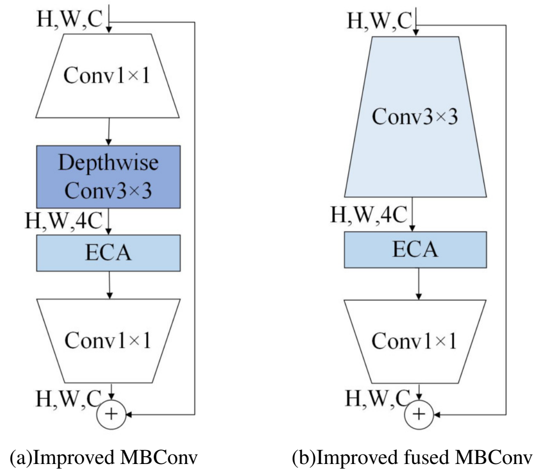

Figure 4. The structure of the improved MBConv and the improved fused MBConv.

Figure 5. The clothing image styles in the clothing8 dataset.

the size of the feature map at the time of input to the current stage. The Layers parameter represents the number of times the IFMBConv block or IMBConv block is repeated. Note that the size of the feature map output at stage 7 is 7×7 and the number of channels is 384.

EXPERIMENTS

In this section, our method FFENet will conduct comparative experiments on two datasets, clothing 8 and fashion-mnist, and verify the rationality of the structural design by conducting ablation experiments on the clothing 8 dataset. Clothing 8 is our own small sample dataset built for complex scenes, while fashion-mnist is a public clothing dataset. The performance of our model in specific scenarios is verified on the clothing 8 dataset, and the experiments on the fashion-mnist dataset are used to verify the performance of our model on regular large datasets.

Implement Details

System configuration information for the experimental platform. The system version is Windows 10, the processor is an Intel(R) Core(TM) i9-12900KF CPU @ 3.20GHz and the GPU is an NVIDIA GeForce RTX 3090 Ti 24GB. Conda environment relies on python 3.8. The optimizer used is the SGD optimizer, which the initial learning rate is 0.01 and the decay coefficient of the optimizer is 0.0001. The input

Model	Precision(%)↑								mPrecision(%)↑
	Dress	Jacket	Pant	Polo	Shirt	T-shirt	Tank Top	Warmcloth	
GoogleNet	73.33	92.42	97.75	87.32	90.62	92.68	91.94	79.38	86.33
Resnet-101	82.67	95.52	82.80	82.3	80.6	82.7	91.90	90.4	89.74
DenseNet-201	78.57	95.71	98.75	78.21	84.72	82.93	81.82	82.02	85.34
EfficientNet-B7	78.87	95.38	94.38	78.31	85.51	**87.95**	89.23	89.02	87.33
ViT-L	25.74	44.58	49.21	22.45	36.36	36.21	36.21	40.62	34.86
Swin-L	**95.65**	89.33	95.79	83.82	78.33	77.92	86.05	81.17	86.01
ConvNext-L	75.00	**100.00**	96.51	81.82	92.42	82.95	88.06	88.10	88.11
EfficientNetV2-L	78.26	96.97	**100.00**	82.28	94.03	86.36	**95.16**	88.64	90.21
FFENet(ours)	91.07	**100.00**	97.98	**91.55**	**96.67**	82.67	91.67	**96.20**	**93.53**
Model	Recall(%)↑								mRecall(%)↑
	Dress	Jacket	Pant	Polo	Shirt	T-shirt	Tank Top	Warmcloth	
GoogleNet	79.71	88.41	92.13	94.20	83.82	86.21	80.00	81.40	85.74
Resnet-101	89.86	92.75	97.75	89.86	85.29	87.36	81.43	89.53	89.23
DenseNet-201	79.71	97.10	88.76	88.41	81.71	78.16	77.14	84.88	85.48
EfficientNet-B7	81.16	89.86	94.38	94.20	86.76	83.91	82.86	84.88	87.25
ViT-L	37.68	53.62	69.66	15.94	52.94	24.14	20.0	15.12	36.14
Swin-L	83.02	95.71	91.92	79.17	74.60	83.33	92.5	83.13	85.42
ConvNext-L	82.61	89.86	93.26	91.30	89.71	83.91	84.29	87.06	87.75
EfficientNetV2-L	78.26	92.75	**97.75**	94.20	92.65	87.36	84.29	91.76	89.88
FFENet(ours)	**86.96**	**95.65**	**97.75**	**97.10**	**94.12**	**90.80**	**91.43**	**92.94**	**93.34**
Model	Accuracy(%)↑								mAccuracy(%)↑
	Dress	Jacket	Pant	Polo	Shirt	T-shirt	Tank Top	Warmcloth	
GoogleNet	94.40	97.86	98.68	96.87	97.53	95.39	96.71	94.23	96.46
Resnet-101	96.71	95.68	99.34	97.36	97.36	**97.20**	97.03	95.22	97.36
DenseNet-201	95.22	99.18	98.19	95.88	97.03	94.56	95.39	95.22	96.33
EfficientNet-B7	95.39	98.35	98.35	96.38	96.87	96.05	96.87	96.38	96.83
ViT-L	80.56	87.15	85.01	84.18	84.35	83.03	83.36	84.84	84.84
Swin-L	80.14	88.14	87.97	85.61	85.10	85.10	86.96	84.93	85.49
ConvNext-L	94.88	98.84	98.51	96.70	98.02	95.21	96.86	96.53	96.94
EfficientNetV2-L	95.05	98.84	**99.67**	97.03	98.51	96.20	97.69	97.19	97.52
FFENet(ours)	**97.17**	**99.34**	**99.67**	**97.85**	**99.01**	97.19	**98.68**	**97.85**	**98.35**

Table 2. Comparison of classification performance on the clothing 8 validation set. Results that surpass all competing methods are bold font. The upward arrow next to the parameter in the table indicates that the larger the parameter, the better.

network has 224×224 image pixels, batch size is 8. GoogleNet, ResNet, DenseNet, EfficientNet, ConvNext, EfficientNetV2, ViT and SwinT were chosen to compare the classification of the models.

Dataset

Clothing 8.

Clothing 8 dataset is a dataset consisting of images related to clothe-ware. It is a dataset assembled in this paper by collecting open source datasets from the kaggle website and some images from deepfashion, and finally combining them with a series of data we crawled on the web ourselves. The clothing 8 dataset has training set of 4550 examples and a val set of 606 examples. Each example is associated with a label from 8 classes. The 8 categories are skirt, jacket, pants, polo, shirt, tank top, t-shirt and warmcloth, see Figure 5 for details.

Fashion-mnist.

Fashion-mnist dataset is a dataset consisting of images related to clothe-ware, shoes, and bag. The fashion-mnist dataset has a training set of 60000 examples and a test set of 10000 examples. Each example is a 28×28 gray-scale image associated with a label from 10 classes. The 10 categories are Angle Boot, Bag, Coat, Dress, Pullover, Sandal, Shirt, Sneaker, Trouser and T-shirt.

Evaluation Criterion

We usually call the prediction is correct and positive as true positive (TP). A false positive (FP) is a prediction that is false and positive. If the prediction os correct and the result is negative, it is called true negative (TN). A false negative (FN) is when the prediction is false and negative. Based on the above theories, the evaluation indexes of the text are as follows. Accuracy is the proportion of correct prediction results in total prediction, which specific calculation method is shown in Equation (2). Precision is the

Model	Model Accuracy(%)↑
GoogleNet	85.83
Resnet-101	89.46
DenseNet-201	85.33
EfficientNet-B7	87.31
ViT-L	36.24
SwinT-L	74.1
ConvNext-L	87.79
EfficientNetV2-L	90.01
FFENet(ours)	**93.4**

Table 3. Comparison of classification performance on the clothing 8 validation set. Results that surpass all competing methods are bold font. The upward arrow next to the parameter in the table indicates that the larger the parameter, the better.

percentage of positive predictions that are correct, which specific calculation method is shown in Equation (3). Recall is the percentage of all positive events that correctly predicted the result, which specific calculation method is shown in Equation (4). Model accuracy is equal to the number of correct predictions (NCP) of all kinds divided by the total number of verified pictures (TNVP), which specific calculation method is shown in Equation (5). It is the parameter most often used to evaluate the quality of model training. The equations are as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$Model \quad Accuracy = \frac{NCP}{TNVP} \tag{5}$$

where accuracy here is calculating the probability that a single category is correct, whereas model accuracy is calculating the probability that the entire model is correct. In addition, we include the number of model parameters and model complexity as indicators to analyze our model in the ablation experiment.

Comparison Evaluation on Clothing8 Dataset

The number of training rounds for our experiments rounds on the clothing 8 dataset is 100. The robustness of the model is very important, so we verify the performance of our model on the clothing 8 validation set. The three indicators compared in Table 2 are Precision, Recall and Accuracy, and it can be seen that the effect of our model is better than these models. In this complex scene small sample data set, it can be seen that the classification model based on convolution is better than the classification model based on tranformer. Firstly, in terms of the prediction mPrecision metric, our model outperforms the best EfficientNetV2 by 3.32%. Secondly, from the average recall, EfficientNetV2 has the best effect, but our model is 3.46% better than ConvNext. Finally, our model is also the best in terms of mAccuracy metric, our mAccuracy is 98.35%, which is 0.82% better than the best existing model EfficientNetV2 in the table.

Table 3 shows the comparison results of Model Accuracy between our model and other existing models on the clothing 8 validate seen from Table 3 that ResNet, ConvNeXt, and EfficientNetV2 achieve preferably performance in the existing methods, and their accuracy rates are 90.01%, 89.46% and 87.79%, respectively. However, our model EfficientNetV2 achieves an accuracy of 93.4%, 3.39% better than the best model EfficientNetV2. The performance of our model on this small sample dataset of complex scenes is impressive.

Model	Model Accuracy(%)↑	mAccuracy(%)↑	mRecall(%)↑	mPrecision(%)↑
GoogleNet	88.18	97.63	88.18	88.34
ResNet	90.00	98.00	90.00	90.05
DenseNet	91.11	98.22	91.11	91.15
EfficientNet	93.87	98.78	93.89	93.88
ViT	86.70	97.33	86.66	86.64
SwinT	90.08	98.07	90.35	90.33
ConvNext	93.86	98.75	93.76	93.73
EfficientNetV2	93.93	98.79	93.93	93.98
FFENet(ours)	**94.62**	**98.92**	**94.62**	**94.62**

Table 4. Comparison of classification performance on the fashion-mnist test set. Results that surpass all competing methods are bold font. The upward arrow next to the parameter in the table indicates that the larger the parameter, the better

DCT-FDE Module	ACs	ECA	Model Accuracy(%)↑	Params(M)↓
			90.01	117.24
✓			90.76	117.26
✓	✓		91.91	117.36
✓	✓	✓	**93.40**	**95.16**

Table 5. Comparison of classification performance on the clothing 8 dataset. ACs stands for channel information for adjusting the constructed network. Results that surpass all competing methods are bold font. The upward arrow next to the parameter in the table indicates that the larger the parameter, the better. The downward arrow next to the parameter in the table indicates that the smaller the parameter, the better

Comparison Evaluation on Fashion-mnist Dataset

The number of training rounds for our conduct experiments on the fashion-mnist dataset is 50. Our model is compared with some existing classification models for the same volume on the fashion-mnist test set, and these experiments are not pre-trained. Among them, ViT and SwinT adopt the largest size model, and ConvNeXt adopts base model, largest model used by EfficientNet and EfficientNetV2. Transformer works well on very large datasets, but the transformer model do not work well on fashion-mnist dataset, so we used the largest volume of Transformer classification model to compare with convolutional classification model.

As can be seen from Table 4, in terms of model accuracy, our model outperforms the best model EfficientNetV2-L by 0.69%. According to the mAccuracy metric, our model is 0.13% better than the best model ConvNext. Then, it also has good effects from the indicators of mRecall and mPrecision. Its mRecall and mPrecision are both 94.62%, 0.69% and 0.64% higher than the current best algorithm respectively. From these comparative experiments, we can see that our model works well even on datasets with simple backgrounds.

Ablation study

Our ablation experiments are conducted on the clothing 8 dataset, and the number of training rounds is set to 100. The other experimental settings are the same as those in subsection . We choose MBConv block and fusion MBConv block to build the network structure, and the specific information can be found in subsection . Table 5 shows our improvement process. We add DCT-FDE module and can see a 0.75% improvement in accuracy. We think about why the accuracy is not improved a lot. Considering that the number of channels of our DCT-FDE module output feature map is 48, we adjust the number of input and output channels of the first and second stages of the network to be 48, so that the information of the feature map output by our DCT-FDE module can be fully learned. The number of channels in the previous first and second stage are less than 48, so the learned feature map information must be insufficient. By adjusting the number of channels in the network we get another 1.15% improvement. Because the SE module inside the MBConv block and fused MBConv block has the operation of compression channel, so we replace the SE block with the ECA module. ECA module is also a channel attention mechanism, but the ECA module has no operation to compress the channel. Experiments show that the model accuracy is improved by 1.49% after replacing the SE module with the ECA module. And the final model parameters decreased by 22.08M compared with the initial model.

Table 6 shows our other ablation experiment, in which the influence of DCT block size on the accuracy

DCT Block Size	Model Accuracy(%)↑	Params(M)↓	GFLOPs↓
2 ×2	90.586	**117.25**	**12.33**
4 ×4	**90.760**	117.26	12.46
8 ×8	90.759	117.30	12.98

Table 6. Comparison of classification performance on the clothing 8 dataset. Results that surpass all competing methods are bold font. The upward arrow next to the parameter in the table indicates that the larger the parameter, the better. The downward arrow next to the parameter in the table indicates that the smaller the parameter, the better.

of the final model is discussed. According to the data in the table, we can find that the accuracy rate of 2×2 block is lower than that of 4×4 block. Meanwhile, compared with 2×2 block, the number of parameters and computational complexity in 4×4 block are not much improved. From the data in the table, if the 8×8 DCT block size is used, the accuracy is also decreased a little compared with the 4×4 block size, and the number of parameters and the computational complexity are increased, so we choose 4×4 block as our block size.

CONCLUTIONS

In this work, we mainly study how to improve the performance of clothing classification in complex scenes. Since the clothing classification largely depends on the clothing texture information and contour information. Different frequency bands in the frequency domain store the image texture, contour and other information respectively that have confirmed by previous studies and our experiments. If this information can be extracted and learned through this feature in the frequency domain, it is likely to improve the performance of clothing classification. Therefore, we propose a discrete cosine transform feature extraction module combined with a fully convolutional backbone algorithm, which is a clothing classification network based on frequency-spatial feature enhancement network. Our proposed algorithm can effectively improve the accuracy of image classification. Finally, we conduct extensive experiments on two datasets: clothing 8 and fashion-mnist. The experiments show that the network we constructed has excellent performance both on our clothing dataset of complex scenes and regular clothing dataset.

Our method also has limitations, such as for simple background clothing images, our boost is not particularly significant. These issues will be addressed in our future work.

ACKNOWLEDGEMENT

This work was supported by national natural science foundation of China (No. 62202346), Hubei key research and development program (No.2021BAA042), open project of engineering research center of Hubei province for clothing information (No. 2022HBCI01), Wuhan applied basic frontier research project (No. 2022013988065212), MIIT's AI Industry Innovation Task unveils flagship projects (Key technologies, equipment, and systems for flexible customized and intelligent manufacturing in the clothing industry), and Hubei science and technology project of safe production special fund (Scene control platform based on proprioception information computing of artificial intelligence).

REFERENCES

Al Shehri, W. (2022). Alzheimer's disease diagnosis and classification using deep learning techniques. *PeerJ Computer Science*, 8:e1177. doi: 10.7717/peerj-cs.1177.

Bai, X., Xue, R., Wang, L., and Zhou, F. (2019). Sequence sar image classification based on bidirectional convolution-recurrent network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9223–9235. doi: 10.1109/TGRS.2019.2925636.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359.

Borhanuddin, B., Jamil, N., Chen, S., Baharuddin, M., Tan, K., and Ooi, T. (2019). Small-scale deep network for dct-based images classification. In *2019 4th International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–6. IEEE.

Bruna, J. and Mallat, S. (2013). Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886.

Cheung, W. and Hamarneh, G. (2009). *n*-sift: *n*-dimensional scale invariant feature transform. *IEEE Transactions on Image Processing*, 18(9):2012–2021. doi: 10.1109/TIP.2009.2024578.
Déniz, O., Bueno, G., Salido, J., and De la Torre, F. (2011). Face recognition using histograms of oriented gradients. *Pattern recognition letters*, 32(12):1598–1603.
Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., and Guo, B. (2022). Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134. IEEE. doi: 10.1109/CVPR52688.2022.01181.
Gueguen, L., Sergeev, A., Kadlec, B., Liu, R., and Yosinski, J. (2018). Faster neural networks straight from jpeg. *Advances in Neural Information Processing Systems*, 31.
Hassan, M. R., Huda, S., Hassan, M. M., Abawajy, J., Alsanad, A., and Fortino, G. (2022). Early detection of cardiovascular autonomic neuropathy: A multi-class classification model based on feature selection and deep learning feature fusion. *Information Fusion*, 77:70–80. doi: 10.1016/j.inffus.2021.07.010.
He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. IEEE. doi: 10.1109/CVPR.2016.90.
Hua, W., Dai, Z., Liu, H., and Le, Q. (2022). Transformer quality in linear time. In *International Conference on Machine Learning*, pages 9099–9117. PMLR.
Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708. IEEE. doi: 10.1109/CVPR.2017.243.
Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J. (2021). 1d convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 151:107398.
Li, Q., Shen, L., Guo, S., and Lai, Z. (2020). Wavelet integrated cnns for noise-robust image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7245–7254. IEEE. doi: 10.1109/CVPR42600.2020.00727.
Liu, M., Jiao, L., Liu, X., Li, L., Liu, F., and Yang, S. (2020). C-cnn: Contourlet convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2636–2649. doi: 10.1109/TNNLS.2020.3007412.
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022. IEEE. doi: 10.1109/ICCV48922.2021.00986.
Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104. IEEE. doi: 10.1109/CVPR.2016.124.
Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986. IEEE. doi: 10.1109/CVPR52688.2022.01167.
Liu, Z., Xu, J., Peng, X., and Xiong, R. (2018). Frequency-domain dynamic pruning for convolutional neural networks. *Advances in neural information processing systems*, 31.
Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567.
Ölçer, N., Ölçer, D., and Sümer, E. (2023). Roof type classification with innovative machine learning approaches. *PeerJ Computer Science*. doi: 10.7717/peerj-cs.1217.
Pan, S., Gupta, T. K., and Raza, K. (2023). Batts: a hybrid method for optimizing deep feedforward neural network. *PeerJ Computer Science*, 9:e1194. doi: 10.7717/peerj-cs.1194.
Pang, C.-Y., Zhou, R.-G., Hu, B.-Q., Hu, W., and El-Rafei, A. (2019). Signal and image compression using quantum discrete cosine transform. *Information Sciences*, 473:121–141.
Qin, Z., Zhang, P., Wu, F., and Li, X. (2021). Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792.
Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520. IEEE. doi: 10.1109/CVPR.2018.00474.
Shajini, M. and Ramanan, A. (2022). A knowledge-sharing semi-supervised approach for fashion clothes classification and attribute prediction. *The Visual Computer*, 38(11):3551–3561.
Sun, L., Zhao, G., Zheng, Y., and Wu, Z. (2022). Spectral–spatial feature tokenization transformer for

hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14. doi: 10.1109/TGRS.2022.3144158.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). Supplementary material for 'eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA*, pages 13–19. IEEE. doi: 10.1109/CVPR42600.2020.01155.

Xu, K., Qin, M., Sun, F., Wang, Y., Chen, Y.-K., and Ren, F. (2020). Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1740–1749. IEEE. doi: 10.1109/CVPR42600.2020.00181.

Zhang, J., Liu, L., Huang, D., Fu, X., and Huang, Q. (2016). Clothing co-segmentation based on hog features and e-svm classifier. In *2016 6th International Conference on Digital Home (ICDH)*, pages 16–19. IEEE. doi: 10.1109/ICDH.2016.013.

Zhou, Z.-H. (2022). Open-environment machine learning. *National Science Review*, 9(8):nwac123. doi:10.1093/nsr/nwac123.