Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

FSPDF: Few-shot learning with progressive dual-domain feature fusion via self-supervised learning

Dongqing Li, Jie Jin, Linhua Zou, Hong Zhao *

School of Computer Science, Minnan Normal University, Zhangzhou, Fujian, 363000, China
The Key Laboratory of Data Science and Intelligence Application, Minnan Normal University, Zhangzhou, Fujian, 363000, China

ARTICLE INFO

Keywords:
Few-shot learning
Feature fusion
Discrete wavelet transform
Self-supervised learning

ABSTRACT

Few-shot learning (FSL) aims to develop models that can generalise to new tasks using only a few labelled examples. Recently, feature fusion methods have achieved great success in FSL by aggregating features from different sources. However, these methods generally rely on efficient feature extractors to capture the intrinsic patterns of the data and focus solely on single-domain spatial features. In this paper, we propose a progressive dual-domain feature fusion strategy via self-supervised learning (FSPDF). FSPDF consists of dual-domain feature learning (DFL) and dual-domain feature fusion (DFF) and leverages progressive dual-domain fusion to address the issue of insufficient feature representation. The DFL strategy obtains an efficient feature extractor through a dual-domain self-supervised module, optimising generalisation to new data and fully extracting dual-domain features. The DFF strategy leverages the previously obtained feature extractor and incorporates discrete wavelet transform to enrich and fuse dual-domain features, providing a more diverse feature representation. The effectiveness of FSPDF is demonstrated across three commonly used benchmark FSL datasets. For instance, in the 5-way 1-shot setting, FSPDF improves performance by 3.47% on CUB-200-2011 and 1.05% on miniImageNet. Code is available at https://github.com/fhqxa/FSPDF.

1. Introduction

Few-shot learning (FSL) aims to create models capable of generalising to new tasks using only a limited number of annotated examples, thereby addressing the challenge of scarce data in machine learning [1, 2]. Traditional algorithms often demand substantial labelled data to train on new tasks, a requirement that is both expensive and time-intensive to fulfil [3]. This issue is particularly pronounced in fields such as biology and medicine. In contrast, FSL adapts rapidly to new tasks by fine-tuning pretrained models on small labelled datasets. FSL has achieved significant advancements in areas such as medicine [4,5], natural language processing [6,7], and fault diagnosis [8–10].

Current FSL methods primarily consist of meta-learning and feature fusion techniques. Meta-learning focuses on acquiring generalised knowledge from a variety of tasks to improve performance on new challenges [11]. These approaches can be categorised into optimisation-based methods, metric learning-based methods, and data augmentation-based methods [12]. Optimisation-based methods emphasise setting the initial state of the model to enable rapid adaptation to new tasks with minimal iterations. Various strategies have been developed to learn initialisation techniques [2] and parameter adjustment mechanisms [13], facilitating swift adaptation to novel challenges. Metric learning-based methods focus on constructing a feature space that facilitates classification using distance metrics [1,14]. This approach enables the model to perform effectively in FSL scenarios without requiring extensive fine-tuning. In contrast, data augmentation-based methods [15] aim to enhance the generalisation capability of the model in few-shot settings by increasing the diversity of the training data. Although meta-learning approaches have achieved significant breakthroughs in the FSL field, they face limitations in addressing sample feature diversity.

Different from meta-learning approaches, feature fusion methods enhance a generalisation of the model to new tasks by integrating multi-source feature information [16]. These methods are categorised into inter-layer-based and attention mechanism-based feature fusion techniques [17]. Inter-layer-based feature fusion, a fundamental aspect of modern network architectures, leverages interactions between layers [18]. For example, Ding et al. [19] and Piao et al. [20] concentrated on extracting and fusing features through innovative network structures to improve the capacity of the model to capture multi-scale information. In contrast, attention mechanism-based feature fusion methods

* Corresponding author at: School of Computer Science, Minnan Normal University, Zhangzhou, Fujian, 363000, China.
E-mail addresses: g2023061041@stu.mnnu.edu.cn (D. Li), g2022061016@stu.mnnu.edu.cn (J. Jin), g2023061023@stu.mnnu.edu.cn (L. Zou), zh1127@mnnu.edu.cn (H. Zhao).

https://doi.org/10.1016/j.knosys.2025.113389
Received 19 September 2024; Received in revised form 18 February 2025; Accepted 18 March 2025
Available online 2 April 2025

improve model performance by selectively integrating the most relevant features from various layers or sources. These methods enable the model to flexibly identify and prioritise the most important features for the specific task at hand [21,22].

The feature fusion methods described above have proven effective and have achieved initial success in FSL. However, these approaches overlook two key limitations: (1) They rely on large labelled datasets to train feature extractors. When training data is scarce, the generalisation ability of the feature extractor diminishes, making it difficult to capture critical features in the data. (2) They focus exclusively on single spatial domain features, neglecting the frequency domain information in images. Addressing these limitations presents opportunities to enhance model performance when processing diverse datasets.

To address these issues, we propose a two-stage progressive dual-domain feature fusion strategy via self-supervised learning (FSPDF). FSPDF incorporates a two-stage framework consisting of dual-domain feature learning (DFL) and dual-domain feature fusion (DFF). First, to overcome the limitations identified, we design a DFL strategy that employs a dual-domain self-supervised module to predict rotation angles. This approach enables the model to learn robust feature extractors from unlabelled data across both spatial and frequency domains. This phase minimises the dependency on large quantities of labelled data and improves the ability of the model to generalise to new datasets. To address the second issue, we propose the DFF, which utilises the feature extractors trained in the first stage and applies wavelet transforms to deeply extract frequency domain features. This approach captures features from both spatial and frequency domains and performs their fusion. By independently extracting and fusing features from these dual domains, the model provides a more diverse and comprehensive feature representation for FSL tasks.

We present experimental results on three widely used datasets to demonstrate the effectiveness of FSPDF. In the 5-way 1-shot setting, FSPDF surpasses the baseline model by at least 1.05% on miniImageNet, 0.53% on CIFAR-FS, and 3.47% on CUB-200-2011. The primary contributions of our work are as follows:

- From a data perspective, we introduce a dual-domain self-supervised module. By predicting rotation angles, the model learns feature extractors for both spatial and frequency domains using unlabelled data. This approach allows the model to fully leverage the dual-domain information present in the data.
- From a feature perspective, we propose a dual-domain feature fusion strategy. By integrating the feature extractors from the previous stage with wavelet transforms, we extract frequency domain features. This approach captures and fuses features from both spatial and frequency domains, enhancing the optimisation of feature representation.
- We validate the performance of FSPDF through a series of experiments on three commonly used FSL datasets. The results demonstrate that our approach outperforms many existing FSL methods.

The remainder of this paper is organised as follows: Section 2 offers a brief review of previous work on FSL. Section 3 describes the FSPDF model in detail. Section 4 presents the experimental results and analysis across three datasets. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Related work

In this section, we review the relevant literature on FSL based on meta-learning, feature fusion, and self-supervised learning.

2.1. *Few-shot learning based on meta-learning*

Meta-learning-based FSL aims to adapt to new tasks by leveraging knowledge from prior tasks, addressing scenarios with limited data [1,2]. Based on meta-learning principles, existing methods can be classified into three categories: optimisation-based, metric learning-based, and data augmentation-based approaches [12].

Optimisation-based meta-learning approaches focus on achieving effective initialisation by training on diverse tasks, facilitating rapid adaptation to new tasks [23]. Finn et al. [2] proposed a meta-learning algorithm tailored for rapid adaptation to new tasks, and it required only a few samples and iterations. In contrast, Qin et al. [24] introduced an innovative meta-learning method that segments input images into multiple blocks, focuses on the locations of category-specific objects, and significantly reduces the number of model parameters. Building on this research direction, Rusu et al. [13] proposed learning an optimised initialisation of model parameters, enabling adaptation to new tasks with just a few gradient steps.

Metric learning-based meta-learning methods leverage similarities between images to construct an embedding space that clusters samples of the same class while separating those of different classes. Vinyals et al. [14] introduced matching networks, which utilise attention mechanisms and memory to evaluate support and query samples within the embedding space. Snell et al. [1] designed a new embedding representation that clusters feature points of each class around a separate prototype representation. From the attention perspective, Kang et al. [25] developed a relational embedding method that employs attention mechanisms to assess the relationships between support and query samples.

Data augmentation-based meta-learning methods enhance the existing few labelled samples by applying transformations and expansions, increasing the training data available to the model [26]. Traditional data augmentation techniques typically involve operations such as translation, flipping, cropping, scaling, and rotation of images [27,28]. However, the aforementioned methods are typically tailored to specific datasets and are difficult to be applied across other datasets. To address this limitation, Hariharan et al. [29] tackled the challenge of scarce data for new tasks by generating additional synthetic data, referred to as “hallucinated” data. To enhance data augmentation within a meta-learning framework, Ni et al. [30] proposed a task-specific data augmentation method. This approach transfers samples from the base dataset to new tasks and processes them using an encoder–decoder paradigm, ensuring that the resulting features remain semantically meaningful. Inspired by the above work, Cho et al. [31] developed an adaptive data augmentation network that autonomously selects the most effective augmentation techniques to optimise performance in FSL.

These methods significantly improve performance in FSL by optimising various stages of the training process. In contrast to the aforementioned approaches, our work emphasises innovation at the feature level. Specifically, we focus on fusing dual-domain features to achieve a more diverse and expressive feature representation.

2.2. *Few-shot learning based on feature fusion*

Feature fusion-based FSL enhances model generalisation by integrating information from multiple feature domains. These approaches can be categorised into inter-layer fusion and attention mechanism-based fusion methods.

Inter-layer fusion methods allow the model to process data at varying granularity levels by leveraging detailed information from shallow layers and abstract representations from deeper layers. Ding et al. [19] proposed a feature fusion framework that employs a relational network to integrate inter-layer features, extracting multi-scale image features. Similarly, Piao et al. [20] developed a module that uses residual connections to merge complementary multi-layer data from both depth and RGB sources. Building on the concept of multi-layer and multi-scale feature fusion, Sun et al. [32] proposed a network that preserves high-resolution representations throughout its architecture by parallelly connecting and fusing inter-layer features spanning high to low resolutions.

Attention mechanism-based feature fusion methods improve model performance by selectively integrating the most relevant features from various layers or sources. These strategies are inspired by the human visual attention process. Lin et al. [21] leveraged the inherent multi-scale and hierarchical features of convolutional neural networks (CNNs) to approximate an image pyramid, thereby achieving high-resolution semantic feature representation. Similarly, Sinha et al. [22] developed a module that processes multi-scale features and their concatenations to generate multi-scale attention maps, ensuring a consistent scale for feature context aggregation. Building on these concepts, Shermin et al. [33] introduced a two-stage attention mechanism that refines cross-layer features through an iterative exchange of information, identifying local visual features.

Inter-layer fusion methods integrate deep abstract features with shallow detailed features, while attention-based methods intelligently select and combine key features. Although these approaches excel in specific areas, they are limited to single-domain features and fail to leverage the benefits of multi-domain integration. To bridge this gap, we expand feature representation by fusing information from both frequency and spatial domains.

2.3. Self-supervised learning

Self-supervised learning improves the learning capabilities of the model by designing pretext tasks that do not require additional labelled samples during training. These pretext tasks encourage the model to develop a deeper understanding of the data, driving it to learn useful feature representations while naturally solving these tasks, which subsequently benefits downstream applications [34]. Existing self-supervised learning approaches can be broadly categorised into contrastive and prediction-based methods [35].

Contrastive-based self-supervised learning captures the intrinsic structure of data by comparing pairs of samples. Chen et al. [36] introduced a classic framework that enhances visual representation using contrastive loss, aiming to maintain consistency across different augmented views of the same image. Instead of relying on positive and negative sample pairs for comparison, Grill et al. [37] introduced an innovative network that iteratively predicts the representation of one augmented view from another without the need for negative pairs. To enhance the efficiency of self-supervised learning, Tong et al. [38] utilised a joint embedding formulation, achieving a significant reduction in training epochs by two orders of magnitude.

Prediction-based self-supervised learning enhances representations by training the model to predict specific attributes or future states of the input data [39]. Drawing inspiration from the spatial context of images, Doersch et al. [40] proposed a method for inferring the relative positions of patches within an image, thereby enabling the model to learn valuable visual features. Rather than predicting masked images, Wei et al. [41] proposed a masked feature prediction approach, in which the model is trained by obscuring parts of the input image and tasked with predicting the features of the masked sections. Shifting the focus from spatial arrangements to colourisation, Zhang et al. [42] introduced a method for predicting the colours of grayscale images. Building on the rotation invariance property of image data, Gidaris et al. [39] developed a task involving the prediction of image rotation angles. Building on this foundation, Mangla et al. [35] introduced manifold mixup in downstream tasks to achieve smoother and more generalisable representations. Assran et al. [43] developed a joint embedding architecture for self-supervised learning, which predicts relationships between different views of an image.

Building on previous self-supervised learning tasks, we designed a dual-domain self-supervised module to more effectively extract and fuse dual-domain features for downstream tasks. Unlike the aforementioned methods, our approach introduces an additional branch to process frequency domain information, enabling deeper learning of dual-domain features. Additionally, we integrate a dual-domain fusion strategy in downstream tasks, leveraging the complementary characteristics of spatial and frequency domain features to achieve a more diverse and robust feature representation.

3. The FSPDF model

In this section, we first present a brief overview of the entire model. Then, we provide a detailed explanation of the two-stage module of the model. Finally, we perform model analysis and learning.

3.1. Framework overview

From both data and feature perspectives, we propose a two-stage progressive feature fusion strategy. The basic structure of FSPDF is illustrated in Fig. 1. FSPDF comprises two main stages: DFL and DFF.

(1) Dual-domain feature learning (DFL): From the data perspective, we design a dual-domain self-supervised module to train more robust feature extractors. This module independently predicts rotation angles for numerous unlabelled samples in both spatial and frequency domains.

(2) Dual-domain feature fusion (DFF): From the feature perspective, we perform dual-domain feature extraction and fusion using the feature extractors trained in the previous stage. Wavelet transform is employed to extract low-frequency domain features, while traditional CNNs are used to capture spatial domain features. Finally, the fusion of these dual-domain features creates a more diverse and comprehensive representation.

3.2. Dual-domain feature learning (DFL)

In this stage, we employ the rotation self-supervised task as a pretext task for FSPDF. The objective is to enhance the performance of the dual-domain fusion strategy in data-scarce scenarios from two perspectives: rotational invariance and visual perceptibility. The theoretical foundation of rotational invariance encourages the model to prioritise high-level semantic information, such as shapes, edges, and structures. This property ensures that the model remains attuned to the core features of the image, regardless of variations in rotation angles. In contrast to invariance, visual perceptibility focuses on associating rotation angles with the primary semantics of the image. This approach enables the model to capture the global structure of the image while reinforcing the correlation between the dual-domain features.

The auxiliary objective of FSPDF is to estimate the applied rotation angle by analysing the transformations in the input image after rotation at various angles. We train the dual-domain feature extractors using an unlabelled dataset. During the spatial domain self-supervised learning process, we employ a standardised approach to process spatial domain images, converting them into a three-channel RGB colour space. Conversely, during the frequency domain self-supervised learning process, we first convert the images into the YUV colour space. We then apply 4:2:2 sub-sampling to the YUV-formatted images, splitting them into 24 channels.

The dataset settings for the self-supervised auxiliary stage are as follows. A dataset comprising samples and corresponding labels composes the dataset of a self-supervised pretext task. The sample set $D = \{(x_i, y_i)|i = 1, \ldots, m\}$, where x_i is the ith sample, y_i is the label of sample x_i, and m is the corresponding sample numbers. In the context of our study, the r is the identity of the random rotation image angle. The set C^r represents the complete collection of 0°, 90°, 180°, and 270° rotation angles, where $C^r = \{r_j, j = 1, \ldots, p\}$, and r_j is the jth rotational angle.

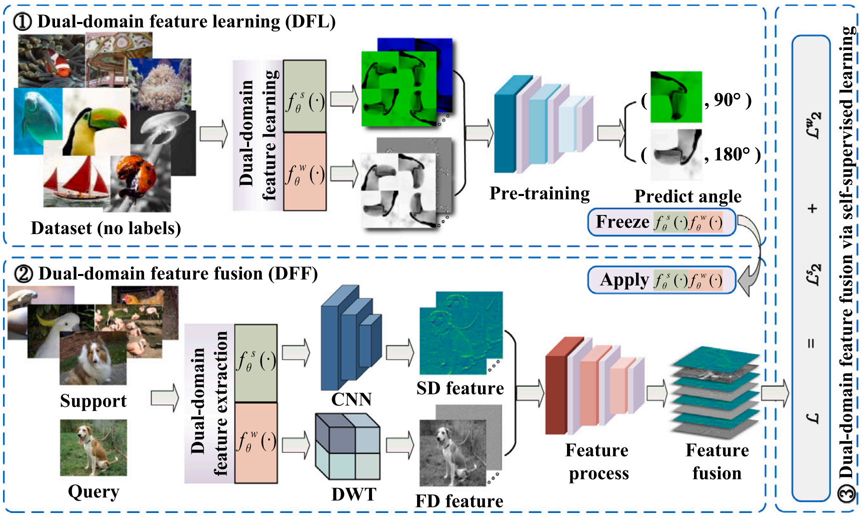

① Dual-domain feature learning (DFL)
Dataset (no labels)
Dual-domain feature learning
$f^s_\theta(\cdot)$
$f^w_\theta(\cdot)$
Pre-training
(, 90°)
(, 180°)
Predict angle
Freeze $f^s_\theta(\cdot) f^w_\theta(\cdot)$
② Dual-domain feature fusion (DFF)
Apply $f^s_\theta(\cdot) f^w_\theta(\cdot)$
Support
Query
Dual-domain feature extraction
$f^s_\theta(\cdot)$
$f^w_\theta(\cdot)$
CNN
DWT
SD feature
FD feature
Feature process
Feature fusion
$\mathcal{L} = \mathcal{L}^s_2 + \mathcal{L}^w_2$
③ Dual-domain feature fusion via self-supervised learning

Fig. 1. Overview of FSPDF. $f^w_\theta(\cdot)$ and $f^s_\theta(\cdot)$ represent the feature extractors in the frequency domain and spatial domain, respectively. $\mathcal{L}^w_2$ and $\mathcal{L}^s_2$ denote the losses in the frequency and spatial domains during the DFF stage, respectively, while $\mathcal{L}$ is the total loss.

The set D^r represents D after being subjected to random rotations at four distinct angles. Each rotated sample x^r_i in the set D^r corresponds to the angle of rotation r, and the angle of rotation r is the label of the rotated sample x^r_i in this stage of the task. This methodology allows for systematically exploring the invariance of the dataset under rotation.

The loss of DFL phase on the spatial domain $\mathcal{L}^s_{rot}$ is as follows:

$$\mathcal{L}^s_{rot} = \frac{1}{|C^r|}\sum_{i=1}^{m}\sum_{j=1}^{p}\mathcal{L}(K_c\left(f^s_\theta\left(x^r_i\right)\right), r_j), \tag{1}$$

where $|\cdot|$ represents the cardinality of a set, $\mathcal{L}(\cdot,\cdot)$ is the standard cross entropy loss, $K_c(\cdot)$ is a cosine classifier, and the spatial domain feature extractor $f^s_\theta(\cdot)$ adopts a conventional CNN image classification network structure.

The loss of DFL phase on the frequency domain $\mathcal{L}^w_{rot}$ is as follows:

$$\mathcal{L}^w_{rot} = \frac{1}{|C^r|}\sum_{i=1}^{m}\sum_{j=1}^{p}\mathcal{L}(K_c\left(f^w_\theta\left(x^r_i\right)\right), r_j), \tag{2}$$

where the frequency domain feature extractor $f^w_\theta(\cdot)$ processes the frequency information of the image by the two-dimensional haar discrete wavelet transform (DWT).

The loss function $\mathcal{L}_c$ for predicting the true category of rotated images takes the following form:

$$\mathcal{L}_c = \mathcal{L}\left(x^r_i, y_i\right). \tag{3}$$

In this image classification scenario, we combine the auxiliary loss function for predicting rotation angles with the standard classification loss function. The parameters $\mathcal{L}^s_{rot}$, $\mathcal{L}^w_{rot}$, and $\mathcal{L}_c$ are controlled using the parameter λ to obtain the corresponding self-supervised losses for the spatial and frequency domains.

The total loss functions $\mathcal{L}^s_1$ and $\mathcal{L}^w_1$ in the DFL stage are represented as follows:

$$\mathcal{L}^s_1 = \lambda\mathcal{L}^s_{rot} + (1-\lambda)\mathcal{L}_c, \tag{4}$$

$$\mathcal{L}^w_1 = \lambda\mathcal{L}^w_{rot} + (1-\lambda)\mathcal{L}_c, \tag{5}$$

where the parameter λ modifies the contribution of self-supervised loss to classification loss.

3.3. Dual-domain feature fusion (DFF)

In this phase, the dual-domain fusion task serves as the core component of FSPDF. The goal is to address the FSL challenge by fully leveraging the complementary information within the image. The theoretical foundation of information complementarity allows the model to focus on local details in the spatial domain (such as shapes and edges) while capturing global patterns in the frequency domain (such as textures and frequency distributions). Spatial and frequency domain features inherently exhibit a complementary relationship, offering multi-level perceptions of the data from different perspectives. By fusing these dual-domain features, the model harnesses their respective strengths, resulting in a more comprehensive and enriched feature representation.

We establish the FSL problem using the N-way K-shot training and test scenarios. We divide the dataset into base classes and novel classes. The training process for FSL models unfolds in two main stages. The first phase involves training a network to utilise data from the base sample set $D_b = \{(x_i, y_i)\}_{i=1}^{m_b}$, where m_b is the corresponding base sample numbers. The second phase includes modifying the network for the novel sample set $D_n = \{(x_i, y_i)\}_{i=1}^{m_n}$, where m_n is the corresponding base sample numbers. FSL algorithms aim to cultivate a rich set of feature representations leveraging the extensive labelled data from the base classes D_b. These learned representations are easily transferable and adaptable to novel classes D_n with minimal labelled samples.

By simulating the working principle of the human visual system, we perform multi-layer convolution on the input image and obtain final spatial domain features. The loss function $\mathcal{L}^s_e$ of the spatial domain feature extractor $f^s_\theta(\cdot)$ is represented as follows:

$$\mathcal{L}^s_e = \mathcal{L}\left(K_c(f^s_\theta(x_i)), y_i\right). \tag{6}$$

The spatial domain total loss function $\mathcal{L}^s_2$ for the DFF stage is denoted as:

$$\mathcal{L}^s_2 = \gamma\mathcal{L}^s_e + (1-\gamma)\mathcal{L}^s_1, \tag{7}$$

where the parameter γ is designed to modulate the impact of the DFL strategy on leveraging spatial domain features for classification predictions.

Fig. 2 illustrates how a two-dimensional wavelet transform decomposes an image with a resolution of $H \times W$ resolution. In this process, $L(\cdot)$ and $H(\cdot)$ represent the low-pass and high-pass filters, respectively, which extract low-frequency and high-frequency information from the image. The downsampling operation $\downarrow 2$ is then applied to both the low-frequency and high-frequency components. First, a one-dimensional filter is applied to decompose the original image horizontally, resulting in low-frequency and high-frequency components. Then, the output from this step is further decomposed in the vertical direction,

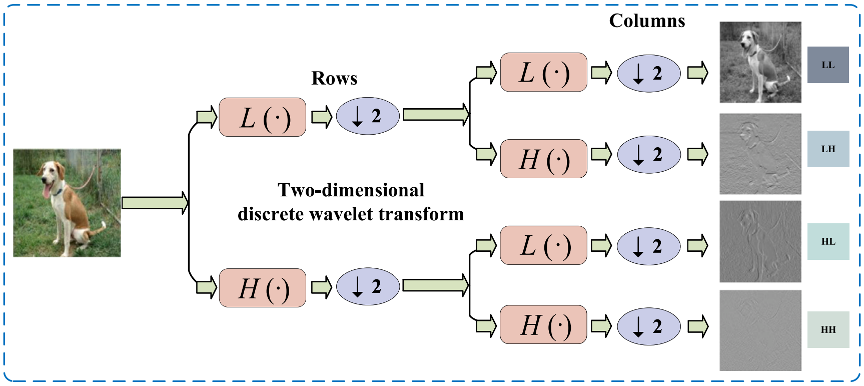


Fig. 2. The process of two-dimensional wavelet decomposition.

yielding four components: LL, LH, HL, and HH, which represent varying combinations of low- and high-frequency information. After discrete wavelet transform (DWT) frequency domain processing, the input image resolution is reduced to $\frac{H}{2} \times \frac{W}{2}$.

To simplify image processing, we use different convolution kernels instead of $L(\cdot)$ and $H(\cdot)$ filters to perform the DWT operation on the image. We utilise four distinct convolutional kernels k_{LL}, k_{LH}, k_{HL}, and k_{HH} to extract the corresponding low-frequency and high-frequency components by the two-dimensional haar wavelet definition. These kernels are delineated as follows:

$$k_{LL} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, k_{LH} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, k_{HL} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, k_{HH} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{8}$$

Using k_{LL} to extract the low-frequency component, the FSPDF model captures the overall structure of the image, providing a coarse representation. The remaining kernels k_{LH}, k_{HL}, and k_{HH} focus on extracting the high-frequency components, which are critical for identifying intricate textures and edges.

By giving the input image x_i and moving the convolution kernel on the image x_i, we obtain four frequency domain component bands x_{LL}, x_{LH}, x_{HL}, and x_{HH}. The representations of low-frequency bands are as follows:

$$x_{LL} = k_{LL} \otimes x_i, \tag{9}$$

where "⊗" represents convolution operations. In addition, The representations for high-frequency bands are detailed as follows:

$$x_{LH} = k_{LH} \otimes x_i, \quad x_{HL} = k_{HL} \otimes x_i, \quad x_{HH} = k_{HH} \otimes x_i. \tag{10}$$

We propose the following frequency feature extraction process. First, during DWT, we select the approximate components from the low-frequency x_{LL} to represent the frequency domain features. These components are utilised as the frequency domain feature representation, while the remaining high-frequency components are discarded.

Fig. 3 illustrates the impact of different frequency domain components on visual recognition. Through DWT, the input image is divided into four components: one low-frequency and three high-frequency parts. Despite the low-frequency component occupying only one-quarter of the image resolution, it still retains the main features of the image, making it easily recognisable. In contrast, the high-frequency components, which occupy three-quarters of the resolution, make it challenging to discern the primary object. The high-frequency components primarily contain edges and textures, which have a limited impact on image recognition. In contrast, the low-frequency component retains the main contours and global information of the image. Additionally, its lower resolution reduces computational costs, enhancing processing efficiency.

Subsequently, the input image size is assumed to be $H \times W$ and the resolution size is $\frac{H}{2} \times \frac{W}{2}$ after DWT frequency domain processing. The frequency domain feature extractor $f_\theta^w(\cdot)$ input size is $\frac{H}{2} \times \frac{W}{2}$. The loss function $\mathcal{L}_e^w$ of the frequency domain feature extractor $f_\theta^w(\cdot)$ is represented as follows:

$$\mathcal{L}_e^w = \mathcal{L}\left(W_c(f_\theta^w(x_{LL})), y_i\right). \tag{11}$$

The frequency domain total loss function $\mathcal{L}_2^w$ for the DFF stage is denoted as:

$$\mathcal{L}_2^w = \gamma \mathcal{L}_e^w + (1-\gamma)\mathcal{L}_1^w, \tag{12}$$

where the parameter γ adjusts the influence of the self-supervised approach in utilising frequency domain features for making classification forecasts.

The final loss function can be expressed as follows:

$$\mathcal{L} = \eta \mathcal{L}_2^s + (1-\eta)\mathcal{L}_2^w, \tag{13}$$

where the parameter η controls the influence of the spatial domain feature and the frequency domain feature for the dual-domain fusion features.

Traditional inter-layer fusion methods are ineffective for dual-domain features, as forced fusion between network layers often results in a decline in classification accuracy. To address this issue, we employ a direct feature fusion approach, combining the normalised spatial and frequency domain features extracted by the network using a concatenation operation. During the fusion process, we first create a new array to store the fused dual-domain features. The spatial and frequency domain features are then sequentially input into this array, ensuring that both sets of features are stored within the same data structure, thereby forming a comprehensive joint feature representation.

3.4. Model analysis and learning

The FSPDF method incorporates self-supervised auxiliary tasks to facilitate DFF. Algorithm 1 provides the pseudocode for the model training process. First, during the DFL stage, the self-supervised rotation task, outlined in lines 3 and 4, randomly rotates the image by 0°, 90°, 180°, 270° and predicts the corresponding rotation angle. The model leverages rotational invariance to learn more diverse image features, resulting in more robust feature extractors for the subsequent stage. Second, during the DFF stage, features are extracted from both the spatial and frequency domains, as described in lines 9 and 10. In line 11, DFF is performed to generate richer feature representations for classification tasks. The loss function guides the iterative optimisation of model parameters, with their final adjustments made in line 12.

Compared with conventional FSL approaches, FSPDF exhibits the following characteristics. First, FSPDF utilises a dual-domain self-supervised module to predict image rotation angles, which aids in exploring data characteristics across both frequency and spatial domains. This module produces more effective dual-domain feature extractors

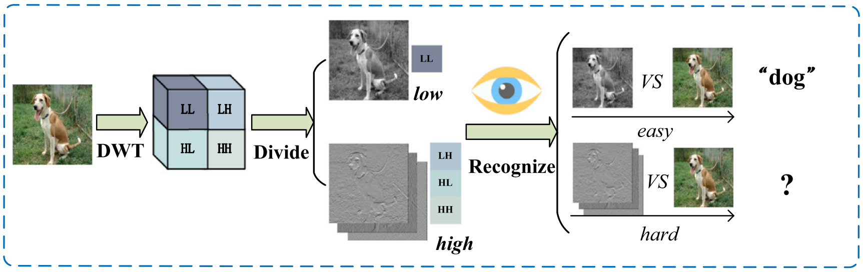


Fig. 3. Image decomposition and feature recognition process based on wavelet transform.

Algorithm 1 FSPDF.

```
Input: The base training set $D_b$ in each episode, with $N$ classes and $K$ support examples for each class.
Output: The best parameter set $\theta$ of FSPDF model.
Initialise feature extractor $f_\theta^s(\cdot)$ and $f_\theta^w(\cdot)$;
//Dual-domain feature learning stage
for $epoch \leq K_1$ do
    Predict image rotation angle by Eq. (1) and Eq. (2);
    Predict image category after rotation by Eq. (3);
    Calculate the loss of DFL stage by Eq. (4) and Eq. (5);
    Update and freeze feature extractors $f_\theta^s(\cdot)$ and $f_\theta^w(\cdot)$;
end for
//Dual-domain feature fusion stage
for $K_1 < epochs \leq K_2$ do
    Extract spatial domain feature by Eq. (6);
    Extract frequency domain feature by Eq. (11);
    Dual-domain features fusion by Eq. (13);
    Update parameter set $\theta$ of model by $\mathcal{L}$;
end for
return $\theta$.
```

for the subsequent stage. Second, unlike methods that rely on single-domain feature representations, FSPDF combines feature extractors trained and optimised in the previous stage with DWT technology to extract and fuse dual-domain features. This approach enhances the diversity of feature representation.

4. Experimental results and analysis

In this section, we provide a detailed description of the experimental setup and result analysis for the FSPDF algorithm: (1) datasets and experimental setup; (2) performance comparison with related methods; (3) ablation experiment; (4) analysis of low-frequency and high-frequency regions in DWT; (5) analysis of different wavelet basis selections in DWT; (6) parameter sensitive analysis; and (7) visualisation of the model.

4.1. Datasets and experimental setup

We evaluate our model on three datasets: miniImageNet, CUB-200-2011, and CIFAR-FS. (1) miniImageNet is a widely used benchmark for FSL, containing 100 categories with 600 images per category, each at a resolution of 84 × 84. In our experiments, we allocate 64 classes for training, 16 for validation, and 20 for testing. (2) CUB-200-2011 is focused on fine-grained bird species classification, containing 11,788 images across 200 species. The dataset is split into 130 training categories, 20 validation categories, and 50 test categories. (3) CIFAR-FS, derived from CIFAR-100, includes 100 categories. We randomly divide these into 64 for training, 16 for validation, and 20 for testing, with all images resized to 32 × 32.

Table 1
Comparison of accuracy (%) with different FSL approaches in 5-way 1-shot and 5-shot experimental settings on CUB-200-2011.

Model	Backbone	5-way 1-shot	5-way 5-shot
MatchingNet [14]	ConvNet	61.16 ± 0.89	72.86 ± 0.70
ProtoNet [1]	ConvNet	66.36 ± 1.00	82.03 ± 0.59
MAML [2]	ConvNet	66.26 ± 1.05	78.82 ± 0.70
RelationNet [44]	ConvNet	64.38 ± 0.94	80.16 ± 0.64
RISE [45]	ConvNet	67.42 ± 0.96	82.86 ± 0.59
CSTS [46]	ConvNet	60.83 ± 0.45	77.12 ± 0.44
DeepEMD [47]	ResNet-12	79.27 ± 0.29	89.80 ± 0.51
AFHN [48]	ResNet-18	70.53 ± 1.01	83.95 ± 0.63
Exemplar [49]	WRN-28-10	71.58 ± 0.32	84.63 ± 0.57
Baseline++ [50]	WRN-28-10	70.40 ± 0.81	82.92 ± 0.78
Rotation [39]	WRN-28-10	77.61 ± 0.86	89.32 ± 0.46
S2M2$_R$ [35]	WRN-28-10	76.12 ± 0.87	88.62 ± 0.48
FSPDF (ours)	WRN-28-10	$\mathbf{79.59 \pm 0.77}$	$\mathbf{90.91 \pm 0.79}$

We employ a two-stage training strategy, setting $\lambda = 0.5$, $\gamma = 0.5$ and $\eta = 0.5$. The self-supervised phase is trained for 400 epochs, followed by 400 episodes in the meta-training stage, with $K_1 = 400$ and $K_2 = 800$. All experiments are conducted using the PyTorch framework on an NVIDIA GeForce RTX 2080 Ti and a 3.60 GHz Intel CPU.

4.2. Performance comparison with related methods

We evaluate the performance of FSPDF by comparing it with various models across three commonly used datasets. S2M2$_R$ [35], which serves as the foundational model and is most similar to our approach, is included in the comparison. The best results are highlighted in bold.

Regarding computational complexity, our model utilises a standard convolutional network architecture, which is also employed in many existing models such as [3,23,24]. Based on [35], the primary focus of our work is on performing inference and training the model architecture using the proposed dual-domain fusion strategy, without incorporating additional networks. The time required for wavelet image processing, when calculated using a formula, is almost negligible. Therefore, our model shares the same algorithmic complexity as existing models. Moreover, this allows us to transfer our model to other frameworks, such as [8] and [9], as our main contributions lie in wavelet image processing and DFF.

Table 1 presents the results of our experiments on CUB-200-2011. By comparing FSPDF with alternative FSL models, we make the following observations:

(1) Table 1 shows that FSPDF achieves significant improvements on CUB-200-2011, increasing accuracy by 3.47% in the 1-shot setting and 2.29% in the 5-shot setting compared to the baseline. This performance boost underscores the advantages of our DFF approach.

(2) Spatial domain features capture local image details, such as the geometric shapes of bird beaks and feather textures. In contrast, DWT-based frequency domain analysis extracts global structural features, like the overall shape and posture of birds. From the dual-domain fusion perspective, combining these two types of features enhances the feature

Table 2
Comparison of accuracy (%) with different FSL approaches in 5-way 1-shot and 5-shot experimental settings on CIFAR-FS.

Model	Backbone	5-way 1-shot	5-way 5-shot
MatchingNet [14]	ConvNet	51.32 ± 0.85	68.93 ± 0.74
ProtoNet [1]	ConvNet	55.50 ± 0.70	72.00 ± 0.60
MAML [2]	ConvNet	58.90 ± 1.90	71.50 ± 1.00
RelationNet [44]	ConvNet	55.00 ± 1.00	72.00 ± 0.60
RISE [45]	ConvNet	60.05 ± 0.96	78.44 ± 0.64
CSTS [46]	ConvNet	62.47 ± 0.47	81.12 ± 0.42
DeepEMD [47]	ResNet-12	46.47 ± 0.78	63.22 ± 0.71
AFHN [48]	ResNet-18	68.32 ± 0.93	81.45 ± 0.87
Exemplar [49]	WRN-28-10	70.05 ± 0.17	84.01 ± 0.22
Baseline++ [50]	WRN-28-10	67.50 ± 0.64	80.08 ± 0.32
Rotation [39]	WRN-28-10	70.66 ± 0.20	84.15 ± 0.14
$S2M2_R$ [35]	WRN-28-10	73.09 ± 0.19	85.55 ± 0.13
FSPDF (ours)	WRN-28-10	**73.56 ± 0.20**	**86.84 ± 0.17**

Table 3
Comparison of accuracy (%) with different FSL approaches in 5-way 1-shot and 5-shot experimental settings on miniImageNet.

Model	Backbone	5-way 1-shot	5-way 5-shot
MatchingNet [14]	ConvNet	43.56 ± 0.84	55.31 ± 0.73
ProtoNet [1]	ConvNet	50.37 ± 0.83	67.33 ± 0.67
MAML [2]	ConvNet	50.96 ± 0.92	66.09 ± 0.71
RelationNet [44]	ConvNet	51.84 ± 0.88	64.55 ± 0.70
RISE [45]	ConvNet	53.22 ± 0.81	69.41 ± 0.67
CSTS [46]	ConvNet	62.38 ± 0.48	79.77 ± 0.44
DeepEMD [47]	ResNet-12	**65.91 ± 0.82**	82.41 ± 0.56
AFHN [48]	ResNet-18	62.38 ± 0.72	78.16 ± 0.56
Exemplar [49]	WRN-28-10	62.20 ± 0.45	78.80 ± 0.15
Baseline++ [50]	WRN-28-10	57.53 ± 0.10	72.99 ± 0.43
Rotation [39]	WRN-28-10	63.90 ± 0.18	81.03 ± 0.11
$S2M2_R$ [35]	WRN-28-10	64.88 ± 0.18	83.18 ± 0.11
FSPDF (ours)	WRN-28-10	65.82 ± 0.19	**84.29 ± 0.87**

representation and improves the ability of the model to detect subtle differences in fine-grained datasets.

(3) Compared with the classic Exemplar algorithm, FSPDF improves classification accuracy by 8.01% in 1-shot tasks and 6.28% in 5-shot tasks. From the perspective of DFL, the pretraining strategy enhances the sensitivity of the model to subtle image variations. This approach strengthens the ability of the model to capture fine image details and provides more stable and reliable feature representations for DFF in downstream tasks.

Our experimental results on CIFAR-FS are shown in Table 2.

We reach the following conclusions:

(1) Table 2 shows the classification accuracy of the FSPDF model on CIFAR-FS, with improvements of 0.47% in the 1-shot setting and 1.29% in the 5-shot setting over the baseline.

(2) In the 1-shot task, we observe a modest accuracy improvement of 0.47%. This slight improvement is attributed to the low image resolution of the CIFAR-FS dataset (32 × 32 pixels), which limits the ability of FSPDF to capture fine texture and edge details. Additionally, the mixing of foreground and background in the images complicates feature extraction, making it more challenging for FSPDF to learn effective features.

(3) Although RISE and CSTS are tailored for specific tasks, FSPDF still achieves higher accuracy. This suggests that FSPDF may have broader applicability across different few-shot learning scenarios.

The performance on miniImageNet is summarised in Table 3.

The following observations are made from the evaluation of FSPDF alongside other FSL models:

(1) Table 3 shows the performance of FSPDF on miniImageNet. In the 1-shot setting, FSPDF achieves a 0.94% improvement in classification accuracy. In the 5-shot setting, the improvement increases to 1.11%. The higher image resolution of the miniImageNet dataset allows FSPDF to capture more detailed and comprehensive features, enhancing performance through the DFF strategy.

Table 4
Ablation experimental results on miniImageNet, CUB-200-2011, and CIFAR-FS. SD and FD, respectively, represent the spatial domain and frequency domain.

SD	FD	CUB-200-2011		CIFAR-FS		miniImageNet	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
–	–	76.12	88.62	73.09	85.55	64.88	83.18
✓	–	69.52	81.93	72.79	85.48	61.44	83.59
–	✓	77.33	89.24	70.04	85.46	63.74	82.19
✓	✓	**79.59**	**90.91**	**73.56**	**86.84**	**65.82**	**84.29**

(2) The 5-way 1-shot classification accuracy of the DeepEMD model on this dataset is higher than our classification accuracy. This model introduces an FSL algorithm that identifies the best match between image regions. Despite the use of a more advanced backbone network, the improvement in accuracy is minimal. This situation occurs because we do not explicitly account for the correspondence between images, whereas DeepEMD addresses this issue by comparing feature matrices across different images to establish relationships.

4.3. Ablation experiment

We conduct ablation experiments to evaluate the effectiveness of spatial domain, frequency domain, and dual-domain fused features for classification in the FSPDF model. Table 4 presents the results of these ablation experiments, highlighting the impact of different modules on performance in 5-way 1-shot and 5-shot settings across the miniImageNet, CUB-200-2011, and CIFAR-FS datasets.

These results lead to the following conclusions:

(1) Spatial domain features are essential for the FSPDF model, as they capture and preserve the overall structure and layout of the image. Understanding spatial information helps the model interpret the relative positions of elements within the image, allowing it to accurately recognise and comprehend complex spatial relationships between objects.

(2) Features derived from the frequency domain are crucial for the FSPDF model. Through wavelet transformation, these features decompose images into blocks that range from low to high frequencies. This decomposition enables FSPDF to perform multi-scale analysis and process rich image information across different scales.

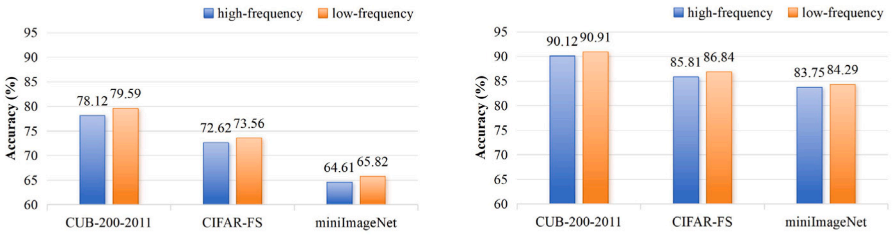
(3) Our experimental results demonstrate optimal performance with dual-domain fusion. FSPDF outperforms the baseline on CUB-200-2011, CIFAR-FS, and miniImageNet. Specifically, in the 1-shot learning scenario, FSPDF achieves improvements of 3.47% on CUB-200-2011, 0.47% on CIFAR-FS, and 1.05% on miniImageNet. These improvements increase to 2.29% for CUB-200-2011, 1.29% for CIFAR-FS, and 1.11% for miniImageNet in the 5-shot learning scenario. This demonstrates that the dual-domain fusion strategy leverages the complementary characteristics of the spatial and frequency domains, enhancing classification performance in FSL scenarios.

Overall, the ablation studies emphasise the unique and complementary roles of spatial and frequency domain features in improving classification performance. Furthermore, the dual-domain fusion strategy in FSPDF enhances their strengths while mitigating their weaknesses.

4.4. Analysis of low-frequency and high-frequency regions in DWT

We perform an ablation study to evaluate the effects of low-frequency and high-frequency regions in DWT on classification performance. Fig. 4 illustrates the results for 5-way 1-shot and 5-way 5-shot settings on miniImageNet, CUB-200-2011, and CIFAR-FS datasets, leading to the following conclusions:

(1) The DWT categorises features into low-frequency approximation components and high-frequency detail components (horizontal, vertical, and diagonal). As illustrated in Figs. 4(a) and 4(b), low-frequency



(a) 5-way 1-shot

(b) 5-way 5-shot

Fig. 4. Evaluation of different regions in DWT on CUB-200-2011, miniImageNet, and CIFAR-FS datasets.

Table 5
The ablation experiment results in different wavelets based on miniImageNet, CUB-200-2011, and CIFAR-FS.

Wavelet Base	CUB-200-2011		CIFAR-FS		miniImageNet	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Haar	**79.59**	**90.91**	75.28	**86.84**	**65.98**	84.29
Daubechies	79.32	90.13	75.56	86.73	64.58	85.70
Biorthogonal	79.21	90.68	74.89	86.26	65.06	**85.93**
Symlets	79.44	90.46	**76.11**	85.91	64.85	84.08

components are crucial for improving classification accuracy. Compared with methods that rely on high-frequency components, FSPDF with low-frequency components improves classification accuracy. Significantly, despite the number of high-frequency components being three times greater than the low-frequency components, low-frequency components still play a crucial role in classification tasks.

(2) In 1-shot and 5-shot settings of the CUB-200-2011 dataset, employing solely low-frequency components enhances classification accuracy by 1.47% and 0.79%, respectively, over high-frequency components. Starting from the fine-grained nature of the dataset, images in CUB dataset typically have high intra-class consistency and low background interference. Therefore, using low-frequency components as the main feature of frequency domain analysis can preserve the main structure and significant features of the image.

In summary, we evaluate the importance of different frequency components in FSL classification tasks. On these datasets, we observe that abandoning high-frequency components and focusing on low-frequency components reduces noise and enhances classification accuracy. The effectiveness of this strategy indicates that low-frequency components have unique advantages in capturing global structure and visual information, especially when facing complex backgrounds and multi-object scenes.

4.5. Analysis of different wavelet basis selections in DWT

To validate the performance of FSPDF across different wavelet bases, we conduct evaluations using the Haar, Daubechies, Biorthogonal, and Symlets wavelet bases, as shown in Table 5. These tests are performed on three popular FSL benchmark datasets: miniImageNet, CUB-200-2011, and CIFAR-FS, covering 5-way 1-shot and 5-way 5-shot classification settings.

The following conclusions can be drawn from these results:

(1) Analysis of the experimental results across the three datasets shows consistent performance of the four wavelet bases in both 1-shot and 5-shot scenarios. This consistency indicates that FSPDF adapts well to different wavelet transform algorithms. FSPDF utilises these wavelet bases to enhance feature extraction and classification, enabling rapid learning from individual samples and knowledge accumulation across multiple samples.

(2) The Haar wavelet base demonstrates exceptional performance across all three datasets. Specifically, it achieves accuracies of 79.59% in the CUB-200-2011 1-shot task and 90.91% in the 5-shot task. In addition, the Haar wavelet base shows distinct advantages in the CIFAR-FS 5-way 5-shot setting and the miniImageNet 5-way 1-shot setting.

(3) The superior performance of the Haar wavelet base can be attributed to its characteristics. As one of the simplest wavelet bases, the Haar wavelet base has a straightforward functional form that is computationally efficient, making it faster in data processing compared to other wavelet bases.

(4) The compact support of the Haar wavelet base provides excellent localisation properties, which help capture local image features—an essential aspect for distinguishing different bird categories in CUB-200-2011. Additionally, the orthogonality of the Haar wavelet base proves effective in processing coarse-grained datasets such as CIFAR-FS and miniImageNet, reducing redundancy in feature representation and capturing critical information in images.

(5) Compared with the Haar wavelet base, Daubechies, Biorthogonal, and Symlets wavelet bases theoretically provide more detailed handling and better frequency domain localisation capabilities as their complexity increases. Experimental results show that the Symlets wavelet base achieves the best accuracy on CIFAR-FS under 5-way 1-shot tasks with 76.11%, and the Biorthogonal wavelet base performs best on miniImageNet under 5-way 5-shot tasks with 85.93%. However, in most other scenarios, their performance does not surpass that of the simpler Haar wavelet base. This suggests that the increased complexity of the wavelet base does not always lead to improved performance in FSL tasks.

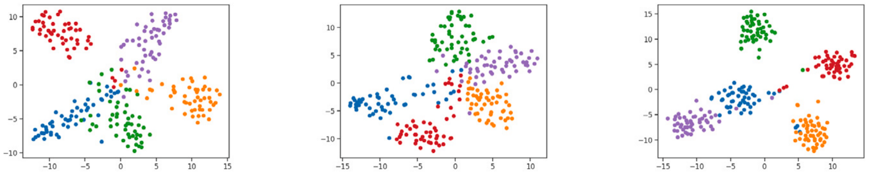
Our experimental results indicate that selecting a wavelet base requires careful consideration of its complexity and theoretical properties in the context of FSL tasks. It is also important to evaluate how well the wavelet base aligns with the characteristics of the dataset, as well as its effectiveness and efficiency in practical applications.

4.6. Parameter sensitive analysis

In this section, we discuss the effectiveness of the γ and η parameters (with values ranging from 0.3 to 0.7) on the classification accuracy of FSPDF. The γ parameter controls the influence of the DFL stage on the DFF stage, and the η parameter controls the trade-off between spatial and frequency domain features. In addition, the λ parameter setting is in accord with the baseline configuration, thereby no further analysis is performed here. Tables 6 and 7 present the classification accuracy for different parameter settings in the 5-way 1-shot and 5-way 5-shot tasks, respectively. The results demonstrate that FSPDF is sensitive to variations in γ and η parameters.

We have the following observations:

(1) From a longitudinal perspective, the variability in 1-shot experimental results is higher than that observed in 5-shot experiments. In the 1-shot experiments, the highest classification accuracy (79.94) is achieved when $\gamma = 5$ and $\eta = 5$. In the 5-shot experiments, the highest accuracy (92.02) is achieved with $\gamma = 7$ and $\eta = 5$. These results suggest that an appropriate weighting of self-supervised tasks facilitates more effective dual-domain feature learning in downstream tasks. However, when the γ value is too low, the allocation of weight



(a) Spatial domain (b) Frequency domain (c) Dual-domain fusion

Fig. 5. The t-SNE visualisation of spatial domain, frequency domain, and dual-domain of CUB-200-2011 under the 5-way 1-shot setting. Different colours represent different categories.

Table 6
Parameter sensitivity analysis for different values of η and γ under a 5-way 1-shot task on CUB-200-2011.

Parameters	$\eta = 3$	$\eta = 4$	$\eta = 5$	$\eta = 6$	$\eta = 7$
$\gamma = 3$	78.99	79.11	78.35	75.95	71.76
$\gamma = 4$	78.60	78.79	78.39	76.07	71.90
$\gamma = 5$	78.05	78.34	**79.94**	75.85	72.32
$\gamma = 6$	78.56	78.87	78.68	77.01	73.23
$\gamma = 7$	77.97	78.58	78.61	77.20	74.30

Table 7
Parameter sensitivity analysis for different values of η and γ under a 5-way 5-shot task on CUB-200-2011.

Parameters	$\eta = 3$	$\eta = 4$	$\eta = 5$	$\eta = 6$	$\eta = 7$
$\gamma = 3$	91.28	91.69	91.80	91.19	89.11
$\gamma = 4$	91.44	91.83	92.01	91.22	89.16
$\gamma = 5$	91.34	91.70	91.84	91.22	89.11
$\gamma = 6$	91.31	91.71	92.00	91.53	89.75
$\gamma = 7$	91.32	91.85	**92.02**	91.51	89.88

to the self-supervised task may become excessive, causing the model to over-rely on rotational information.

(2) From a horizontal perspective, dual-domain fusion effectiveness diminishes as the η parameter increasingly emphasises spatial features. However, at $\eta = 5$, feature complementarity is optimal, confirming the importance of balancing spatial and frequency domains to improve fusion performance.

(3) Excessively high or low γ values cause the model to either overly rely on or neglect rotational information. An appropriate γ parameter enables the model to learn rich features from rotational predictions, provides prior information for downstream tasks, and guides the extraction of dual-domain features. Additionally, the η parameter controls the fusion weights between spatial and frequency domain features, balancing global information with locally complementary features. Selecting suitable η values can maximise the complementarity of dual-domain features.

4.7. Visualisation of the model

To better understand the contribution of dual-domain features for the task, we visualised the features extracted from the spatial domain, frequency domain, and fused dual-domain features for the 5-way 1-shot classification task on CUB-200-2011. Fig. 5 shows the visualisation results of the FSPDF model, with five randomly selected categories, each category using 50 test samples. Different colours represent different categories in the figure.

The visualisation results of spatial domain features shown in Fig. 5(a) display good clustering performance for classification tasks, with relatively clear boundaries between different categories. However, some confusion exists in the central area of the image, suggesting that solely relying on spatial domain features for classification hard captures all the critical information of the data.

In Fig. 5(b), we observe that compared with spatial domain features, the distribution of samples in the frequency domain feature visualisation appears more dispersed, resulting in a boundary distortion situation. This demonstrates difficulties distinguishing different categories only depending on frequency domain features, especially in scenarios where high-frequency detail features play a decisive role in classification labels. This circumstance is linked to our strategy of selecting low-frequency parts as frequency domain features. We are more inclined to pay attention to low-frequencies as they better capture the global structure of images and reduce noise interference.

In Fig. 5(c), we observe that features fused from dual-domain display the best clustering results, with clearer boundaries between categories. This shows that dual-domain feature fusion utilises the complementarity of spatial and frequency domains, enhancing the ability of the model to differentiate between categories. The dual-domain fusion strategy enriches the dimensions of the features and compensates for details that might be overlooked by single-domain features, providing the model with a more diverse feature representation.

5. Conclusions and future work

In this paper, we introduce FSPDF. FSPDF addresses inadequate feature representation in scarce sample scenarios by employing a DFL and DFF approach. DFL uses a dual-domain self-supervised module to develop a feature extractor that enhances feature extraction and learning for subsequent stages. Subsequently, DFF leverages the previously obtained feature extractor and incorporates DWT to enrich and fuse dual-domain features. We conduct experiments in 1-shot and 5-shot settings across three standard few-shot classification datasets to validate the effectiveness of FSPDF. In the future, we plan to optimise feature fusion by assigning variable weights to features based on decision processes and further enhance the exploration of image frequency domain information through combinations of multiple wavelet bases and adaptive wavelet base methods.

CRediT authorship contribution statement

Dongqing Li: Writing – original draft, Software, Methodology, Conceptualization. **Jie Jin:** Validation, Methodology, Investigation. **Linhua Zou:** Validation, Investigation. **Hong Zhao:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62376114 and the Fujian Province Undergraduate Universities Teaching Reform Research Project under Grant No. FBJY20240103.

Data availability

Data will be made available on request.

References

[1] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, Adv. Neural Inf. Process. Syst. 30 (2017) 1690–1703.

[2] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conference on Machine Learning, 2017, pp. 1126–1135.

[3] J.C. Su, S. Maji, B. Hariharan, When does self-supervision improve few-shot learning? in: European Conference on Computer Vision, 2020, pp. 645–666.

[4] F. Teng, Q. Zhang, X. Zhou, J. Hu, T. Li, Few-shot ICD coding with knowledge transfer and evidence representation, Expert Syst. Appl. 238 (2024) 121861.

[5] L. Qain, Y. Bouteraa, T. Vaiyapuri, Y. Haung, A turning point few-shot learning for COVID-19 diagnosis, Eng. Appl. Artif. Intell. 133 (2024) 108337.

[6] Q. Ye, B.Y. Lin, X. Ren, Crossfit: A few-shot learning challenge for cross-task generalization in NLP, in: Conference on Empirical Methods in Natural Language Processing, 2021, pp. 7163–7189.

[7] X. Shi, S. Xue, K. Wang, F. Zhou, J. Zhang, J. Zhou, C. Tan, H. Mei, Language models can improve event prediction by few-shot abductive reasoning, Adv. Neural Inf. Process. Syst. 36 (2024) 29532–29557.

[8] Y. Zhang, D. Han, P. Shi, Semi-supervised prototype network based on compact-uniform-sparse representation for rotating machinery few-shot class incremental fault diagnosis, Expert Syst. Appl. 255 (2024) 124660.

[9] C. Qiu, T. Tang, T. Yang, M. Chen, Learning to generalize with latent embedding optimization for few-and zero-shot cross domain fault diagnosis, Expert Syst. Appl. (2024) 124280.

[10] Y. He, D. He, Z. Lao, Z. Jin, J. Miao, Z. Lai, Y. Chen, Few-shot fault diagnosis of turnout switch machine based on flexible semi-supervised meta-learning network, Knowl.-Based Syst. (2024) 111746.

[11] P. Zhu, Z. Zhu, Y. Wang, J. Zhang, S. Zhao, Multi-granularity episodic contrastive learning for few-shot learning, Pattern Recognit. 131 (2022) 108820.

[12] Y. Wang, Q. Yao, J.T. Kwok, L.M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM Comput. Surv. 53 (3) (2020) 1–34.

[13] A.A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, R. Hadsell, Meta-learning with latent embedding optimization, in: International Conference on Learning Representations, 2018, pp. 1–15.

[14] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, Matching networks for one shot learning, Adv. Neural Inf. Process. Syst. 29 (2016) 3637–3645.

[15] J. Zhou, Y. Zheng, J. Tang, L. Jian, Z. Yang, FlipDA: Effective and robust data augmentation for few-shot learning, in: Annual Meeting of the Association for Computational Linguistics, 2022, pp. 8646–8665.

[16] H. Yang, S. Qiu, A novel dynamic contextual feature fusion model for small object detection in satellite remote-sensing images, Information 15 (4) (2024) 210–230.

[17] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, K. Barnard, Attentional feature fusion, in: IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3560–3569.

[18] X. Liu, L. Li, F. Liu, B. Hou, S. Yang, L. Jiao, GAFNet: Group attention fusion network for PAN and MS image high-resolution classification, IEEE Trans. Cybern. 52 (10) (2021) 10556–10569.

[19] Y. Ding, X. Tian, L. Yin, X. Chen, S. Liu, B. Yang, W. Zheng, Multi-scale relation network for few-shot learning based on meta-learning, in: International Conference on Computer Vision Systems, 2019, pp. 343–352.

[20] Y. Piao, W. Ji, J. Li, M. Zhang, H. Lu, Depth-induced multi-scale recurrent attention network for saliency detection, in: IEEE/CVF International Conference on Computer Vision, 2019, pp. 7254–7263.

[21] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[22] A. Sinha, J. Dolz, Multi-scale self-guided attention for medical image segmentation, IEEE J. Biomed. Heal. Informatics 25 (1) (2020) 121–130.

[23] B. Shi, W. Li, J. Huo, P. Zhu, L. Wang, Y. Gao, Global-and local-aware feature augmentation with semantic orthogonality for few-shot image classification, Pattern Recognit. 142 (2023) 109702.

[24] Z. Qin, H. Wang, C.B. Mawuli, W. Han, R. Zhang, Q. Yang, J. Shao, Multi-instance attention network for few-shot learning, Inform. Sci. 611 (2022) 464–475.

[25] D. Kang, H. Kwon, J. Min, M. Cho, Relational embedding for few-shot classification, in: IEEE/CVF International Conference on Computer Vision, 2021, pp. 8822–8833.

[26] P. Doubinsky, N. Audebert, M. Crucianu, H. Le Borgne, Semantic generative augmentations for few-shot counting, in: IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 5443–5452.

[27] S. Benaim, L. Wolf, One-shot unsupervised cross domain translation, Adv. Neural Inf. Process. Syst. 31 (2018) 2108–2118.

[28] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T. Lillicrap, Meta-learning with memory-augmented neural networks, in: International Conference on Machine Learning, 2016, pp. 1842–1850.

[29] B. Hariharan, R. Girshick, Low-shot visual recognition by shrinking and hallucinating features, in: IEEE/CVF International Conference on Computer Vision, 2017, pp. 3018–3027.

[30] R. Ni, M. Goldblum, A. Sharaf, K. Kong, T. Goldstein, Data augmentation for meta-learning, in: International Conference on Machine Learning, 2021, pp. 8152–8161.

[31] W. Cho, E. Kim, Improving augmentation efficiency for few-shot learning, IEEE Access 10 (2022) 17697–17706.

[32] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, J. Wang, High-resolution representations for labeling pixels and regions, 2019, arXiv preprint arXiv:1904.04514.

[33] T. Shermin, S.W. Teng, F. Sohel, M. Murshed, G. Lu, Integrated generalized zero-shot learning for fine-grained classification, Pattern Recognit. 122 (2022) 108246.

[34] J. Rajasegaran, S. Khan, M. Hayat, F.S. Khan, M. Shah, Self-supervised knowledge distillation for few-shot learning, in: British Machine Vision Conference, 2021, pp. 1995–2006.

[35] P. Mangla, N. Kumari, A. Sinha, M. Singh, B. Krishnamurthy, V.N. Balasubramanian, Charting the right manifold: Manifold mixup for few-shot learning, in: IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2218–2227.

[36] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, 2020, pp. 1597–1607.

[37] J.B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent-a new approach to self-supervised learning, Adv. Neural Inf. Process. Syst. 33 (2020) 21271–21284.

[38] S. Tong, Y. Chen, Y. Ma, Y. Lecun, Emp-ssl: Towards self-supervised learning in one training epoch, 2023, arXiv preprint arXiv:2304.03977.

[39] S. Gidaris, P. Singh, N. Komodakis, Unsupervised representation learning by predicting image rotations, in: International Conference on Learning Representations, 2018, pp. 1–16.

[40] C. Doersch, A. Gupta, A.A. Efros, Unsupervised visual representation learning by context prediction, in: IEEE/CVF International Conference on Computer Vision, 2015, pp. 1422–1430.

[41] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, C. Feichtenhofer, Masked feature prediction for self-supervised visual pre-training, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14668–14678.

[42] R. Zhang, P. Isola, A.A. Efros, Colorful image colorization, in: European Conference on Computer Vision, 2016, pp. 649–666.

[43] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, N. Ballas, Self-supervised learning from images with a joint-embedding predictive architecture, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15619–15629.

[44] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.

[45] Z. Yan, Y. An, H. Xue, Reinforced self-supervised training for few-shot learning, IEEE Signal Process. Lett. (2024).

[46] H. Zhao, Y. Su, Z. Wu, W. Ding, CSTS: Exploring class-specific and task-shared embedding representation for few-shot learning, IEEE Trans. Neural Networks Learn. Syst. (2024) 1–13.

[47] C. Zhang, Y. Cai, G. Lin, C. Shen, Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12203–12213.

[48] K. Li, Y. Zhang, K. Li, Y. Fu, Adversarial feature hallucination networks for few-shot learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13470–13479.

[49] A. Dosovitskiy, P. Fischer, J.T. Springenberg, M. Riedmiller, T. Brox, Discriminative unsupervised feature learning with exemplar convolutional neural networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (9) (2016) 1734–1747.

[50] W. Chen, Y. Liu, Z. Kira, Y.F. Wang, J. Huang, A closer look at few-shot classification, in: International Conference on Learning Representations, 2019, pp. 1–24.