sciendo

JAISCR, 2026, Vol. 16, No. 3, pp. 237 – 256
10.2478/jaiscr-2026-0012

A VISUALLY EXPLAINABLE DYNAMIC SIMILARITY NETWORK FOR FEW-SHOT CLASSIFICATION

Zirui Pei[1], Zuqiang Meng[2,*], Tingting Diao[2], Peng Miao[2], Yifan Meng[2], Chaohong Tan[3]

[1]*College of Computer, Electronics and Information, Guangxi University*
Nanning, 530004, China

[2]*College of Computer, Electronics and Information, Guangxi University,*
Nanning, 530004, China

[3]*Guangxi Key Laboratory of Digital Infrastructure,*
Guangxi Zhuang Autonomous Region Information Center
Nanning, 530000, China

[*]*E-mail: zqmeng@126.com*

Submitted: 31st June 2025; Accepted: 27th December 2025

Abstract

Few-shot learning (FSL) aims to transfer knowledge from known to unknown categories using limited samples. However, the opaque nature of neural networks makes it challenging to discern the knowledge learned by the model, and existing methods often lack explainability, limiting their reliable application in high-stakes fields such as medical diagnosis and autonomous driving. To address this, we propose a visually explainable dynamic similarity network (VEDSNet), which achieves a balance of performance, explainability, and efficiency through a lightweight architecture (approximately 6.8M parameters, built on a ViT-Tiny backbone). The Feature Decomposition Module (FDM) generates fine-grained, semantically meaningful representations via parallel feature learning, providing intuitive visual insights into the model's decisions. The Dynamic Metric Module (DMM) employs a sample-adaptive dual-metric strategy to enhance discrimination with limited data, switching to a single metric for efficiency when data is sufficient. Experiments on standard datasets demonstrate that VEDSNet achieves high classification accuracy while providing clear visual explanations of its decision-making process, making it suitable for efficient deployment in resource-constrained scenarios.

Keywords: metric learning, attention mechanism, few-shot learning, explainable ai

1 Introduction

Deep learning models have achieved groundbreaking advances in computer vision [1], but their exceptional performance often relies on large-scale labeled datasets. This poses significant challenges in practical applications, such as medical diagnosis [56, 57]. To address data scarcity, FSL, which significantly reduces dependence on labeled samples, has gained widespread attention [2]. The core architecture of FSL typically comprises two components: 1) a feature extractor generating high-dimensional semantic representations, and 2) a sim-

ilarity metric function computing feature similarity between support and query samples. Existing FSL methods follow several directions: one enhances discriminative features by integrating attention mechanisms between the feature extractor and similarity function [3, 4]. However, these methods often rely on a single low-resolution feature representation, containing redundant information that limits the capture of subtle features. Another approach decomposes features into multiple fine-grained representations [34] to achieve comprehensive embeddings, improving classification accuracy but often neglecting model explainability. This lack of explainability hinders understanding and verification of the model's decision-making rationale, leading to trust issues. For instance, in image classification, a model may correctly classify a "person holding a dog" as "dog" but base its decision on human-dog interaction patterns rather than dog-specific features, raising concerns about generalization and reliability in high-stakes domains like medical diagnosis and autonomous driving.

Moreover, feature similarity calculation remains a key challenge in FSL. Many methods rely on a single similarity metric [41, 34], which, while efficient, struggles in complex feature distribution scenarios. Dual-similarity networks, such as BSNet [28], integrate complementary metrics to improve performance in fine-grained FSL tasks, offering a new perspective for metric learning. The choice of feature extractor is also critical. Traditional FSL methods use deep convolutional neural networks (e.g., ResNet) to extract local features and form semantic representations. Recent approaches introduce innovative architectures like Vision Transformers (ViT) and Wavelet Attention Networks (WAN) [5, 6, 7, 8, 9] to capture long-range dependencies. However, these models often have large parameter sizes and high computational costs, posing challenges for deployment in resource-constrained environments. Thus, selecting a lightweight feature extractor while maintaining performance is a pressing research direction in FSL.

To address these challenges, we propose an explainable FSL framework, Visually Explainable Dynamic Similarity Network (VEDSNet), leveraging a lightweight architecture (approximately 6.8M parameters, built on a ViT-Tiny backbone [10]) to balance performance, explainability, and efficiency. VEDSNet incorporates two key modules: the FDM and the DMM. The FDM adapts attention mechanisms to decompose visual features into fine-grained semantic regions through parallel feature learning and iterative updates, enabling interpretable visualizations of decision-relevant areas (Figure 1), where blue areas indicate suppressed regions and red areas highlight attended regions. This decomposition reveals the model's decision rationale in FSL tasks, validated through controlled experiments to enhance transparency over existing methods while preserving computational efficiency within the lightweight VEDSNet architecture. The DMM optimizes existing metric-learning approaches [28] by dynamically adapting its similarity strategy based on sample size. In low-data regimes, it balances learnable and fixed metrics using a correlation-based coefficient γ, validated via ablations to enhance training stability and avoid local optima. In data-sufficient scenarios, it switches to a single efficient metric, reducing computational complexity. Ablation studies confirm the robustness of γ and the K-based switching threshold. This adaptive strategy, combined with a lightweight backbone, ensures high classification accuracy and efficient deployment in resource-constrained settings. The main contributions of this study are summarized as follows:

1. We propose an explainable FSL framework that achieves strong classification accuracy while providing qualitative and quantitative insights into the model's decision-making process.
2. We propose an end-to-end trainable framework based on the Vision Transformer, which manages the model's parameter size and computational demands while delivering solid performance.
3. We propose an FDM that breaks down visual features into attention-weighted components for visualization, helping the model focus on relevant regions while preserving classification accuracy.
4. We design a DMM that adjusts its similarity strategy based on data availability, using dual pathways to improve discrimination with limited data and switching to a single pathway for efficiency when data is sufficient.

Figure 1. FSL using the FDM showing attention visualization across different feature groups. Images are from the Tiered-ImageNet dataset [18].

2 Related work

2.1 Few-shot learning

Few-shot classification is a key challenge in deep learning, where the goal is to build a classifier with strong generalization using very limited labeled data. In this setting, each class has only a few training examples, causing traditional deep learning methods, which rely on large datasets, to perform poorly. To tackle data scarcity, researchers have developed various techniques, with metric learning [11, 12] and transfer learning [13] being particularly effective, as they improve few-shot classification performance and domain adaptability. Transfer learning, a well-established approach in FSL, leverages knowledge from a source domain with abundant labeled data to enhance model performance in a target domain with scarce annotations [14]. State-of-the-art visual models [15, 16] typically follow a "pre-train then fine-tune" approach: pre-training on a large-scale dataset like ImageNet-21k [17], then fine-tuning on smaller target datasets such as Tiered-ImageNet [18] or CIFAR-FS [19]. To improve knowledge transfer, various fine-tuning strategies have been proposed. Some methods focus on classifier-only fine-tuning [20], freezing the pre-trained feature extractor (e.g., a ResNet backbone) and adjusting only the final fully connected layer. However, this approach depends heavily on semantic similarity between source and target tasks; if the tasks differ significantly, transfer effectiveness decreases. Other methods use parameter-efficient tuning, inserting lightweight trainable modules into the pre-trained network [21, 22]. While these reduce the number of trainable parameters, they add architectural complexity and may face limitations in fine-grained tasks, leading to performance constraints. Metric learning methods [23, 24, 25] aim to create an optimized embedding space where samples from the same class cluster tightly and different classes are well separated. Siamese networks [26] pioneered deep metric learning for few-shot classification, using a twin-tower architecture and contrastive loss to establish a foundation. Matching Networks [29] introduced attention-based weighting of support samples for an end-to-end k-NN framework. Prototypical Networks [27] simplified computation by using class prototypes (mean embeddings), improving within-class consistency. Relation Networks [49] advanced this by replacing fixed metrics (e.g., Euclidean or cosine) with a learnable nonlinear comparator (MLP), better capturing complex relationships. Recent studies [28, 48] have shown that a dual-metric mechanism can enhance performance on fine-grained few-shot benchmarks, inspiring further exploration in metric learning. Our method builds on these established approaches, combining their strengths to achieve improved discrimination and generalization.

2.2 Explainable ai

Explainable AI methods are generally divided into two categories: post-hoc explanation methods and intrinsic explanation methods. Post-hoc methods typically generate feature importance heatmaps using backpropagation gradients, such as Layer-wise Relevance Propagation (LRP), network gradients, or other attribution techniques [43]. In contrast, intrinsic explanation methods embed transparency into the model design. Some approaches use human-readable structures like decision trees [44] or attention mechanisms [32], while others adopt concept-based methods, such as Concept Bottleneck Models [45, 55], or visualize the regions the model focuses on through activation maps and attribution techniques [46, 43]. In this work, we adopt an intrinsic explainability approach and propose an explainable FSL framework, VEDSNet, based on the self-attention mechanism. It incorporates an FDM to extract informative and discriminative re-

gions and a DMM to enhance classification performance. Experimental results show that VED-SNet achieves strong performance on several few-shot classification benchmarks while providing visual explanations that effectively highlight potential prediction errors, which is especially valuable in high-stakes domains like healthcare.

3 Model

3.1 Problem definition

Given a dataset for training a few-shot image classification model, the dataset consists of three components: a training set D_{train}, a support set D_{support}, and a test set D_{test}. The training set contains independent class spaces where each class is associated with a large number of labeled image examples. These classes in D_{train} are defined as base classes C_{base}. In contrast, the support set D_{support} and test set D_{test} share the same class space, which is disjoint from D_{train}. The classes in D_{support} and D_{test} are defined as novel classes C_{novel}. If the support set contains M novel classes, each with K image examples, this FSL problem is termed an M-way K-shot learning task. The goal of FSL is to train an image classification model using D_{train} and D_{support} such that the model can accurately classify test images from D_{test} into the novel classes even when K is small.

3.2 General form of metric-learning-based method

Euclidean Distance Constraint: $d_{\text{L2}}(u,v) = ||u - v||_2$. The Euclidean distance assigns equal weighting to all dimensions in the feature space, assuming unit weights for each dimension. Cosine Similarity Constraint: $d_{\cos}(u,v) = 1 - (u^T v)/||u||||v||$. Cosine similarity disregards feature magnitude by normalizing vectors, effectively measuring distances on a unit hypersphere. Mahalanobis Distance Constraint: $d_M(u,v) = \sqrt{(u-v)^T \Sigma^{-1} (u-v)}$. The Mahalanobis distance accounts for feature covariance, but its quadratic form limits its ability to model non-linear manifolds.

3.3 Overview

In this section, we elaborate on the design details of the proposed FDM and DMM, with the model architecture shown in Figure 2. We extract low-level features using a lightweight vision Transformer structure [10]. The input image $X_{\text{raw}} \in \mathbb{R}^{H\times W\times 3}$, where h is the arithmetic square root of H, is processed through the backbone network to output feature maps $F_{\text{raw}} \in \mathbb{R}^{H\times W\times C}$. To better adapt to the sequential processing of the FDM, the features are reshaped to $F \in \mathbb{R}^{C\times W\times h\times h}$. Subsequently, the reshaped features are fed into a 1×1 convolutional layer, compressing the channel dimension to D, resulting in compressed features $f \in \mathbb{R}^{D\times W\times h\times h}$.

3.4 Feature decomposition module

Then the compressed features are input into the FDM. To maintain the integrity of the feature space structure, we generate positional encoding P to preserve spatial information. The features combined with positional encoding yield enhanced features $f_{pe} = f + p$. The initial values for each feature group are sampled from a normal distribution, and through expansion operations, L groups of features $S \in \mathbb{R}^{B\times L\times D}$ are generated. This multi-group parallel design enables the model to simultaneously explore multiple locally optimal regions in the feature space, as shown in Figure 3.

We use the initialized learnable feature vectors as Query, linear projections of the position-enhanced features f_{pe} as Key, and the compressed features as Value. Attention is calculated on each spatial dimension based on Query and Key, using a normalization function to obtain:

$$A = \phi(Q \cdot K^T) \odot \varepsilon(Q \cdot K^T) \tag{1}$$

where ϕ is the softmax function that normalizes attention weights across slots to ensure competitive feature allocation, and ε is the sigmoid function that provides element-wise gating to modulate attention activation strength. The slot representations are then updated iteratively using the computed attention weights:

$$S_{t+1} = \text{GRU}(V \cdot A^T, S_t) \tag{2}$$

The value of S is updated through a GRU mechanism, which dynamically integrates histori-

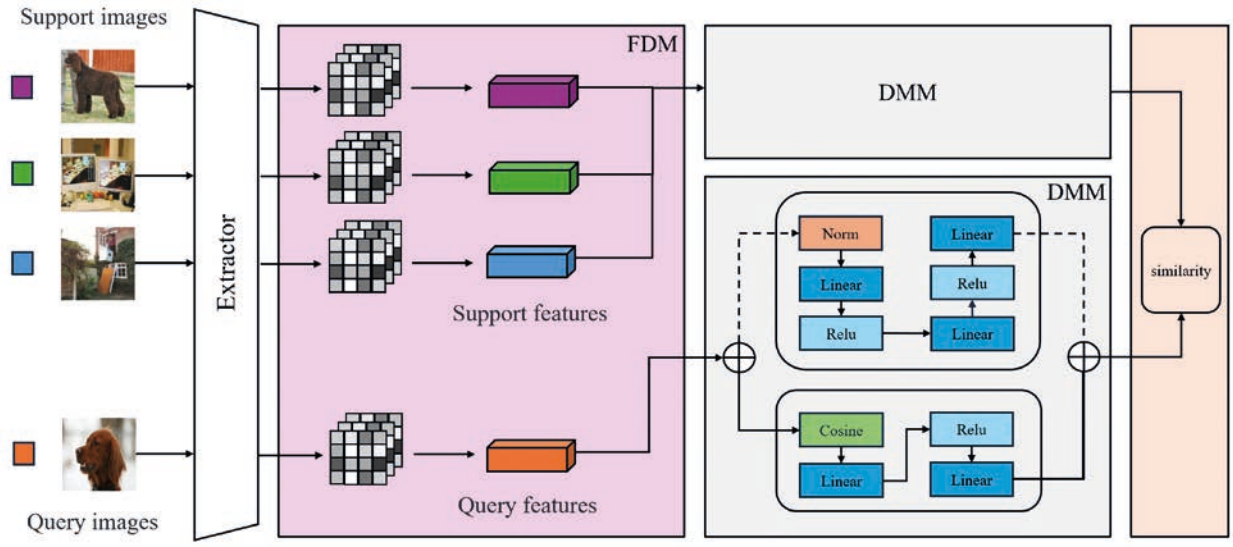

Figure 2. The overall structure of VEDSNet is as follows. In the FDM, input features are replicated into *L* copies, which, through parallel processing paths and the application of attention mechanisms, enable the model to capture diverse feature information. Each copy learns different focus points through self-attention mechanisms, decomposing complex visual information into explainable components, thereby enhancing the model's discriminative capabilities. Subsequently, the aggregated features are fed into the DMM, which employs different metric strategies based on sample quantities.

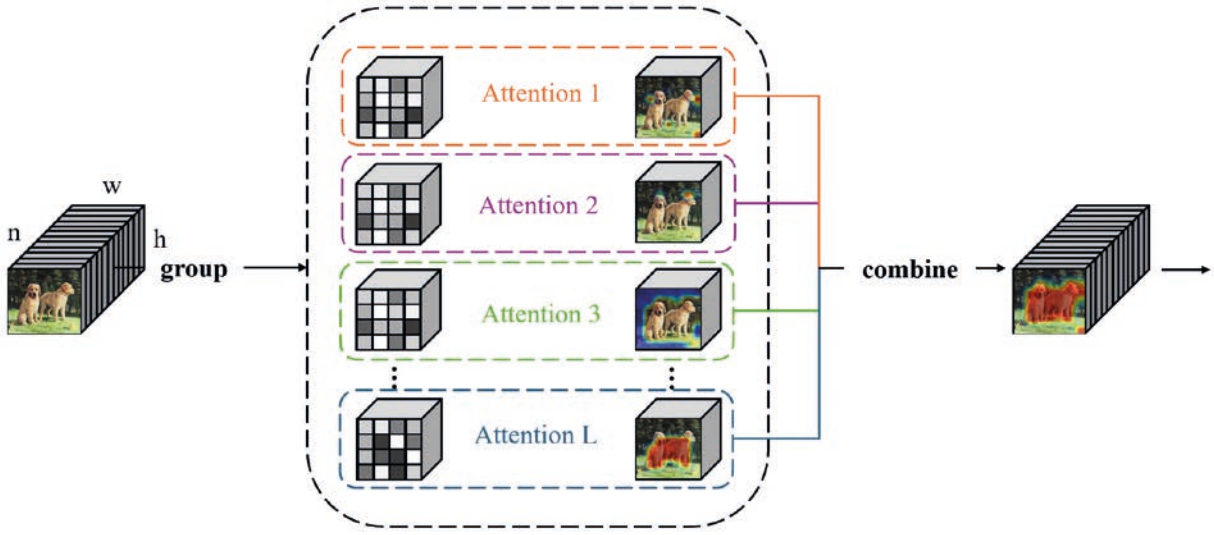

Figure 3. Feature Decomposition Module (FDM).

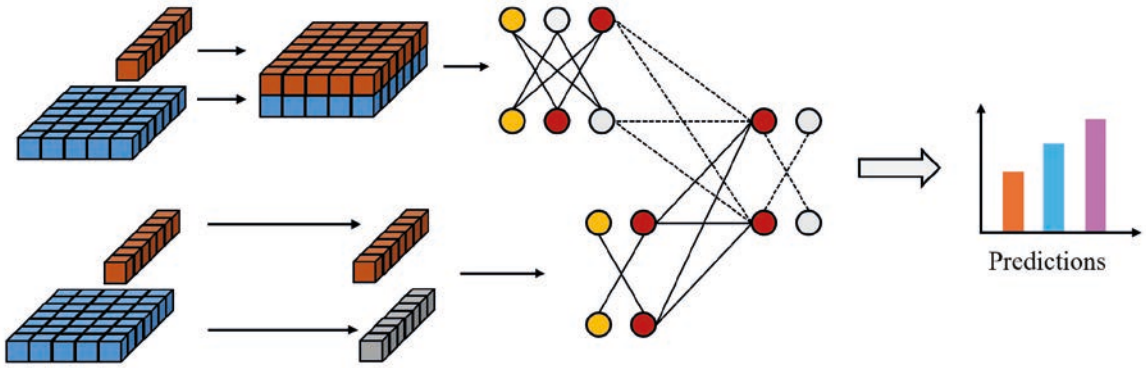

Figure 4. Dynamic Metric Module (DMM).

cal states with current updates via its gating mechanism. This process is repeated three times (we will analyze the threshold in section 5.5). Through iterative optimization, we obtain the final attention weights. To further enhance the model's discriminative capability, we introduce learnable weight vectors $W \in \mathbb{R}^{L\times 1}$ in the model to filter attention for each group effectively. These weight vectors, through an adaptive learning mechanism, dynamically weigh the attention of each slot, thereby distinguishing important slots from redundant ones:

$$a = A \cdot W \tag{3}$$

where L is the total number of slots, initialized as an all-one vector. As a quantitative index of feature importance, the numerical range of the weight coefficient W is significantly related to the direction of feature activation.

To better perform metric learning, we have done a few things: the attention weights are divided into support weights A_s and query weights A_q. Mean operations are performed on A_s and A_q along the first dimension (i.e., the sample dimension), represented as:

$$a_s = \frac{1}{N}\sum_{i=1}^{N} A_s^i \tag{4}$$

$$a_q = \frac{1}{N}\sum_{i=1}^{N} A_q^i \tag{5}$$

where A_s^i and A_q^i are tensors for each sample in A_s and A_q. The mean operation results in a tensor with shape (1, C, H, W) (i.e., averaging along the sample dimension).

Subsequent feature fusion is performed:

$$f_s = \frac{1}{HW}\sum_{ij} a_s^{(i,j)} F^{(i,j)} \tag{6}$$

$$f_q = \frac{1}{HW}\sum_{ij} a_q^{(i,j)} F^{(i,j)} \tag{7}$$

3.5 Dynamic metric module

This section describes a dynamic hybrid metric learning framework to compute similarities between feature pairs, as shown in Figure 4. In FSL, metric learning aims to quantify similarity between support and query samples. Conventional methods use fixed metric functions, such as Euclidean distance, cosine similarity, or Mahalanobis distance, to compare embedded features. However, these metrics rely on specific geometric assumptions, such as isotropic feature spaces for Euclidean distance or magnitude invariance for cosine similarity. These assumptions may not align with the complex, nonlinear feature distributions in real-world tasks, especially when sample sizes vary, limiting the model's ability to capture cross-category or cross-domain variations.

Using support features f_s and query features f_q from the feature extraction network, we develop a dual-pathway metric system comprising a geometric constraint pathway and an adaptive learning pathway. The geometric constraint pathway preserves intrinsic geometric structure via prototype-based similarity computation:

$$C_k = \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\frac{1}{HW}\sum_{i,j} a_s^{(i,j)} f^{(i,j)}\right) \tag{8}$$

where N_s denotes the number of support samples per class, and k is the class index. A multilayer perceptron (MLP) extracts high-level feature representations from low-dimensional class prototypes to enhance class discriminability:

$$S_{\mathrm{g}}(f_s, f_q) = \sigma(\mathrm{MLP}_{geo}(\xi(C_k))) \tag{9}$$

where $S_{\mathrm{g}}(\cdot)$ is the geometric constraint metric function that preserves spatial relationships in the embedding space.

For tasks with high sample complexity, geometric constraints alone are insufficient to capture task-specific similarity patterns. The adaptive learning pathway uses a learnable non-linear metric network to identify complex relationships in the data:

$$S_{\mathrm{a}}(f_s, f_q) = \sigma(\mathrm{MLP}_{adapt}([f_s; f_q])) \tag{10}$$

where $S_{\mathrm{a}}(\cdot)$ is a non-linear metric network that captures intricate feature interactions. The support features f_s and query features f_q are concatenated along the channel dimension, forming $[f_s; f_q] \in \mathbb{R}^{2C}$. Task-specific MLPs apply non-linear transformations to explore cross-modal associations.

To adapt to varying sample complexity and feature distributions, we introduce a learnable statistical stability coefficient. The adaptive fusion weight is computed as:

$$X_{\mathrm{pooled}} = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} X_{\mathrm{raw}}^{(i,c,h,w)} \tag{11}$$

$$\gamma = \sigma(\mathrm{MLP}(\mathrm{LayerNorm}(X_{\mathrm{pooled}}))) \tag{12}$$

where $\gamma \in [0,1]$ is a statistical stability coefficient that reflects feature correlation patterns and balances the contributions of geometric and adaptive pathways. We apply layer normalization to stabilize feature statistics before computing γ, followed by a sigmoid activation to constrain γ within $[0,1]$.

This approach optimizes computational efficiency while maintaining discriminative performance. The adaptive metric selection strategy is defined as:

$$S = \begin{cases} \gamma \cdot S_{\mathrm{g}}(f_s, f_q) + (1-\gamma) \cdot S_{\mathrm{a}}(f_s, f_q) & \text{if } K \leq 3 \\ S_{\mathrm{g}}(f_s, f_q) & \text{if } K > 3 \end{cases} \tag{13}$$

We will analyze the threshold in section 5.1.

3.6 Train our model

We introduce a two-stage training paradigm for the VEDSNet architecture to ensure robust feature extraction and optimal decision-making. The initial stage focuses on pre-training the FDM to establish a reliable foundation for feature representation. The subsequent stage incorporates the DMM for end-to-end joint optimization. The training process is guided by a composite loss function that integrates attention regularization with classification objectives to achieve balanced and effective learning.

To promote effective allocation of attention weights, we define the attention loss as follows:

$$\mathcal{L}_{\mathrm{att}} = \left(\frac{1}{B \cdot L \cdot T} \sum_{b=1}^{B} \sum_{l=1}^{L} \sum_{t=1}^{T} \sigma(\mathbf{A}_{b,l,t}) \right)^{p} \tag{14}$$

where $\mathbf{A}_{b,l,t}$ represents the attention weight associated with the t-th input position in the l-th slot of the b-th batch, B denotes the batch size, L signifies the number of attention slots, T indicates the length of the input sequence (corresponding to the number of spatial positions in the feature maps), p is a hyperparameter that modulates the nonlinearity of the loss, and σ represents the activation function.

The cross-entropy loss serves as the primary objective for optimizing classification performance. Within the FSL framework, we leverage ground-truth labels from the query set to supervise the model, minimizing the divergence between predicted and actual label distributions:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log\left(\sigma(\mathbf{y}_{\mathrm{pred}})_{i,c}\right) \tag{15}$$

where $\mathbf{y}_{\mathrm{pred}}$ denotes the predicted logits, $y_{i,c}$ represents the ground-truth label for the i-th sample and c-th class, N is the number of query set samples, C is the number of classes, and σ is the activation function.

The total loss function harmonizes classification accuracy and attention regularization through a weighting hyperparameter λ:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CE}} + \lambda \cdot \mathcal{L}_{\mathrm{att}} \tag{16}$$

This composite loss function ensures that the model optimizes both discriminative accuracy and attention coherence effectively.

This two-stage training approach first stabilizes feature extraction within the FDM during pre-training, followed by comprehensive optimization of the DMM in the joint training phase. The carefully designed loss function enables VEDSNet to achieve robust and accurate performance in FSL tasks, balancing feature discriminability with effective decision-making.

4 Experiments

This section evaluates the performance of the proposed VEDSNet framework against state-of-the-art methods to demonstrate its effectiveness. Through rigorous experimental analysis, we validate the superiority and robustness of our approach in FSL tasks.

4.1 Datasets

We conduct experiments on four mainstream benchmark datasets, including Mini-ImageNet [29], Tiered-ImageNet [18], CUB-200-2011 [47] and CIFAR-FS [19]. Mini-ImageNet [29] contains 100 classes randomly divided into 64 base classes, 16 validation classes, and 20 novel classes, with each class containing 600 images. Tiered-ImageNet is divided into 351 base classes (training), 97 validation classes, and 160 novel classes (testing), maintaining the independence of superclass hierarchies during partitioning (i.e., all subclasses of the same superclass appear only in the same group). CIFAR-FS includes 100 classes randomly divided according to standard few-shot partitioning strategies into 64 base classes, 16 validation classes, and 20 novel classes, with each class containing 600 images. CUB-200-2011 consists of 200 classes divided into 100 base classes, 50 novel validation classes, and 50 novel test classes. There are approximately 11,788 images.

4.2 Experimental details

This research utilized a computer equipped with one RTX 4090 GPU (24GB memory). Following the practice in most literature, we evaluate our model on 2,000 episodes of 5-way classification tasks. These tasks are constructed by randomly selecting five classes from D_{base}, and then sampling support set and query set images from these classes, where the number of samples in the support set is K = 1 or 5, and the number of samples in the query set is M = 15. We report the average accuracy across 2,000 tasks, each with $K \times M = 75$ queries, along with 95% confidence intervals. We adopt the ViT-Tiny architecture pre-trained on ImageNet-21k as the backbone network $F_{(\mathrm{x})}$. We remove the classification head from ViT-Tiny and add dimension permutation operations, ultimately restoring the sequence features into spatial feature maps F, with the number of feature maps being 320. The model training is divided into two steps.

Training of the FDM: we pre-train the FDM with an embedding dimension $d = 64$ and seven groups. Images are sampled from subclasses of the base dataset, and the FDM is trained in a standard classification task, following the methodology [32]. During this phase, all parameters of the backbone network $F_{(\mathrm{x})}$ are fixed. The initial learning rate is set to 10^{-4}, which is reduced to 10^{-5} at the 40th epoch, with training conducted over 60 epochs.

Overall Training of the Model: we perform end-to-end fine-tuning of the backbone network $F_{(\mathrm{x})}$, the FDM, and the positional encoding module. The backbone and positional encoding parameters are optimized with a learning rate of 10^{-5}, while other trainable modules use an initial learning rate of 10^{-4}, which undergoes a 10-fold decay at the 10th epoch. Training spans 30 epochs, with each epoch involving 500 5-way task scenarios. We compute training accuracy, training loss, validation accuracy, and validation loss per epoch, using the cross-entropy loss function for classification. Model performance is evaluated on a validation set comprising 1,000 task scenarios, and the parameters yielding the best validation performance are saved.

The model is implemented in PyTorch with the AdaBelief optimizer [52]. Input images are resized to 224×224 pixels and augmented via random horizontal flips, affine transformations (rotation $\pm 15°$, scaling $[0.8, 1.0]$, translation $\pm 10\%$), and Gaussian blur ($\sigma \in [0, 3.0]$), implemented using the `imgaug` library. Training on the tiered-ImageNet dataset comprises 500 training and 1000 validation episodes per epoch, taking approximately 210 minutes for 30 epochs. This configuration ensures fair comparison and reproducibility on standard GPU hardware.

Table 1. Average accuracy with 95% confidence intervals for 5-way tasks on Tiered-ImageNet dataset.

Method	Backbone	Params	1-shot	5-shot
FRN[23]	ResNet-12	12M	72.06±0.22	86.89±0.14
TPMM[30]	ResNet-12	12M	72.24±0.70	86.55±0.63
SUN[6]	ViT-S	12M	72.99±0.50	86.74±0.33
SSFormers[31]	ResNet-12	12M	72.52±0.25	86.61±0.18
FewTURE[8]	Vit-Small	22M	72.96±0.92	86.43±0.67
MetaQDA[7]	WAN-28-10	36M	74.33±0.65	89.56±0.79
CPEA[9]	ViT-S/16	22M	**76.93±0.70**	90.12±0.45
OM[33]	WAN-28-10	36M	71.54±0.29	87.79±0.46
GML[38]	ResNet-12	12M	71.09±0.22	85.08±0.16
SP-GloVe[5]	ViT-T	10M	74.68±0.50	88.64±0.31
SP-SBERT[5]	ViT-T	10M	73.31±0.50	88.56±0.32
Ours	ViT-Tiny	5M	**76.58±0.39**	**94.64±0.33**

Table 2. Average accuracy with 95% confidence intervals for 5-way tasks on CIFAR-FS dataset.

Method	Backbone	Params	1-shot	5-shot
MetaQDA[7]	WAN-28-10	36M	75.83±0.88	88.79±0.75
SSFormers[31]	ResNet-12	12M	74.50±0.21	86.61±0.23
HCTransformers[37]	ViT-S	21M	78.88±0.18	90.50±0.09
FewTURE[8]	Vit-Small	22M	76.10±0.88	86.14±0.64
FewTURE[8]	Swin-Tiny	29M	77.76±0.81	88.90±0.59
SP-CLIP[5]	ViT	10M	82.18±0.40	88.24±0.32
SUN[6]	ViT-S	12M	78.37±0.46	88.84±0.32
CPEA[9]	ViT-S/16	22M	77.82±0.66	88.98±0.45
GML[38]	ResNet-12	12M	69.61±0.23	84.04±0.16
SP-SBERT[5]	ViT-T	10M	81.32±0.40	88.31±0.32
SP-GloVe[5]	ViT-T	10M	81.62±0.41	88.32±0.32
Ours	ViT-Tiny	5M	**82.78±0.45**	**91.71±0.36**

Table 3. Average accuracy with 95% confidence intervals for 5-way tasks on Mini-ImageNet dataset.

Method	Backbone	Params	1-shot	5-shot
FRN[23]	ResNet-12	12M	66.45±0.19	82.83±0.13
TPMM[30]	ResNet-12	12M	67.64±0.63	83.44±0.43
MetaQDA[7]	WAN-28-10	36M	67.83±0.64	84.28±0.69
SUN[6]	ViT-S	12M	67.80±0.45	83.25±0.30
SSFormers[31]	ResNet-12	12M	67.25±0.24	82.75±0.20
FGFL[35]	ResNet-12	12M	69.14±0.80	86.01±0.62
TST_MFL[36]	ResNet-12	12M	60.42±0.82	80.03±0.58
FA-adapter[34]	ResNet-12	12M	68.36±0.45	83.44±0.28
GML[38]	ResNet-12	12M	65.51±0.20	81.13±0.14
SP-SBERT[5]	ViT-T	10M	**70.70±0.42**	83.55±0.30
FewTURE[8]	ViT-Tiny	5M	67.32±0.41	81.10±0.61
Ours	ViT-Tiny	5M	**70.38±0.47**	**88.53±0.65**

Table 4. Average accuracy with 95% confidence intervals for 5-way tasks on the CUB dataset. Parameter counts (Params) are reported as provided in the original papers. For methods where this information is not available, the parameter field is denoted as "–".

Method	Backbone	Params	1-shot	5-shot
SaberNet[42]	Swin-4	–	84.25±0.78	92.77±0.45
SRM[41]	ResNet-12	12M	83.82±0.18	93.45±0.10
FA-adapter[34]	ResNet-12	12M	84.59±0.56	93.43±0.43
FRN[23]	ResNet-12	12M	83.55±0.19	92.92±0.10
ESPT[40]	ResNet-12	12M	**85.45±0.18**	94.02±0.09
LCCRN[50]	ResNet-12	12M	82.97±0.19	93.63±0.10
MCL-Katz[51]	ResNet-12	12M	**85.11±0.63**	93.18±0.35
TDM[3]	ResNet-12	12M	84.36±0.19	93.37±0.10
Ours	ViT-Tiny	5M	**84.63±0.47**	**95.04±0.27**

4.3 Experimental results

Tables 1, 2, 3, and 4 present the performance of our proposed VEDSNet framework compared to state-of-the-art methods for 1-shot and 5-shot tasks on the Mini-ImageNet [29], CIFAR-FS [19], Tiered-ImageNet [18], and CUB-200-2011 [47] datasets. Our approach demonstrates competitive performance relative to existing FSL methods while additionally providing interpretable visual explanations. In contrast to many advanced methods that rely on large-scale backbone networks with extensive parameter counts, our model employs a lightweight architecture. This efficiency is attributed to the design of our network, which effectively leverages intrinsic data information while maintaining strong generalization capabilities, thereby validating its efficacy in addressing FSL tasks.

5 Analysis

This section substantiates the efficacy of the proposed VEDSNet framework through comprehensive ablation studies and visual analysis. By systematically evaluating the contributions of individual components and providing interpretable visual insights, we validate the robustness and effectiveness of our approach in FSL tasks.

5.1 Analysis of dynamically switching thresholds

To verify the reasonableness of the threshold switching in the DMM, we performed ablation studies. Specifically, we examined the performance implications of various fixed thresholds alongside continuous scheduling strategies across $K \in \{1,2,3,4,5,10\}$ to mitigate potential biases in threshold determination. These investigations were conducted on the Tiered-ImageNet dataset to assess performance in a FSL setting. Consistent with the primary experiments, we adopted a ViT-Tiny backbone architecture and identical training protocols, employing 5-way classification episodes. Evaluation was based on 2,000 randomly sampled test episodes, with results reported as mean classification accuracy (%) accompanied by 95% confidence intervals. As shown in Table 5

5.2 Computational efficiency analysis

To assess the FDM, we compared a grouped structure with an ungrouped baseline. Both configurations maintain identical model sizes and parameter counts, confirming that the grouped structure introduces no additional parameter overhead. The grouped FDM exhibits slightly higher inference latency compared to the ungrouped baseline, attributed to the computational cost of coordinating parallel pathways and their integration. In batch processing, the grouped configuration shows marginally higher average processing time but demonstrates lower variability, indicating greater stability. Peak memory usage remains consistent across both configurations, highlighting the lightweight nature of the grouping mechanism. The throughput of the grouped configuration is slightly lower than that of the ungrouped version, consistent with its latency profile. These minor trade-offs

Table 5. Performance metrics are evaluated for varying values of K to validate the scheduling threshold. Average accuracy with 95% confidence intervals is reported for 5-way few-shot classification tasks on the tieredImageNet dataset.

Method	K**=1**	K**=2**	K**=3**	K**=4**	K**=5**	K**=10**
Single-Metric	63.70±0.43	77.35±0.53	84.25±0.48	93.29±0.45	94.64±0.33	95.77±0.23
Dual-Metric	76.58±0.39	82.37±0.78	86.25±0.36	88.77±0.47	90.41±0.39	95.58±0.29

in latency and throughput are offset by the grouped FDM's enhanced stability and richer feature representations, which are particularly advantageous for applications requiring high explainability and robustness in FSL tasks.

5.3 Analysis of metric methods

To thoroughly evaluate the efficacy of the DMM, we performed extensive ablation studies on three widely recognized benchmark datasets: Mini-ImageNet, CIFAR-FS, and Tiered-ImageNet. Figure 5 presents the outcomes of these experiments, comparing four distinct configurations: (1) LE, which relies exclusively on a learnable distance metric; (2) CO+MLP, which employs a cosine similarity metric integrated with a multi-layer perceptron; (3) LE&CO, a direct combination of both metrics; and (4) Ours, the proposed dynamic hybrid approach. The results consistently highlight the superiority of our dynamic hybrid method. In the 1-shot setting, our approach significantly outperforms the individual metric configurations across all datasets, yielding notable enhancements in classification accuracy. In the 5-shot setting, our method continues to demonstrate a robust performance advantage, underscoring the effectiveness of the dynamic balancing mechanism in adapting to varying numbers of support samples. These findings affirm the enhanced robustness and generalization capabilities of our approach in FSL tasks.

5.4 Set number of groups

The number of feature groups, denoted as L, is a critical hyperparameter in the VEDSNet model. A smaller L reduces classification accuracy, indicating the need for a sufficient number of feature groups to enable effective feature decomposition. Conversely, an excessively large L leads to training instability across all datasets. The optimal L must be tuned for each dataset, with $L = 7$ often achieving a balance between discriminative power and stability, as shown in Figure 6. This sensitivity to L represents a limitation of the VEDSNet model. The absence of the FDM (w/o FDM) significantly degrades performance, underscoring its importance for robust classification.

5.5 Effect of recurrent update depth in the FDM

To examine the impact of iterative updates within the FDM, we conducted experiments on the Mini-ImageNet dataset under the 5-way 5-shot setting, evaluating different GRU update depths ranging from one to five iterations. Performing three GRU updates achieved the best balance between representation refinement and computational cost. Fewer iterations led to insufficient interaction among slot states, resulting in less stable concept separation, while deeper updates produced diminishing returns and slightly increased overfitting risk. We found negligible variation in runtime and memory consumption across different update depths, indicating that the recurrent refinement process adds minimal computational overhead. Therefore, we adopt three recurrent updates as the default setting, as it provides consistent performance gains on Mini-ImageNet without incurring notable computational or generalization penalties.

Table 7. Effect of GRU update depth (T) in the FDM on performance and efficiency.

T	1	2	3	4	5
Acc(%)	85.37	86.28	**88.53**	65.53	40.41

5.6 Explainability Analysis

The FDM, which encodes input images into multiple concept-like feature groups and visualizes their distinct contributions to decision-making, is a

Table 6. The performance metrics of the VEDSNet framework's FDM configurations are systematically evaluated. Improvements in classification accuracy are primarily attributed to the grouping strategy employed within the FDM. We report the average accuracy with 95% confidence intervals for 5-way few-shot classification tasks on the tieredImageNet dataset. The model size, representing the memory footprint, is quantified to highlight the computational efficiency of the proposed configurations.

Method	Accuracy	Params	Size	Latency	Batch Time	Peak Mem	Throughput
Single-Path	90.44±0.30	6.8M	28.14MB	32.96ms	58.10±1.60	1238.35MB	30.33episodes/sec
Multi-Branch	94.64±0.33	6.8M	28.14MB	33.26ms	58.53±1.68	1238.36MB	30.07episodes/sec

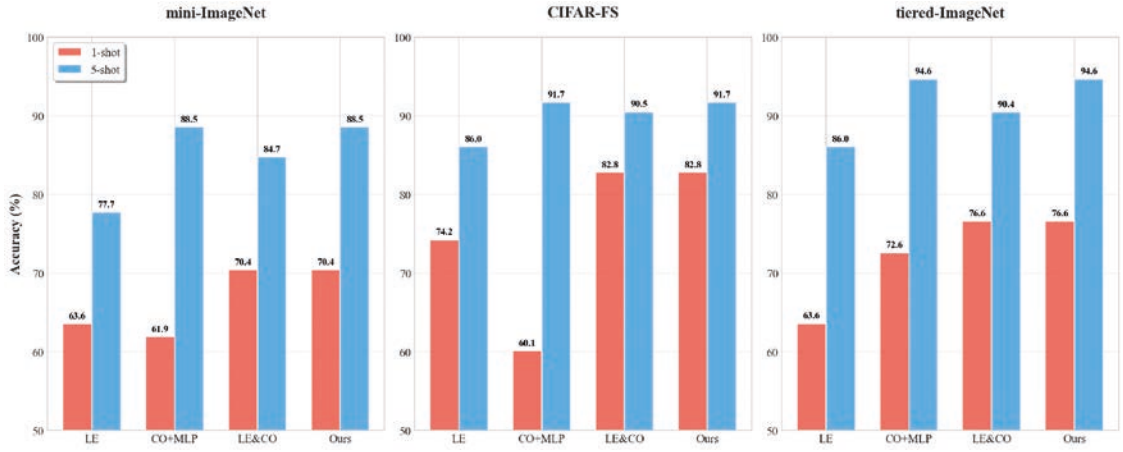

Figure 5. Ablation experiments on DMM on three datasets.

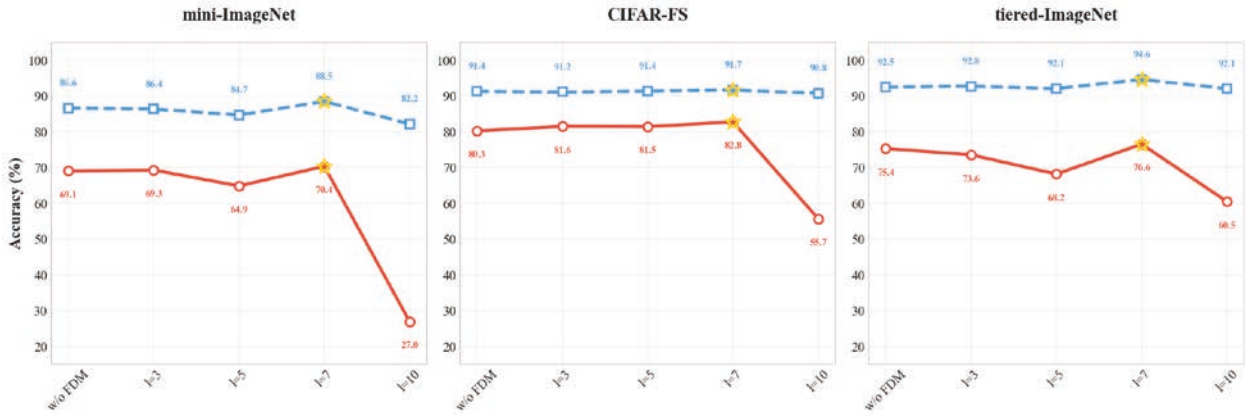

Figure 6. Ablation study on different layer configurations for few-shot classification accuracy (%). The blue line represents the 5-shot task, and the red line represents the 1-shot task.

key component. This section provides a comprehensive analysis of the model's explainability.

5.6.1 Consistency and distinctiveness of feature groups

VEDSNet is designed to learn disentangled yet complementary feature groups, each capturing a specific aspect of the visual structure. As illustrated in Figure 7, every feature group consistently attends to semantically coherent regions—such as edge contours, textures, or chromatic transitions—across different samples and even across classes. This consistency suggests that each group forms an internal "concept axis" that remains stable throughout the dataset, while distinctiveness across groups ensures that the FDM collectively covers a broad range of visual semantics. Qualitatively, this behavior resembles human perceptual grouping, where separate visual cues are integrated for holistic recognition.

5.6.2 Concept-Level visualization of feature groups

The FDM provides explainability at two complementary levels. At the global level, the aggregation of all feature groups produces an object-centric attention distribution (Figure 8), indicating that the model has learned to focus primarily on the main object. This global visualization reflects how VEDSNet forms a coherent high-level understanding of the target category. At the concept level, each feature group within the FDM acts as a weighted concept detector that selectively activates specific semantic components contributing to the overall recognition. As shown in Figure 9 and Figure 10, these groups highlight distinct, semantically stable subregions across instances, and their learned weights determine the relative contribution of each concept to the final prediction. This mechanism enables the model to decompose a complex object into multiple interpretable parts while maintaining discriminative focus through the adaptive weighting process. In this way, the combination of weighted concept activation and global attention aggregation yields explanations that are both structurally grounded and semantically meaningful—demonstrating that VEDSNet's explainability arises naturally from its architecture rather than from post-hoc visualization. While our visual explanations are attention-based, their semantic coherence across samples suggests more stable and interpretable concept attribution.

Figure 9. Visualization of the most important class-level conceptual features for bird species, showing that VEDSNet consistently attends to the most discriminative semantic regions—such as the head, wings, and neck patterns—that define each category.

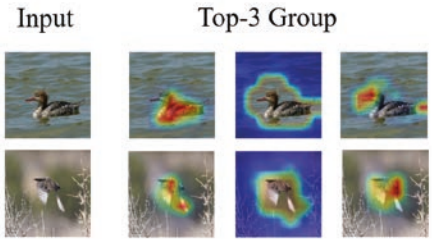

Figure 10. Visualization of the Top-3 activated feature groups for each sample, illustrating how VEDSNet integrates multiple complementary concept groups that attend to distinct semantic regions for holistic recognition.

5.6.3 Limitations of the aggregated visualization

While the aggregated visualization provides valuable explainability, it also reveals certain limitations inherent to the current architecture. As illustrated in Figure 11, the model occasionally produces misleading similarity responses when visually dominant but semantically irrelevant regions in the query and support images coincide. In such cases, the FDM may attend to large texture or shape patterns that appear visually correlated, causing the similarity metric to overemphasize superficial alignments rather than true semantic correspondence. This behavior can lead to incorrect associations between distinct object categories, particularly when the contextual appearance of one object partially resembles another under similar visual conditions. Moreover, the feature groups, while semantically coherent, may not always exhibit full independence. Overlapping activations

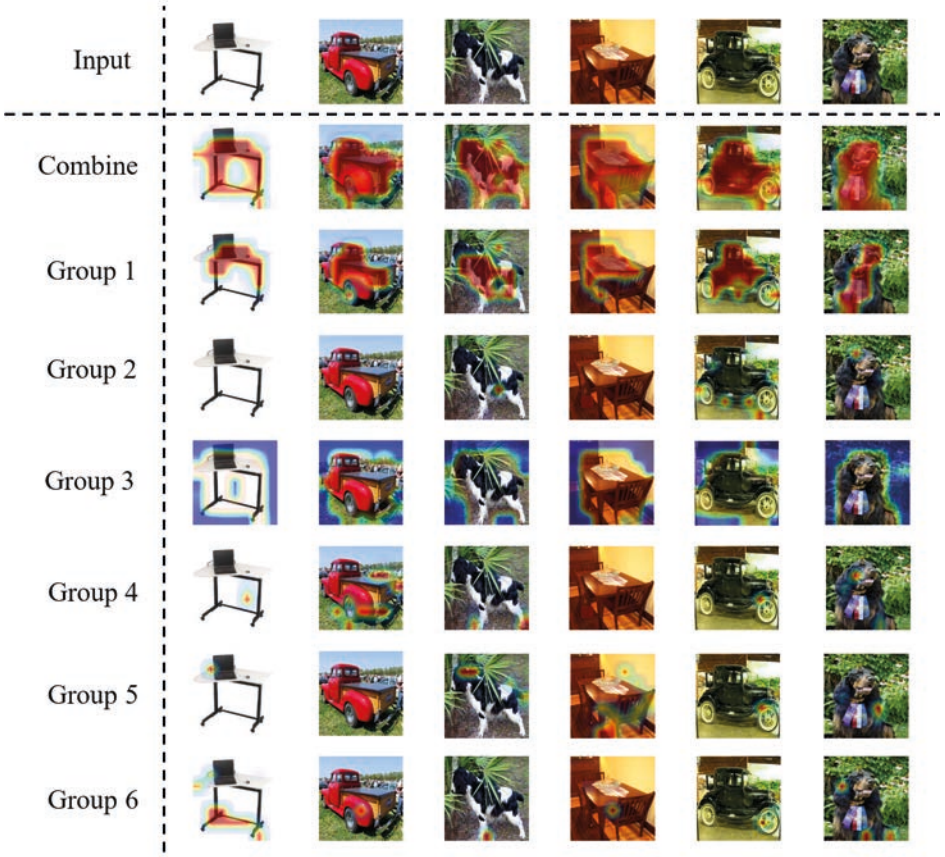

Figure 7. Feature group visualization shows the explainable decision-making process of the FDM, with different groups focusing on distinct visual attributes.

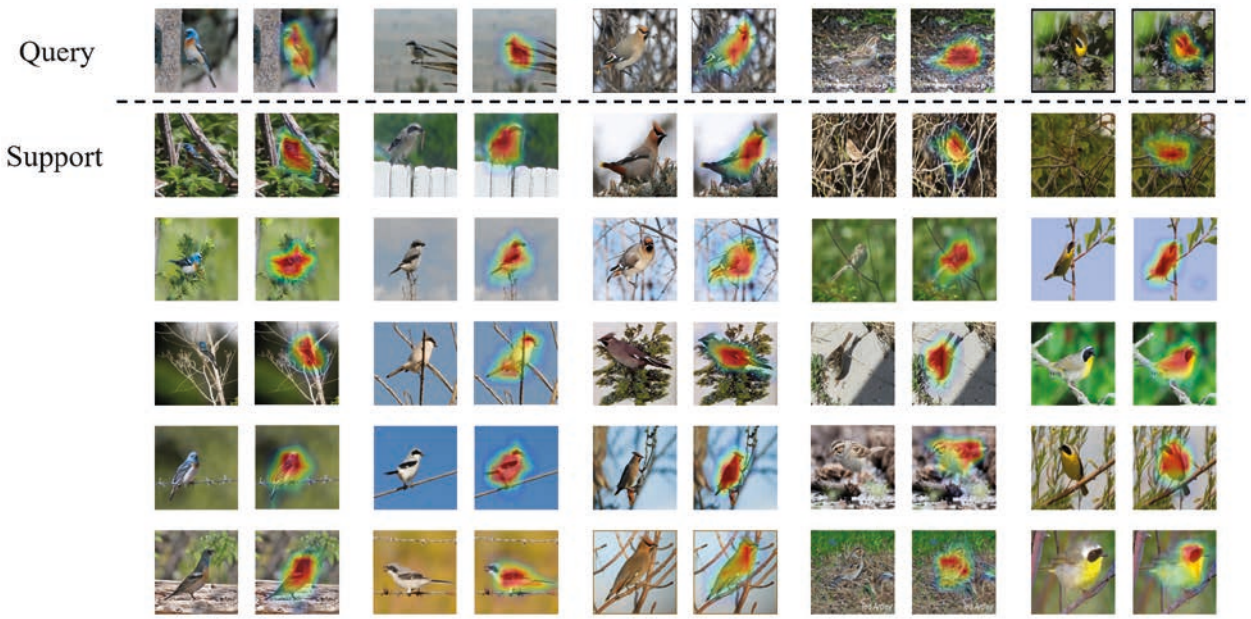

Figure 8. Heatmap visualization of feature similarities between support set and query images, demonstrating local and global similarity distributions used for classification decisions.

among groups can amplify the influence of ambiguous regions and further compound classification errors in visually complex or cluttered scenarios. Future work could address these issues by incorporating context-aware disentanglement objectives or contrastive regularization strategies that explicitly penalize spurious cross-category alignments. Such refinements would enhance the robustness and reliability of explainability in challenging few-shot settings.

5.6.4 Quantitative evaluation

To evaluate the explanatory characteristics of VEDSNet, we conducted comparative experiments against representative explainable AI (XAI) methods. We assessed not only task performance but also objective measures of explanation faithfulness and stability. For fairness, we used DMM and ViT-Tiny as the base architectures and applied established post-hoc XAI techniques on the same encoders wherever feasible. All methods followed matched optimization settings, data splits, and major hyperparameters. While perfect equivalence across pipelines is difficult in practice, this protocol aims to reduce confounding from backbone choice and training procedure so that differences more directly reflect the explanation mechanisms. Experiments were run on CUB over 2000 episodes for 5-way 5-shot tasks. Explanation quality was measured using standard objective metrics (e.g., IAUC and DAUC). Results in Table 8 indicate that VEDSNet achieves explanation performance that is competitive with post-hoc baselines under the controlled settings. Given the shared encoders and training configurations, observed gaps in IAUC/DAUC are more likely attributable to differences in the explanation mechanisms rather than gross disparities in model capacity; however, residual architectural bias cannot be fully ruled out. We evaluate the proposed model using three complementary metrics. Classification Accuracy measures the model's discriminative capability and serves as a baseline indicator of overall task performance. To assess the quality of visual explanations, we adopt two standardized interpretability metrics following Petsiuk et al. [39]. The Insertion Area Under the Curve (IAUC) quantifies how model performance recovers as pixels ranked by explanation-derived importance are progressively reinstated, where higher values indicate stronger alignment between highlighted regions and decision-critical features. Conversely, the Deletion Area Under the Curve (DAUC) measures the rate of accuracy degradation as high-importance pixels are sequentially removed, with lower scores reflecting more precise localization of discriminative evidence.

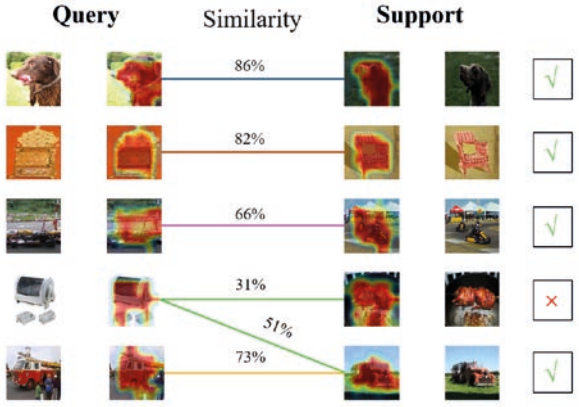

Figure 11. Visualization of similarity responses between query and support images. Correct matches exhibit high similarity and spatially consistent attention, while the failure case (fourth row) demonstrates that visually dominant yet semantically unrelated regions can cause misleading similarity activation, resulting in misclassification.

Table 8. Comparison of different methods on 5-way tasks on Mini-ImageNet dataset.

Method	Precision ↑	IAUC ↑	DAUC ↓
GradCAM [53]	0.8388	0.5989	0.2286
Score-CAM [54]	0.8252	0.5673	0.2487
Ours	**0.8550**	**0.6291**	**0.1943**

5.7 Comparative Analysis Of Distance Metrics

We hypothesize that learnable distance metrics provide robust stability for the VEDSNet model, while cosine similarity enhances overall classification accuracy. To substantiate this hypothesis, we conducted a systematic comparison of different distance metrics to assess their influence on classification performance, using the learnable distance metric as a baseline. The evaluation covered cosine similarity, Mahalanobis distance, and

Euclidean distance across 5-way 1-shot and 5-shot tasks on standard FSL benchmarks.

The results, summarized in Table 9, show that cosine similarity consistently achieves the best performance among all evaluated metrics. This advantage is attributed to its focus on directional alignment rather than magnitude, which proves beneficial in high-dimensional embedding spaces typical of FSL scenarios. The Mahalanobis distance demonstrates moderate performance, benefiting from its ability to capture inter-feature correlations, while the Euclidean distance performs the weakest due to its equal weighting of all feature dimensions, which limits its capacity to model the relative importance of features.

These observations support our design choice to incorporate cosine similarity within the dynamic hybrid metric framework of VEDSNet. By dynamically combining the stability of learnable distance metrics with the discriminative sensitivity of cosine similarity, our model achieves a balanced and robust metric representation, contributing to improved generalization and interpretability in FSL.

Table 9. Different metric methods on the Tiered-ImageNet dataset.

Method	1-shot	5-shot
Cosine Similarity	76.58±0.39	94.64±0.33
Euclidean Distance	73.83±0.43	92.78±0.36
Mahalanobis Distance	75.97±0.45	93.85±0.33

5.8 Fine-tuning Strategy

To systematically evaluate the impact of different training strategies on the performance of the VEDSNet framework, we conducted a comparative study between two common fine-tuning paradigms: (1) Linear Probing, which freezes the parameters of the feature extractor and the FDM while updating only the DMM; and (2) Full Fine-Tuning, which jointly optimizes all network parameters in an end-to-end manner. As shown in Table 10, the full fine-tuning approach consistently outperforms linear probing across both 1-shot and 5-shot settings. The observed improvement is mainly attributed to the adaptive refinement of feature representations enabled by joint parameter optimization.

When the target dataset—such as CIFAR-FS with its fine-grained classification categories—differs substantially from the source domain (e.g., general classification in ImageNet), merely updating the classification layer is insufficient to capture mid-level semantic variations. In contrast, full fine-tuning enables holistic parameter adaptation, allowing the model to dynamically recalibrate feature extraction and metric alignment across all layers, thereby improving generalization and robustness in FSL scenarios.

Table 10. Two different training strategies on the CIFAR-FS dataset.

Method	1-shot	5-shot
Linear probe	81.05±0.57	90.43±0.61
Fine-tune	82.78±0.45	91.71±0.36

Conclusion

In this study, we introduce VEDSNet, a model tailored for explainable FSL classification tasks. Evaluated on four benchmark datasets—Mini-ImageNet, CIFAR-FS, Tiered-ImageNet, and CUB-200-2011—VEDSNet achieves competitive classification performance relative to established FSL methods. The FDM is instrumental in enhancing prediction accuracy by extracting salient information from the backbone network, thereby enabling more precise and discriminative feature representations. Additionally, the DMM effectively mitigates overfitting, contributing to improved classification accuracy. Experimental results confirm the efficacy of these components in delivering robust and interpretable FSL performance. Moving forward, we plan to extend VEDSNet's applicability to a broader range of real-world data modalities, including video, text, and other heterogeneous data types, beyond its current focus on image-based tasks. This expansion aims to enhance the model's versatility and performance across diverse application domains, further contributing to the advancement of explainable FSL.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62266004) and the Open Project Program of Guangxi Key Laboratory of Digital Infrastructure (Grant Number: GXDINBC202401).

References

[1] F. Z. Mehrjardi, A. M. Latif, M. S. Zarchi, R. Sheikhpour, A survey on deep learning-based image forgery detection, Pattern Recognition, 144, 2023, 109778.

[2] X. Li, X. Yang, Z. Ma, J.-H. Xue, Deep metric learning for few-shot image classification: A review of recent developments, Pattern Recognition, 138, 2023, 109381.

[3] S. Lee, W. Moon, J.-P. Heo, Task discrepancy maximization for fine-grained few-shot classification, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 5331–5340.

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

[5] W. Chen, C. Si, Z. Zhang, L. Wang, Z. Wang, T. Tan, Semantic prompt for few-shot image recognition, arXiv preprint arXiv:2303.14123, 2023.

[6] B. Dong, P. Zhou, S. Yan, W. Zuo, Self-promoted supervision for few-shot transformer, In: European Conference on Computer Vision, Springer, 2022, 329–347.

[7] X. Zhang, D. Meng, H. Gouk, T. M. Hospedales, Shallow bayesian meta learning for real-world few-shot recognition, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 651–660.

[8] M. Hiller, R. Ma, M. Harandi, T. Drummond, Rethinking generalization in few-shot classification, In: Advances in Neural Information Processing Systems, 35, 2022, 3582–3595.

[9] F. Hao, F. He, L. Liu, F. Wu, D. Tao, J. Cheng, Class-aware patch embedding adaptation for few-shot image classification, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 18905–18915.

[10] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, L. Yuan, TinyViT: Fast pretraining distillation for small vision transformers, In: European Conference on Computer Vision, Springer, 2022, 68–85.

[11] P. Bateni, R. Goyal, V. Masrani, F. Wood, L. Sigal, Improved few-shot visual classification, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 14493–14502.

[12] A. Parnami, M. Lee, Learning from few examples: A summary of approaches to few-shot learning, arXiv preprint arXiv:2203.04291, 2022.

[13] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proceedings of the IEEE, 109(1), 2020, 43–76.

[14] H.-J. Ye, H. Hu, D.-C. Zhan, F. Sha, Few-shot learning via embedding adaptation with set-to-set functions, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 8808–8817.

[15] T. Ridnik, E. Ben-Baruch, A. Noy, L. Zelnik-Manor, ImageNet-21K Pretraining for the Masses, arXiv preprint arXiv:2104.10972, 2021.

[16] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin et al., Scaling vision transformers to 22 billion parameters, In: International Conference on Machine Learning, PMLR, 2023, 7480–7512.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, 248–255.

[18] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, R. S. Zemel, Meta-learning for semi-supervised few-shot classification, arXiv preprint arXiv:1803.00676, 2018.

[19] L. Bertinetto, J. F. Henriques, P. H. Torr, A. Vedaldi, Meta-learning with differentiable closed-form solvers, arXiv preprint arXiv:1805.08136, 2018.

[20] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, J.-B. Huang, A closer look at few-shot classification, arXiv preprint arXiv:1904.04232, 2019.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen et al., LoRA: Low-Rank Adaptation of Large Language Models, ICLR, 1(2), 2022, 3.

[22] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, In: International Conference on Machine Learning, PMLR, 2019, 2790–2799.

[23] D. Wertheimer, L. Tang, B. Hariharan, Few-shot classification with feature map reconstruction networks, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 8012–8021.

[24] C. Zhang, Y. Cai, G. Lin, C. Shen, DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers,

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 12203–12213.

[25] C. Cao, Y. Zhang, Learning to compare relation: Semantic alignment for few-shot learning, IEEE Transactions on Image Processing, 31, 2022, 1462–1474.

[26] G. Koch, R. Zemel, R. Salakhutdinov et al., Siamese neural networks for one-shot image recognition, In: ICML Deep Learning Workshop, 2, Lille, 2015, 1–30.

[27] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, In: Advances in Neural Information Processing Systems, 30, 2017.

[28] X. Li, J. Wu, Z. Sun, Z. Ma, J. Cao, J.-H. Xue, BSNet: Bi-Similarity Network for Few-shot Fine-grained Image Classification, IEEE Transactions on Image Processing, 30, 2020, 1318–1331.

[29] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra et al., Matching networks for one shot learning, In: Advances in Neural Information Processing Systems, 29, 2016.

[30] J. Wu, T. Zhang, Y. Zhang, F. Wu, Task-aware part mining network for few-shot learning, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 8433–8442.

[31] H. Chen, H. Li, Y. Li, C. Chen, Sparse spatial transformers for few-shot learning, Science China Information Sciences, 66(11), 2023, 210102.

[32] L. Li, B. Wang, M. Verma, Y. Nakashima, R. Kawasaki, H. Nagahara, SCOUTER: Slot Attention-based Classifier for Explainable Image Recognition, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 1046–1055.

[33] G. Qi, H. Yu, Z. Lu, S. Li, Transductive few-shot classification on the oblique manifold, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 8412–8422.

[34] J. Sun, J. Li, Few-shot classification with fork attention adapter, Pattern Recognition, 156, 2024, 110805.

[35] H. Cheng, S. Yang, J. T. Zhou, L. Guo, B. Wen, Frequency guidance matters in few-shot learning, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 11814–11824.

[36] Z. Sun, W. Zheng, P. Guo, M. Wang, TST_MFL: Two-stage training based metric fusion learning for few-shot image classification, Information Fusion, 113, 2025, 102611.

[37] Y. He, W. Liang, D. Zhao, H.-Y. Zhou, W. Ge, Y. Yu, W. Zhang, Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 9119–9129.

[38] T. Wu, T. Kobayashi, Geometric Mean Improves Loss For Few-Shot Learning, arXiv preprint arXiv:2501.14593, 2025.

[39] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, arXiv preprint arXiv:1806.07421, 2018.

[40] Y. Rong, X. Lu, Z. Sun, Y. Chen, S. Xiong, ESPT: A self-supervised episodic spatial pretext task for improving few-shot learning, In: Proceedings of the AAAI Conference on Artificial Intelligence, 37(8), 2023, 9596–9605.

[41] X. Li, Z. Li, J. Xie, X. Yang, J.-H. Xue, Z. Ma, Self-reconstruction network for fine-grained few-shot classification, Pattern Recognition, 153, 2024, 110485.

[42] Z. Li, Z. Hu, W. Luo, X. Hu, SaberNet: Self-attention based effective relation network for few-shot learning, Pattern Recognition, 133, 2023, 109024.

[43] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, K.-R. Müller, Explaining deep neural networks and beyond: A review of methods and applications, Proceedings of the IEEE, 109(3), 2021, 247–278.

[44] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, In: Advances in Neural Information Processing Systems, 35, 2022, 507–520.

[45] B. Wang, L. Li, Y. Nakashima, H. Nagahara, Learning bottleneck concepts in image classification, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 10962–10971.

[46] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, Data Mining and Knowledge Discovery, 37(5), 2023, 1719–1778.

[47] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset, California Institute of Technology, Technical Report CNS-TR-2011-001, 2011.

[48] J. Zeng, Z. Xue, L. Zhang, Q. Lan, M. Zhang, Multistage relation network with dual-metric for few-shot hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing, 61, 2023, 1–17.

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

[50] X. Li, Q. Song, J. Wu, R. Zhu, Z. Ma, J.-H. Xue, Locally-enriched cross-reconstruction for few-shot fine-grained image classification, IEEE Transactions on Circuits and Systems for Video Technology, 33(12), 2023, 7530–7540.

[51] Y. Liu, W. Zhang, C. Xiang, T. Zheng, D. Cai, X. He, Learning to affiliate: Mutual centralized learning for few-shot classification, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 14411–14420.

[52] J. Zhuang, T. Tang, Y. Ding, S. C. Tatikonda, N. Dvornek, X. Papademetris, J. Duncan, Adabelief optimizer: Adapting stepsizes by the belief in observed gradients, In: Advances in neural information processing systems, 33, 2020, 18795–18806.

[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, In: Proceedings of the IEEE international conference on computer vision, 2017, 618–626.

[54] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-cam: Score-weighted visual explanations for convolutional neural networks, In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, 24–25.

[55] J. Qian, T. Wen, M. Ling, X. Du, H. Ding, Pixel-based clustering for local interpretable model-agnostic explanations, Journal of Artificial Intelligence and Soft Computing Research, 15, 2025.

[56] H. Liu, C. Wang, X. Jiang, M. Khishe, A few-shot learning approach for covid-19 diagnosis using quasi-configured topological spaces, Journal of Artificial Intelligence and Soft Computing Research, 14(1), 2023, 77–95.

[57] N. Shakhovska, A. Shebeko, Y. Prykarpatskyy, A novel explainable ai model for medical data analysis, Journal of Artificial Intelligence and Soft Computing Research, 14, 2024.

Zirui Pei received the B.S. degree from Shenyang Normal University, Shenyang, China, in 2022. He is currently pursuing the M.S. degree at the School of Computer, Electronics and Information at Guangxi University, Nanning, China. His current research interests mainly focus on explainable artificial intelligence, few-shot learning, and multimodal learning.
https://orcid.org/0009-0006-4657-3893

Zuqiang Meng received his B.Sc. from Guangxi Normal University, China, in 1997, his M.Sc. from Xiangtan University, China, in 2000, and his Ph.D. from Central South University, China, in 2004. From March 2015 to March 2016, he was a visiting scholar at the University of Kansas, USA. He is currently a Professor in the College of Computer, Electronics and Information at Guangxi University, China. His research interests include cross-media intelligence, multimodal learning, interpretable deep learning, feature fusion, and granular computing.
https://orcid.org/0000-0002-4911-6958

Tingting Diao received the B.S. degree from Guangxi Normal University, Guilin, China, in 2023. She is currently pursuing the M.S. degree at the School of Computer, Electronics and Information at Guangxi University, Nanning, China.Her current research interests mainly focus on sparse mobile crowdsensing, spatiotemporal data inference and prediction, graph neural networks for urban sensing, and multi-scale temporal modeling.
https://orcid.org/0009-0000-0371-0823

Miao Peng received the B.Eng. degree in chemical engineering and technology from Zhejiang University, China, in 2018. He is currently pursuing the M.Sc. degree in artificial intelligence at Guangxi University, China, since 2023. His research interests include deep learning, trustworthy AI, and image generation.
https://orcid.org/0009-0006-2108-5025

Yifan Meng received the Bachelor of Science degree from Qilu University of Technology, Jinan, China, in 2023. He is currently pursuing the M.S. degree at the School of Computer, Electronics and Information, Guangxi University, Nanning, China. His current research interests mainly focus on multimodal classification models, explainable artificial intelligence, large models, and explainable machine learning algorithms for complex data systems.
https://orcid.org/0009-0006-0782-4880

Chaohong Tan, Deputy Director of Guangxi Key Laboratory of Digital Infrastructure, member of the Advisory Committee of Guangxi Electronic Society, and expert in the field of electronic government construction and management in Guangxi. In 1986, he received B.S. degree in Computer Software and Theory from Huazhong University of Technology. His research areas include computer software development, network construction, and system integration.
https://orcid.org/0009-0004-4117-731X