F2SST: Frequency-to-Spatial Semantic Transfer for Few-Shot Image Classification

Xueyi Chen, Bangjun Wang*, Jiaqing Fan, Li Zhang, Fanzhang Li

School of Computer Science & Technology, Soochow University, Suzhou, China
20245227095@stu.suda.edu.cn, {wangbangjun, jqfan, zhangliml, lfzh}@suda.edu.cn

Abstract

Few-shot image classification (FSIC) aims to recognize novel categories from only a few labeled examples, making it inherently challenging under limited supervision. Existing approaches have attempted to alleviate this issue by incorporating explicit semantics like class names or knowledge graphs to guide learning. However, such methods often encounter semantic ambiguity due to their dependence on either overly simplistic semantic priors or resource-intensive external knowledge sources, which limits their potential. In this paper, we explore the frequency domain as an implicit and task-adaptive source of semantic information. We propose F2SST, a Frequency-to-Spatial Semantic Transfer framework that enhances feature learning by leveraging spectral signals as hidden semantics. Specifically, F2SST applies Fast Fourier Transform (FFT) to extract phase-invariant global frequency descriptors, followed by a lightweight Gated Spectral Attention (GSA) module that selectively emphasizes class-relevant frequency components. These enhanced spectral cues are then integrated into the spatial stream through a class-guided fusion mechanism, enabling more robust and semantically aligned representations. Extensive experiments on four standard benchmarks (miniImageNet, tieredImageNet, CIFAR-FS and FC100) demonstrate that F2SST consistently improves performance, validating the effectiveness of frequency-domain semantics in FSIC.

Code — https://github.com/cxy1100/F2SST

Introduction

Deep learning has demonstrated remarkable success in a variety of computer vision applications, particularly in the area of image classification (Hu and Ma 2022; Bär et al. 2024), its heavy dependence on large labeled datasets presents a significant limitation. Unlike humans who can learn new visual concepts from minimal examples, current deep learning systems struggle in low-data regimes. To address this fundamental challenge, the computer vision community has increasingly focused on few-shot learning approaches (Afrasiyabi, Lalonde, and Gagné 2021; Xu and Le 2022; Zhang et al. 2024) that enable effective learning from scarce labeled data.

*Corresponding author.

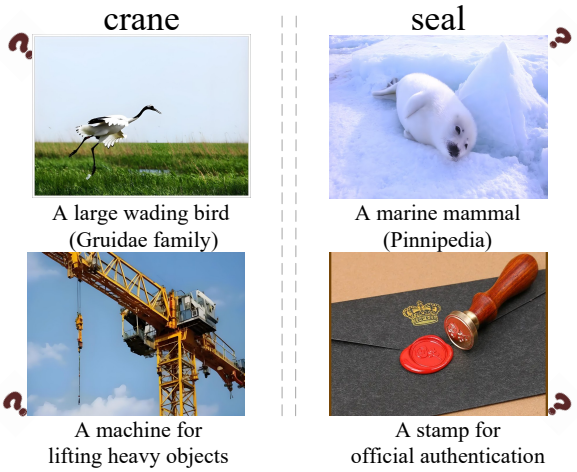


Figure 1: Illustrative examples of semantic ambiguity. The polysemous words “crane” and “seal” are presented with their distinct meanings. These examples demonstrate the semantic ambiguity in textual class names.

Existing works have explored various strategies: metric-based methods (Snell, Swersky, and Zemel 2017; Chen et al. 2019) compare support-query pairs in a learned embedding space; transformer-based models (Hiller et al. 2022; Hao et al. 2023) refine patch-level interactions; and semantic-enhanced methods (Chen et al. 2023) introduce external class descriptions or knowledge graphs to guide feature learning. Semantic-based solutions (Zhang et al. 2021a; Xu and Le 2022) aim to inject higher-level understanding to make up for missing visual diversity. However, they often rely on naive or handcrafted semantics such as textual class names (e.g., “crane” or “seal”) which introduce unavoidable ambiguity. As shown in Figure 1, “crane” may refer to a bird or a construction machine; “seal” might imply either an animal or a stamp. These ambiguities can easily confuse models, especially under few-shot conditions where contextual cues are limited. Moreover, acquiring high-quality semantics through human annotation or external knowledge sources incurs additional cost and may not generalize across domains.

Meanwhile, frequency-domain modeling has been explored in tasks such as dense prediction (Chen et al. 2024), multispectral fusion (Liang et al. 2024), and crowd counting (Chaudhuri et al. 2024), where it improves global consistency and suppresses noise to alleviate semantic ambiguity. While these methods show that spectral features can aid spatial learning, frequency signals are often treated as auxiliary cues and still rely on dense supervision. In contrast, we view low-frequency components as a form of implicit semantics, which serves as a latent prior that captures object-level structures without requiring explicit annotation. This perspective motivates us to investigate how such hidden semantics can be transferred into the spatial domain to support few-shot classification under limited data.

In this paper, we propose F2SST, a Frequency-to-Spatial Semantic Transfer framework that introduces frequency-domain cues as implicit and task-adaptive semantics to enhance representation learning in FSIC. Specifically, F2SST applies Fast Fourier Transform (FFT) to extract phase-invariant global descriptors, followed by a Gated Spectral Attention (GSA) module that selectively emphasizes class-relevant frequency components. These spectral cues are then integrated into the spatial stream through a class-guided fusion mechanism, resulting in more robust and semantically aligned representations under few-shot conditions.

Our contributions can be summarized as follows:

- We propose F2SST, a novel Frequency-to-Spatial Semantic Transfer framework that introduces frequency-domain cues as implicit semantics into few-shot image classification.
- We introduce a lightweight Gated Spectral Attention (GSA) and a spatial fusion with class guidance to selectively emphasize semantically informative spectral components, effectively complementing spatial representations.
- We conduct a systematic analysis of frequency selection strategies and find that low-frequency components consistently outperform other alternatives in semantic alignment and generalization.
- Extensive experiments are conducted on four benchmark datasets to investigate the effectiveness of F2SST, and F2SST achieves state-of-the-art performance.

Related Work

Few-Shot Image Classification. Few-shot image classification (FSIC) focuses on recognizing novel categories using only a few labeled examples. Existing methods can be broadly categorized into three lines: metric-based, optimization-based, and hallucination-based approaches. Metric-based methods aim to learn an embedding space where affinity between support and query samples is measured via distance metrics (Snell, Swersky, and Zemel 2017; Chen et al. 2019). Beyond simple first-order statistics, existing works enrich feature distributions by incorporating second-order moments (Zhang et al. 2019; Xie et al. 2022), or by designing task-adaptive similarity functions (Liu, Cao, and He 2023; Hu et al. 2023). Optimization-based methods seek to learn meta-initializations or adaptation rules that generalize rapidly to new tasks with minimal gradient steps (Finn, Abbeel, and Levine 2017; Yu et al. 2022; Sun and Gao 2023). Hallucination-based methods address data scarcity by generating synthetic examples at either the image level (Zhang et al. 2018) or the feature level (Lazarou, Stathaki, and Avrithis 2022; Bär et al. 2024).

Frequency-Based Visual Representation. The frequency domain provides a complementary perspective to spatial representations by capturing structural and global patterns across different scales. Earlier studies have demonstrated that frequency filtering techniques, such as those based on FFT or DCT, can effectively replace computationally intensive attention mechanisms or enhance convolutional feature quality (Rao et al. 2021; Qin et al. 2021). Subsequent research has expanded the application of frequency modeling to dense prediction and multi-modal fusion tasks. FFT-based token mixers have been proposed as efficient alternatives to spatial self-attention for global token interactions (Tatsunami and Taki 2024). Frequency-phase decoupling mechanisms have been introduced in multispectral and hyperspectral fusion to preserve spatial fidelity, while frequency-aware attention modules have shown improvements in structural consistency for tasks such as crowd counting and semantic segmentation (Liang et al. 2024; Chaudhuri et al. 2024; Chen et al. 2024). These findings collectively highlight the effectiveness of spectral information in capturing high-level semantics and improving representational robustness. In contrast to prior approaches that treat frequency and spatial features separately, our method integrates both domains to inject implicit, task-adaptive semantics into few-shot learning.

Vision Transformers in Few-Shot Image Classification. Vision Transformer (ViT) have demonstrated remarkable capabilities in modeling long-range dependencies (Dosovitskiy et al. 2021; Liu et al. 2021b), yet their effectiveness in few-shot settings is constrained by the lack of inherent inductive biases and dependence on large-scale training data (Liu et al. 2021a). Some advances address these limitations through self-supervised pretraining strategies and specialized few-shot adaptations like FewTURE (Hiller et al. 2022) and CPEA (Hao et al. 2023), which leverage token-level dynamics to improve generalization from limited examples. These developments highlight the potential of transformer architectures when enhanced with mechanisms for capturing both semantics and feature relationships.

Method

Problem Definition

Few-shot image classification aims to recognize novel classes using only a few labeled examples per class. Formally, we assume a label space partitioned into disjoint training and testing classes, i.e., $\mathcal{C}_{\text{train}} \cap \mathcal{C}_{\text{test}} = \emptyset$. The model is trained on $\mathcal{C}_{\text{train}}$ and evaluated on $\mathcal{C}_{\text{test}}$, which contains previously unseen categories with limited supervision. Following standard practice (Vinyals et al. 2016), we formulate FSIC in an episodic manner. Each episode consists of a support set $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N \times K}$ and a query set $\mathcal{Q} =$

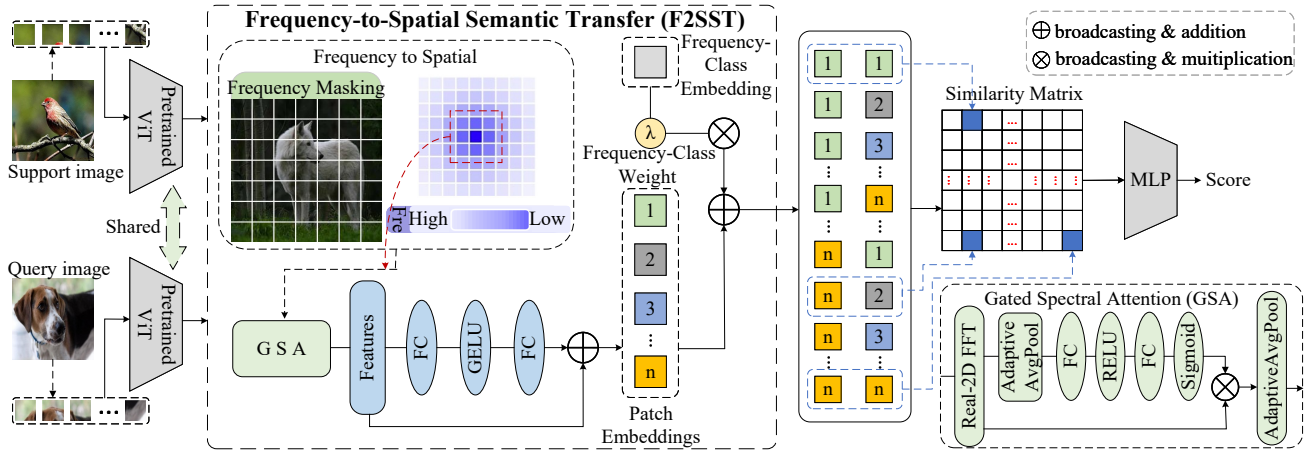


Figure 2: Overview of the F2SST framework. The ViT backbone is pretrained using a self-supervised masked image modeling framework to obtain generalizable patch representations. Given support and query images, spatial patch embeddings are first extracted via the backbone. FFT is applied to capture global frequency features, which are refined by a Gated Spectral Attention (GSA) module. A spatial fusion with class guidance integrates frequency cues back into the spatial domain. Finally, a similarity matrix between support and query patches is computed and processed by a multi-layer perceptron (MLP) for classification.

$\{(x_j^q, y_j^q)\}_{j=1}^{N\times Q}$, where N is the number of classes (N-way), K is the number of labeled support examples per class (K-shot), and Q is the number of query samples per class. The model leverages the support set $\mathcal{S}$ to construct class-level representations and predict the labels of query images in $\mathcal{Q}$. The goal is to train the model using episodes sampled from $\mathcal{C}_{\text{train}}$, such that it generalizes well to episodes drawn from $\mathcal{C}_{\text{test}}$, within the inductive few-shot learning framework.

Overview

The overall pipeline of our proposed F2SST is illustrated in Figure 2. Given a few-shot task composed of a support set and a query set, we employ a pretrained ViT backbone to extract spatial patch embeddings for both. To latent semantic relationships, we transform the patch tokens into the frequency domain using 2D FFT. A Gated Spectral Attention (GSA) module is then applied to selectively enhance class-relevant frequency components while suppressing noisy signals. The refined frequency features are projected back to spatial domain through spatial fusion with class guidance, creating a symbiotic representation where local spatial details are enhanced by global semantic understanding. For classification, we construct a similarity matrix by computing patch-wise interactions between query and support features. This matrix is then flattened and processed through a lightweight MLP to produce the final similarity scores for prediction. By unifying spectral modeling, fusion, and dense similarity reasoning in an end-to-end framework, F2SST naturally encodes hierarchical semantic relationships through frequency-guided feature learning.

Frequency-to-Spatial Semantic Transfer

To enhance representation robustness in low-data regimes, we propose a novel Frequency-to-Spatial Semantic Transfer, which integrates real-valued Fourier representations and a gated spectral attention mechanism to inject implicit semantics into patch-level tokens, as shown in Algorithm 1.

2D Fourier Transform. Given input spatial features $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, we transform them into the frequency domain using the 2D Fast Fourier Transform (FFT). For each channel, the frequency component at location (u, v) is given by:

$$X_f[u,v] = \sum_{m=0}^{H-1}\sum_{n=0}^{W-1} x[m,n]e^{-j2\pi\left(\frac{um}{H}+\frac{vn}{W}\right)}, \quad (1)$$

where (m, n) and (u, v) denote spatial and frequency indices respectively, and j is the imaginary unit. This transformation decomposes the image into orthogonal frequency bases.

Real-Spectrum Projection. To efficiently leverage frequency information while avoiding the complexity of handling complex-valued tensors, we retain only the real part of the 2D FFT spectrum, following the design principle in (Lee-Thorp et al. 2022). This real-spectral projection preserves the most critical structural patterns and eliminates unnecessary imaginary components. To further condense the frequency response into a compact descriptor, we apply global average pooling across the spatial dimensions, yielding a channel-level summary vector that encodes the global distribution of spectral energy.

Gated Spectral Attention. To dynamically emphasize task-relevant frequency channels, we adopt a lightweight channel-wise gating mechanism. The pooled spectrum is first passed through a two-layer fully connected network with ReLU and sigmoid activations, generating soft gates that modulate the importance of each channel. These gates

are then applied multiplicatively to the real-valued spectrum, selectively suppressing noisy or redundant components while preserving salient frequency cues. The filtered spectrum is subsequently projected back to the token space and fused with the original spatial embeddings to enrich downstream representation learning.

Spatial Fusion with Class Guidance. We enhance the spatial sequence $\mathbf{X}_{\text{spatial}}$ by fusing the gated frequency embeddings using a residual MLP:

$$\mathbf{X}_{\text{fused}} = \text{LN}\left(\mathbf{X}_{\text{spatial}} + \text{MLP}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\right)\right), \quad (2)$$

followed by class-guided modulation to boost discrimination:

$$\begin{aligned}\mathbf{X}_{\text{final}}^{(q)} &= \mathbf{X}_{\text{fused}}^{(q)} + \lambda \cdot \mathbf{c}^{(q)},\\ \mathbf{X}_{\text{final}}^{(s)} &= \mathbf{X}_{\text{fused}}^{(s)} + \lambda \cdot \mathbf{c}^{(s)},\end{aligned} \quad (3)$$

where $\lambda = 1.0$ and $\mathbf{c}$ is the frequency-class embedding vector for query/support. This reinforces the consistency between class context and spatial content.

Similarity Matrix Given the frequency-enhanced embeddings from support and query samples, we compute a similarity matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$, where each element measures the similarity between a pair of patch tokens across images:

$$\mathbf{S}_{ij} = \left(d(\mathbf{z}_i^{(S)}, \mathbf{z}_j^{(Q)})\right)^2, \quad (4)$$

where $d(\cdot,\cdot)$ denotes the cosine similarity, and $\mathbf{z}_i^{(S)}$, $\mathbf{z}_j^{(Q)}$ are patch embeddings from support and query images, respectively. This patch-wise dense matching preserves fine-grained structural relationships and avoids reliance on predefined prototypes. The squared similarity emphasizes discriminative differences while remaining norm-invariant.

Training

In F2SST, we follow an episodic training paradigm for few-shot learning. For the c-th support class, there are K labeled images. After computing pairwise similarity scores between the j-th query image and all support images, we first aggregate the similarity score for the j-th query image with respect to the c-th class:

$$s_{cj} = \sum_{k=1}^{K} s_{cjk}, \quad (5)$$

Here, s_{cjk} denotes the similarity between the j-th query image and the k-th support image in class c. We then convert the aggregated similarity scores into class probabilities via a softmax function:

$$p_{cj} = \frac{\exp(s_{cj})}{\sum_{c'=1}^{C}\exp(s_{c'j})}, \quad (6)$$

where C is the number of classes per episode (i.e., the C-way classification setting). The training loss for one episode is computed using cross-entropy between the predicted probabilities and ground-truth labels of query samples:

$$\mathcal{L} = -\frac{1}{CQ}\sum_{j=1}^{CQ}\sum_{c=1}^{C}\mathbb{I}(y_j^{(Q)} = c)\log p_{cj}, \quad (7)$$

In this expression: $y_j^{(Q)}$ is the ground-truth label of the j-th query image. $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the condition holds and 0 otherwise. Q is the number of query images per class in each episode. All trainable parameters, including those in the Gated Spectral Attention module, MLP layers, and fusion units, are optimized by minimizing the loss $\mathcal{L}$ across episodic batches sampled from the training set.

Algorithm 1: F2SST Inference Pipeline for Few-Shot Image Classification

Require: Support set $S = \{(x_i^s, y_i^s)\}_{i=1}^{N\times K}$, Query set $Q = \{x_j^q\}_{j=1}^{N_q}$, Backbone encoder $f(\cdot)$, Spatial Fusion with Class Guidance $\mathcal{F}$, Similarity module $\mathbf{S}$

Ensure: Predicted labels $\{\hat{y}_j^q\}$ for each query image

1: **Extract embeddings:** Use the backbone $f(\cdot)$ to extract spatial features for both support and query images.
2: **Apply FFT:** Transform support and query features to the frequency domain using 2D FFT.
3: **Frequency enhancement:** Use the Gated Spectral Attention (GSA) to emphasize task-relevant frequency channels.
4: **Fusion:** Integrate frequency-enhanced features into spatial domain via $\mathcal{F}$.
5: **Dense similarity:** Compute a dense similarity matrix $\mathbf{S}$ between all query-support patch embeddings using Eq. (4).
6: **Score aggregation:** Flatten $\mathbf{S}$ and feed into a lightweight MLP to obtain similarity scores per query-support pair.
7: **Prediction:** For each query, assign the label of the support class with the highest similarity score.
8: **return** Predicted labels $\{\hat{y}_j^q\}$

Inference

During inference, we process episodes sampled from unseen test classes. For each query image, we compute the class probabilities p_{cj} using Eq. (6). The predicted class for the query image is then assigned as:

$$\hat{y}_j = \arg\max_c p_{cj}, \quad (8)$$

Once trained on the training set, F2SST generalizes directly to novel classes without requiring additional adaptation. In contrast to methods like FewTURE(Hiller et al. 2022) that require test-time optimization, F2SST relies solely on the pre-learned frequency-enhanced spatial features, enabling efficient and scalable one-pass inference. This makes our model well-suited for real-world applications with low-latency constraints.

$$(\mathcal{F}^{(f)}, \mathcal{F}^{(s)}) = \text{Split}(\mathcal{F}^{\text{input}}), \quad (9)$$

Model	Backbone	miniImageNet		tieredImageNet	
		1-shot	5-shot	1-shot	5-shot
ProtoNet (Snell, Swersky, and Zemel 2017)	ResNet-12	62.29 ± 0.33	79.46 ± 0.48	68.25 ± 0.23	84.01 ± 0.56
IEPT (Zhang et al. 2021c)	ResNet-12	67.05 ± 0.44	82.90 ± 0.30	72.24 ± 0.50	86.73 ± 0.34
MELR (Fei et al. 2021)	ResNet-12	67.40 ± 0.43	83.40 ± 0.28	72.14 ± 0.51	87.01 ± 0.35
DMF (Xu et al. 2021)	ResNet-12	67.76 ± 0.46	82.71 ± 0.31	71.89 ± 0.52	85.96 ± 0.35
CNL (Zhao et al. 2021)	ResNet-12	67.96 ± 0.98	83.36 ± 0.51	73.42 ± 0.95	87.72 ± 0.75
PAL (Ma et al. 2021)	ResNet-12	69.37 ± 0.64	84.40 ± 0.44	72.25 ± 0.72	86.95 ± 0.47
COSOC (Luo et al. 2021)	ResNet-12	69.28 ± 0.49	85.16 ± 0.42	73.57 ± 0.43	87.57 ± 0.10
Meta DeepBDC (Xie et al. 2022)	ResNet-12	67.34 ± 0.43	84.46 ± 0.28	72.34 ± 0.49	87.31 ± 0.32
SVAE-Proto (Xu and Le 2022)	ResNet-12	74.84 ± 0.23	83.28 ± 0.40	76.98 ± 0.65	85.77 ± 0.50
FGFL (Cheng et al. 2023)	ResNet-12	69.14 ± 0.80	86.01 ± 0.62	73.21 ± 0.88	87.21 ± 0.61
RENet-ventral (Dong, Zhai, and Zha 2023)	ResNet-12	69.71 ± 0.45	84.23 ± 0.29	73.94 ± 0.48	87.15 ± 0.35
BANs_SLA (Liu et al. 2024)	ResNet-12	70.40 ± 0.44	85.31 ± 0.22	73.85 ± 0.49	87.72 ± 0.33
SemFew (Zhang et al. 2024)	ResNet-12	77.63 ± 0.63	83.04 ± 0.48	78.96 ± 0.80	85.88 ± 0.62
FEAT (Ye et al. 2020)	WRN-28-10	65.10 ± 0.20	81.11 ± 0.14	70.41 ± 0.23	84.38 ± 0.16
PSST (Chen et al. 2021)	WRN-28-10	64.16 ± 0.44	80.64 ± 0.32	–	–
MetaQDA (Zhang et al. 2021d)	WRN-28-10	67.83 ± 0.64	84.28 ± 0.69	74.33 ± 0.65	89.56 ± 0.79
OM (Qi et al. 2021)	WRN-28-10	66.78 ± 0.30	85.29 ± 0.41	71.54 ± 0.29	87.79 ± 0.46
noHub-S (Trosten et al. 2023)	WRN-28-10	82.00 ± 0.26	88.03 ± 0.13	82.85 ± 0.27	88.31 ± 0.16
SIFT (Pan, Xin, and Shen 2024)	WRN-28-10	86.95 ± 0.53	89.22 ± 0.54	77.86 ± 0.77	89.65 ± 0.52
SUN (Dong et al. 2022)	ViT-S/16	67.80 ± 0.45	83.25 ± 0.30	72.99 ± 0.50	86.74 ± 0.33
FewTURE (Hiller et al. 2022)	ViT-S/16	68.02 ± 0.88	84.51 ± 0.53	72.96 ± 0.92	86.43 ± 0.67
CPEA (Hao et al. 2023)	ViT-S/16	71.97 ± 0.65	87.06 ± 0.38	76.93 ± 0.70	90.12 ± 0.45
F2SST (ours)	ViT-S/16	**89.27 ± 0.56**	**97.41 ± 0.18**	**85.25 ± 0.75**	**95.73 ± 0.29**

Table 1: Few-shot classification accuracies (with 95% CI) for the 5-way 1-shot and 5-way 5-shot settings on miniImageNet and tieredImageNet.

where $\mathcal{F}^{(f)} \in \mathbb{R}^{\alpha C \times H \times W}$ and $\mathcal{F}^{(s)} \in \mathbb{R}^{(1-\alpha) C \times H \times W}$ denote the frequency and spatial feature partitions, respectively, with α as a predefined channel split ratio.

Experiments

Datasets

We conduct comprehensive evaluations of our proposed method on four widely-used few-shot classification benchmarks: miniImageNet (Vinyals et al. 2016), tieredImageNet (Ren et al. 2018), CIFAR-FS (Bertinetto et al. 2019), and FC100 (Oreshkin, Rodríguez López, and Lacoste 2018). In line with established protocols (Ye et al. 2020; Hiller et al. 2022), each dataset is partitioned into mutually exclusive training, validation, and test subsets. The label spaces across these splits are non-overlapping, ensuring that categories encountered during training are entirely distinct from those used during evaluation.

Implementation Details

Backbone Pretraining. We adopt ViT-Small as the backbone and initialize it using iBOT (Zhou et al. 2022), a self-supervised masked image modeling approach. The teacher network receives full images, while the student learns to reconstruct masked patch features. Both networks are trained with Adam (Kingma and Ba 2014) optimizer (batch size 512) for 1600 epochs on miniImageNet, CIFAR-FS, and FC100, and 800 epochs on tieredImageNet. Standard augmentations such as random cropping and color jittering are applied during pretraining.

Training. Following the episodic meta-learning paradigm, we train the model under the standard N-way K-shot setting. Each episode consists of N support classes, each with K labeled samples and 15 query samples. We sample 100,000 episodes from the training set. Optimization is performed using Adam with an initial learning rate of 1×10^{-5} and weight decay of 0.001. The learning rate is halved every 5,000 episodes.

Evaluation. We report mean classification accuracy and 95% confidence intervals over 1,000 randomly sampled test episodes under the 5-way 1-shot and 5-shot settings, following the common practice (Zhang et al. 2020; Ye et al. 2020). All models are implemented in PyTorch and trained using two NVIDIA V100 GPUs.

Experimental Results

State-of-the-art Comparison. Table 1 and Table 2 show the comparison results for the 5-way 1-shot and 5-way 5-shot settings on four benchmark datasets respectively. On MiniImageNet, F2SST surpasses SIFT (Pan, Xin, and Shen 2024), the current best method, by 2.32% for 1-shot classification and 8.19% for 5-shot classification. On TieredImageNet, F2SST exceeds noHub-S (Trosten et al. 2023) by 2.40% in 1-shot classification and improves upon

Model	Backbone	CIFAR-FS		FC100	
		1-shot	5-shot	1-shot	5-shot
ProtoNet (Snell, Swersky, and Zemel 2017)	ResNet-12	–	–	41.54 ± 0.76	57.08 ± 0.76
MetaOptNet (Lee et al. 2019)	ResNet-12	72.80 ± 0.70	84.30 ± 0.50	47.20 ± 0.60	55.50 ± 0.60
MABAS (Kim, Kim, and Kim 2020)	ResNet-12	73.51 ± 0.92	85.65 ± 0.65	42.31 ± 0.75	58.16 ± 0.78
RFS (Tian et al. 2020)	ResNet-12	73.90 ± 0.80	86.90 ± 0.50	44.60 ± 0.70	60.90 ± 0.60
BML (Zhou et al. 2021)	ResNet-12	73.45 ± 0.47	88.04 ± 0.33	–	–
Meta-NVG (Zhang et al. 2021b)	ResNet-12	74.63 ± 0.91	86.45 ± 0.59	46.40 ± 0.81	61.33 ± 0.71
RENet (Kang et al. 2021)	ResNet-12	74.51 ± 0.46	86.60 ± 0.32	–	–
TPMN (Wu et al. 2021)	ResNet-12	75.50 ± 0.90	87.20 ± 0.60	46.93 ± 0.71	63.26 ± 0.74
RENet-ventral (Dong, Zhai, and Zha 2023)	ResNet-12	75.82	87.45	–	–
BANs_SLA (Liu et al. 2024)	ResNet-12	77.98 ± 0.45	89.78 ± 0.31	–	–
SemFew (Zhang et al. 2024)	ResNet-12	83.65 ± 0.70	87.66 ± 0.60	54.36 ± 0.71	62.79 ± 0.74
PSST (Chen et al. 2021)	WRN-28-10	77.02 ± 0.38	88.45 ± 0.35	–	–
Meta-QDA (Zhang et al. 2021d)	WRN-28-10	75.83 ± 0.88	88.79 ± 0.75	–	–
SIFT (Pan, Xin, and Shen 2024)	WRN-28-10	81.35 ± 0.75	89.22 ± 0.54	–	–
SUN (Dong et al. 2022)	ViT-S/16	78.37 ± 0.46	88.84 ± 0.32	–	–
FewTURE (Hiller et al. 2022)	ViT-S/16	76.10 ± 0.88	86.14 ± 0.64	46.20 ± 0.79	63.14 ± 0.73
CPEA (Hao et al. 2023)	ViT-S/16	77.82 ± 0.66	88.98 ± 0.45	47.24 ± 0.58	65.02 ± 0.60
F2SST (ours)	ViT-S/16	**91.21 ± 0.53**	**97.71 ± 0.19**	**54.46 ± 0.78**	**77.17 ± 0.58**

Table 2: Few-shot classification accuracies (with 95% CI) for the 5-way 1-shot and 5-way 5-shot settings on CIFAR-FS and FC100.

FFT	Fusion	GSA	1-shot	5-shot
✗	✗	✗	72.59 ± 0.63	88.17 ± 0.36
✓	✗	✗	86.15 ± 0.59	96.51 ± 0.22
✓	✓	✗	86.28 ± 0.60	95.52 ± 0.25
✓	✓	✓	**89.17 ± 0.56**	**97.35 ± 0.19**

Table 3: Ablation on FFT modeling components on miniImageNet (5-way classification). 'FFT' denotes using frequency-domain features, 'Fusion' refers to frequency-to-spatial integration, and 'GSA' is our Gated Spectral Attention (All Frequencies).

CPEA (Hao et al. 2023) by 5.61% in 5-shot classification. On CIFAR-FS, F2SST achieves accuracy gains of 7.56% over SemFew (Zhang et al. 2024) for 1-shot and 7.93% over BANs_SLA (Liu et al. 2024) for 5-shot classification. Although FC100 presents a particularly challenging scenario, F2SST still shows consistent improvements, marginally surpassing SemFew by 0.10% in 1-shot classification while delivering a substantial 12.15% advantage over CPEA in 5-shot classification. These comprehensive results validate that F2SST consistently outperforms previous approaches across all benchmark datasets by effectively introducing the implicit semantics of frequency domain.

Ablation Study

To further analyze the contribution of each component in our F2SST, we conduct a series of ablation studies on the miniImageNet(Vinyals et al. 2016) dataset under the 5-way classification setting.

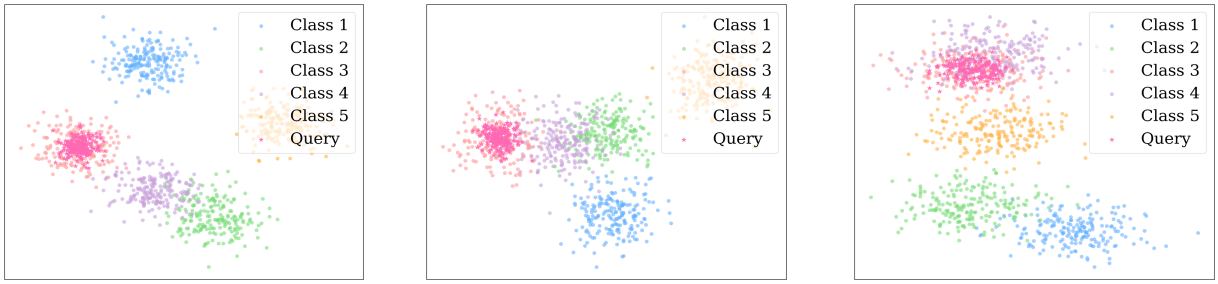
Frequency Selection	1-shot	5-shot
All Frequencies	89.17 ± 0.56	97.35 ± 0.19
Low-Frequency	**89.27 ± 0.56**	**97.41 ± 0.18**
High-Frequency	86.86 ± 0.61	96.74 ± 0.21
Center Block	89.25 ± 0.56	97.40 ± 0.18
Band-Pass	86.70 ± 0.62	96.77 ± 0.20

Table 4: Ablation on frequency selection strategies using F2SST on miniImageNet.

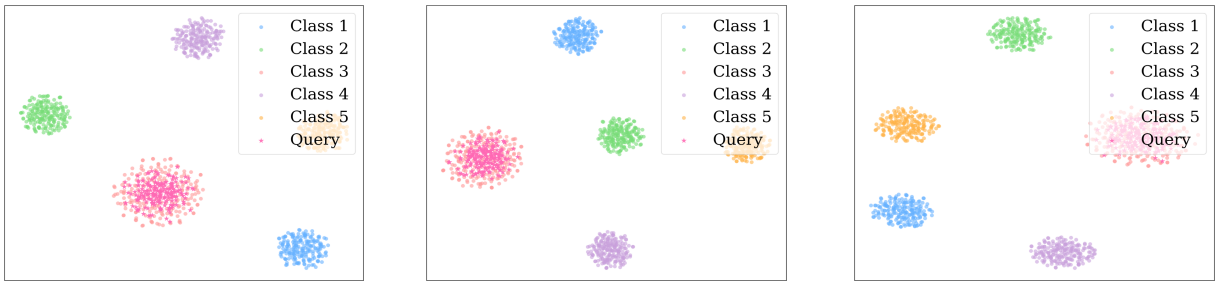
Frequency-Class Weight	1-shot	5-shot
$\lambda = 0.0$	71.51 ± 0.65	86.17 ± 0.40
$\lambda = 0.5$	89.10 ± 0.56	**97.36 ± 0.18**
$\lambda = 1.0$	**89.17 ± 0.56**	97.35 ± 0.19
$\lambda = 2.0$	89.09 ± 0.56	97.04 ± 0.20
$\lambda = 3.0$	89.09 ± 0.57	96.91 ± 0.21

Table 5: Performance comparison with different Frequency-Class Weight λ (All Frequencies).

Frequency Modeling Components. As shown in Table 3, incorporating FFT alone improves performance by capturing global spectral patterns that are more robust to intra-class variations. Adding the fusion module further enhances results by integrating frequency and spatial representations, enabling complementary feature learning. The GSA module brings the most significant individual gain by selectively emphasizing informative frequency channels through learnable gating. These findings validate the effectiveness of our



(a) The raw data of Task 1 (b) The raw data of Task 2 (c) The raw data of Task 3



(d) Task 1 with F2SST (e) Task 2 with F2SST (f) Task 3 with F2SST

Figure 3: Patch embedding visualization of three randomly sampled 5-way 1-shot classification tasks with one query image per class on miniImageNet dataset. (a), (b), and (c) show the visualization on the raw data. (d), (e), and (f) show the corresponding visualization results with F2SST.

unified design that combines spectral modeling, spatial integration, and attention.

Frequency Selection Strategy. To systematically analyze frequency-specific contributions, we partition the spectrum based on radial distance from the DC component. Table 4 reveals distinct characteristics across frequency bands: the superior performance of low-frequency components indicates their predominant encoding of class-discriminative semantic information. Interestingly, high-frequency regions demonstrate limited utility, likely containing primarily fine-grained noise or class-irrelevant local textures. The band-pass strategy's intermediate performance suggests mid-range frequencies capture partial but incomplete semantic patterns. These findings demonstrate that effective spectral modeling requires careful frequency band selection, with low-frequency components proving most critical for semantic representation in few-shot image classification.

Frequency-Class Weight. To examine the effect of the frequency-class weight λ in the fusion module, we test different values ranging from 0.0 (no weighting) to 3.0. Table 5 shows that moderate values like 1.0 achieves optimal performance in the critical 1-shot setting while maintaining competitive 5-shot results. The default is set to be 1.0.

Choices in Eq.(4) Table 6 shows that the squared similarity formulation performs best. Intuitively, this transformation increases the penalty for mismatched features while further amplifying the impact of high-confidence matches, which enhances class separability.

Choices in Eq. (4)	1-shot	5-shot
$d(\cdot,\cdot)^2$	**89.27 ± 0.56**	**97.41 ± 0.18**
$d(\cdot,\cdot)$	86.03 ± 0.57	97.22 ± 0.18
$\lvert d(\cdot,\cdot)\rvert$	89.23 ± 0.56	97.12 ± 0.19

Table 6: Impact of choices in Eq. (4) on the few-shot classification performance (Low-Frequency).

Feature visualization Figure 3 shows the patch embedding visualization results of three randomly sampled 5-way 1-shot classification tasks with/without F2SST. It can be observed that with F2SST, the patch embeddings are clustered by class. This phenomenon indicates that our method can better approximate semantics to classes.

Conclusion

In this paper, we propose F2SST, a framework that exploits frequency-domain signals as implicit semantics for few-shot image classification. Through FFT-based spectral extraction, Gated Spectral Attention (GSA), and class-guided spatial-frequency fusion, F2SST enhances representations under limited supervision. Experiments on four benchmarks demonstrate consistent improvements in both 1-shot and 5-shot settings, with low-frequency components proving most semantically discriminative. Current limitations include heuristic frequency selection and shallow fusion depth, motivating future work on adaptive decomposition and deeper cross-domain integration.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No. 62176172, 61672364); partially by the National Key Research and Development Program of China (Grant No. 2018YFA0701701); and partially by the Natural Science Foundation of Jiangsu Province (Grant for Young Scholars, Grant No. BK20250789).

References

Afrasiyabi, A.; Lalonde, J.-F.; and Gagné, C. 2021. Mixture-based feature space learning for few-shot image classification. In *ICCV*, 9041–9051.

Bär, A.; Houlsby, N.; Dehghani, M.; and Kumar, M. 2024. Frozen feature augmentation for few-shot image classification. In *CVPR*, 16046–16057.

Bertinetto, L.; Henriques, J. F.; Torr, P. H.; and Vedaldi, A. 2019. Meta-learning with differentiable closed-form solvers. In *ICLR*.

Chaudhuri, Y.; Kumar, A.; Buduru, A. B.; and Alshamrani, A. 2024. FGA: Fourier-Guided Attention Network for Crowd Count Estimation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Chen, L.; Fu, Y.; Gu, L.; Yan, C.; Harada, T.; and Huang, G. 2024. Frequency-Aware Feature Fusion for Dense Image Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10763–10780.

Chen, W.; Si, C.; Zhang, Z.; Wang, L.; Wang, Z.; and Tan, T. 2023. Semantic prompt for few-shot image recognition. In *CVPR*, 23581–23591.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2019. A closer look at few-shot classification. In *ICLR*.

Chen, Z.; Ge, J.; Zhan, H.; Huang, S.; and Wang, D. 2021. Pareto self-supervised training for few-shot learning. In *CVPR*, 13663–13672.

Cheng, H.; Yang, S.; Zhou, J. T.; Guo, L.; and Wen, B. 2023. Frequency guidance matters in few-shot learning. In *ICCV*, 11814–11824.

Dong, B.; Zhou, P.; Yan, S.; and Zuo, W. 2022. Self-promoted supervision for few-shot transformer. In *European conference on computer vision*, 329–347. Springer.

Dong, L.; Zhai, W.; and Zha, Z.-J. 2023. Exploring tuning characteristics of ventral stream's neurons for few-shot image classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 534–542.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; and Gelly, S. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fei, N.; Lu, Z.; Xiang, T.; and Huang, S. 2021. MELR: Meta-learning via modeling episode-level relationships for few-shot learning. In *ICLR*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Hao, F.; He, F.; Liu, L.; Wu, F.; Tao, D.; and Cheng, J. 2023. Class-aware patch embedding adaptation for few-shot image classification. In *ICCV*, 18905–18915.

Hiller, M.; Ma, R.; Harandi, M.; and Drummond, T. 2022. Rethinking generalization in few-shot classification. *Advances in neural information processing systems*, 35: 3582–3595.

Hu, M.; Chang, H.; Guo, Z.; Ma, B.; Shan, S.; and Chen, X. 2023. Understanding few-shot learning: Measuring task relatedness and adaptation difficulty via attributes. *Advances in neural information processing systems*, 36: 19397–19409.

Hu, Y.; and Ma, A. J. 2022. Adversarial feature augmentation for cross-domain few-shot classification. In *European conference on computer vision*, 20–37. Springer.

Kang, D.; Kwon, H.; Min, J.; and Cho, M. 2021. Relational embedding for few-shot classification. In *ICCV*, 8822–8833.

Kim, J.; Kim, H.; and Kim, G. 2020. Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning. In *European conference on computer vision*, 599–617. Springer.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *Computer Science*.

Lazarou, M.; Stathaki, T.; and Avrithis, Y. 2022. Tensor feature hallucination for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3500–3510.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-learning with differentiable convex optimization. In *CVPR*, 10657–10665.

Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; and Ontanon, S. 2022. Fnet: Mixing tokens with fourier transforms. In *Proceedings of the 2022 Conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, 4296–4313.

Liang, Y.; Cao, Z.; Deng, S.; Dou, H.-X.; and Deng, L.-J. 2024. Fourier-enhanced implicit neural fusion network for multispectral and hyperspectral image fusion. *Advances in Neural Information Processing Systems*, 37: 63441–63465.

Liu, F.; Yang, S.; Chen, D.; Huang, H.; and Zhou, J. 2024. Few-shot classification guided by generalization error bound. *Pattern Recognition*, 145(000): 12.

Liu, Q.; Cao, W.; and He, Z. 2023. Cycle optimization metric learning for few-shot classification. *Pattern Recognition*, 139: 109468.

Liu, Y.; Sangineto, E.; Bi, W.; Sebe, N.; Lepri, B.; and Nadai, M. 2021a. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34: 23818–23830.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.

Luo, X.; Wei, L.; Wen, L.; Yang, J.; Xie, L.; Xu, Z.; and Tian, Q. 2021. Rectifying the shortcut learning of background for few-shot learning. *Advances in Neural Information Processing Systems*, 34: 13073–13085.

Ma, J.; Xie, H.; Han, G.; Chang, S.-F.; Galstyan, A.; and Abd-Almageed, W. 2021. Partner-assisted learning for few-shot image classification. In *ICCV*, 10573–10582.

Oreshkin, B.; Rodríguez López, P.; and Lacoste, A. 2018. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31.

Pan, M.-H.; Xin, H.-Y.; and Shen, H.-B. 2024. Semantic-based implicit feature transform for few-shot classification. *International Journal of Computer Vision*, 132(11): 5014–5029.

Qi, G.; Yu, H.; Lu, Z.; and Li, S. 2021. Transductive few-shot classification on the oblique manifold. In *ICCV*, 8412–8422.

Qin, Z.; Zhang, P.; Wu, F.; and Li, X. 2021. Fcanet: Frequency channel attention networks. In *ICCV*, 783–792.

Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; and Zhou, J. 2021. Global filter networks for image classification. *Advances in neural information processing systems*, 34: 980–993.

Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-learning for semi-supervised few-shot classification. In *ICLR*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Sun, S.; and Gao, H. 2023. Meta-AdaM: An meta-learned adaptive optimizer with momentum for few-shot learning. *Advances in Neural Information Processing Systems*, 36: 65441–65455.

Tatsunami, Y.; and Taki, M. 2024. Fft-based dynamic token mixer for vision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15328–15336.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking few-shot image classification: a good embedding is all you need? In *European conference on computer vision*, 266–282. Springer.

Trosten, D. J.; Chakraborty, R.; Løkse, S.; Wickstrøm, K. K.; Jenssen, R.; and Kampffmeyer, M. C. 2023. Hubs and hyperspheres: Reducing hubness and improving transductive few-shot learning with hyperspherical embeddings. In *CVPR*, 7527–7536.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wu, J.; Zhang, T.; Zhang, Y.; and Wu, F. 2021. Task-aware part mining network for few-shot learning. In *ICCV*, 8433–8442.

Xie, J.; Long, F.; Lv, J.; Wang, Q.; and Li, P. 2022. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *CVPR*, 7972–7981.

Xu, C.; Fu, Y.; Liu, C.; Wang, C.; Li, J.; Huang, F.; Zhang, L.; and Xue, X. 2021. Learning dynamic alignment via meta-filter for few-shot learning. In *CVPR*, 5182–5191.

Xu, J.; and Le, H. 2022. Generating representative samples for few-shot classification. In *CVPR*, 9003–9013.

Ye, H.-J.; Hu, H.; Zhan, D.-C.; and Sha, F. 2020. Few-shot learning via embedding adaptation with set-to-set functions. In *CVPR*, 8808–8817.

Yu, Y.; Zhang, D.; Li, Y.; and Zhang, Z. 2022. Multi-Proxy Learning from an Entropy Optimization Perspective. In *IJCAI*, 1594–1600.

Zhang, B.; Li, X.; Ye, Y.; Huang, Z.; and Zhang, L. 2021a. Prototype completion with primitive knowledge for few-shot learning. In *CVPR*, 3754–3762.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *CVPR*, 12203–12213.

Zhang, C.; Ding, H.; Lin, G.; Li, R.; Wang, C.; and Shen, C. 2021b. Meta navigator: Search for a good adaptation policy for few-shot learning. In *ICCV*, 9435–9444.

Zhang, H.; Xu, J.; Jiang, S.; and He, Z. 2024. Simple Semantic-Aided Few-Shot Learning. In *CVPR*, 28588–28597.

Zhang, J.; Zhao, C.; Ni, B.; Xu, M.; and Yang, X. 2019. Variational Few-Shot Learning. In *ICCV*, 1685–1694.

Zhang, M.; Zhang, J.; Lu, Z.; Xiang, T.; Ding, M.; and Huang, S. 2021c. IEPT: Instance-level and episode-level pretext tasks for few-shot learning. In *ICLR*.

Zhang, R.; Che, T.; Ghahramani, Z.; Bengio, Y.; and Song, Y. 2018. Metagan: An adversarial approach to few-shot learning. *Advances in neural information processing systems*, 31.

Zhang, X.; Meng, D.; Gouk, H.; and Hospedales, T. M. 2021d. Shallow bayesian meta learning for real-world few-shot recognition. In *ICCV*, 651–660.

Zhao, J.; Yang, Y.; Lin, X.; Yang, J.; and He, L. 2021. Looking wider for better adaptive representation in few-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 10981–10989.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2022. ibot: Image bert pre-training with online tokenizer. In *ICLR*.

Zhou, Z.; Qiu, X.; Xie, J.; Wu, J.; and Zhang, C. 2021. Binocular mutual learning for improving few-shot classification. In *ICCV*, 8402–8411.