



Universiteit
Leiden
The Netherlands

Exploring deep learning for data-efficient and generalizable visual recognition

He, K.

Citation

He, K. (2026, September 22). *Exploring deep learning for data-efficient and generalizable visual recognition*. Retrieved from <https://hdl.handle.net/1887/4310276>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4310276>

Note: To cite this publication please use the final published version (if applicable).

Chapter 4

Attention-guided Feature Pyramid Network for few-shot learning

In the previous chapter, we have studied the influences of a good metric in few-shot learning (FSL) tasks. In this chapter, we turn our attention to strengthening the discriminative capability of backbone features by addressing the following question: how to develop a flexible multi-scale representation framework that can dynamically reconcile feature hierarchies while preserving their discriminative capability under limited supervision (**RQ 2**).

Conventional few-shot learning models are typically built on backbone networks that extract features from a single or fixed-resolution representation space. However, in real-world scenarios, models must adapt to diverse visual patterns and significant scale variations across different categories, making it inadequate for these approaches. In this chapter, we introduce an attention-based feature pyramid network (AFPV) to enhance the multi-scale expressiveness of FSL frameworks. AFPV integrates feature maps from multiple backbone layers through attention-driven fusion modules, enabling the model to selectively aggregate discriminative information across spatial scales while suppressing irrelevant or noisy information. This design improves the model’s ability to construct more stable and generalizable representations, even when training data are extremely limited. Experimental results show that: (1) AFPV significantly strengthens classification robustness across few-shot scenarios; (2) the improved representations transfer effectively to unseen categories.

This chapter is based on the following publication [34]:

- **He, K.**, Bakker, E. M., and Lew, M. S. “Attention-guided Feature Pyramid Network for few-shot learning.” Benelux Conference on Artificial Intelligence and Machine Learning, 2024..

4.1 Introduction

Few-shot learning (FSL) [24, 101, 227] has emerged as a significant challenge in machine learning area where models must rapidly adapt to new tasks with only limited labeled examples. In contrast to conventional deep learning paradigms which rely on large-scale annotated datasets, FSL models aim to efficiently extract and transfer generalizable representations from limited samples. This setting makes FSL particularly valuable in many real-world applications, such as medical diagnosis [263, 264], autonomous driving [265, 266], and remote sensing [267, 268], where data acquisition is time-consuming or infeasible. Over the past decades, various research efforts have been made to improve the representation learning ability under the FSL setting. Metric-based approaches such as Matching Networks [14] and Prototypical Networks [1] design distance metrics in the embedding space, while optimization-based frameworks like MAML [2] enable rapid adaptation to new tasks. In addition, auxiliary training strategies are also explored to alleviate overfitting risks and improve generalization ability in FSL. For instance, self-supervision techniques [269, 270] have been incorporated to generate additional pretext tasks from the limited samples, thereby enriching the feature learning process without requiring extra annotations. Similarly, knowledge distillation [110, 271] has been adapted to transfer knowledge from a large, well-trained teacher model to a compact student model, enhancing the latter’s generalization from few examples.

Among these FSL approaches, attention mechanism also attracts much interest in FSL as it enables models to emphasize discriminative regions while suppress irrelevant background features. For instance, Huang and Choi [272] designed SAPENet, which applies multi-head self-attention to recalibrate local feature maps and enhance prototype representations, thereby improving few-shot classification performance. Similarly, Hou *et al.* [22] introduced a cross-attention-based comparison module that constructs attention maps between support and query features to focus on task-relevant object regions while suppressing background clutter, achieving competitive or state-of-the-art results on benchmarks such as miniImageNet and CUB-200-2011. More recently, FEAT [239] proposed a Transformer-style attention module to learn task-specific embeddings in the feature space, and Yang *et al.* [273] further design a meta-tuned transformer which employs self-attention to jointly model relationships among patches, samples, and tasks, bringing significant improvements for few-shot classification on both in-domain and cross-domain benchmarks. However, these methods often treat features at a single scale, ignoring the hierarchical structure of visual representations that are crucial for distinguishing fine-grained or low-contrast categories in few-shot settings.

Building on these insights, we propose an Attention-Guided Feature Pyramid Network (AFPN) for few-shot image classification. AFPN aims to fully exploit hierarchical and spatial information by integrating attention mechanisms into a multi-scale feature learning framework. As presented in Figure 4.1, multi-scale features often

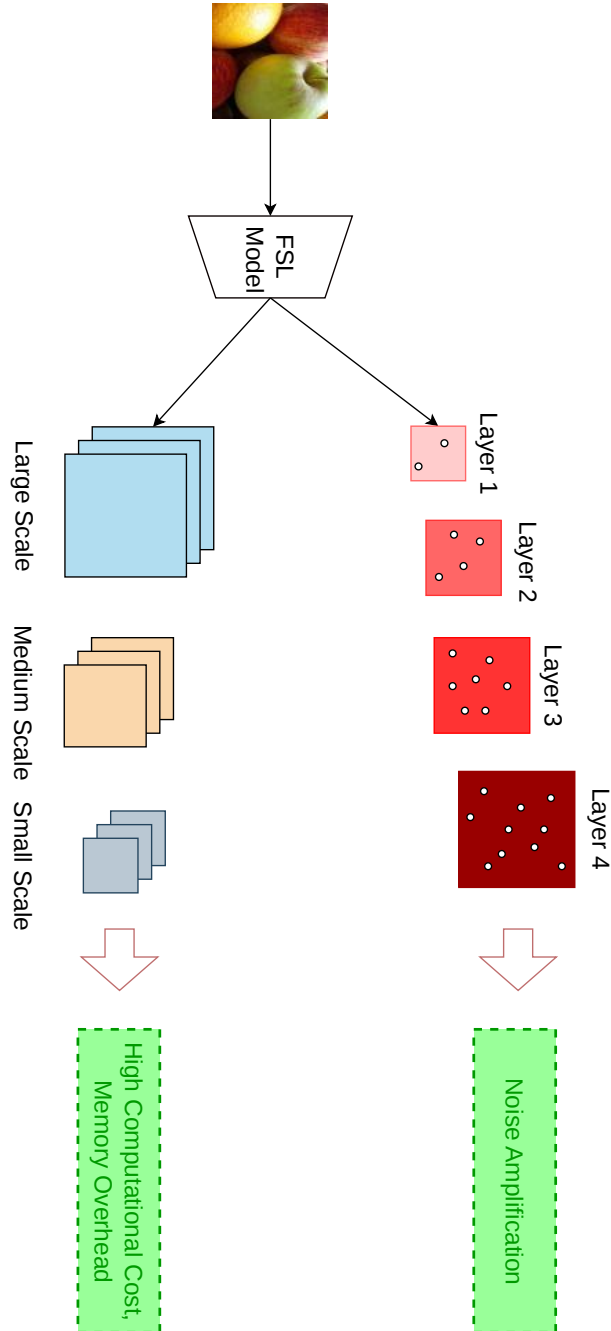


Figure 4.1: Multi-scale features challenge. Large-scale features incur higher computational cost, while deeper layers tend to amplify noise.

4. ATTENTION-GUIDED FEATURE PYRAMID NETWORK FOR FEW-SHOT LEARNING

face high computational cost and noise amplification challenges. Inspired by the success of Feature Pyramid Networks (FPNs) [274] in capturing both fine-grained and high-level semantic cues, AFPN integrates two distinct but complementary attention mechanisms to coordinate effective multi-scale feature interactions. Specifically, we incorporate a parameter-free SimAM attention module [275], which efficiently enhances feature discriminability without introducing additional learnable parameters. It adaptively measures neuron importance based on energy functions, enabling the model to emphasize informative regions while suppressing noisy activations across feature maps. Besides, we design a Cross-layer Spatial Attention Module (CSAM) that explicitly models the interdependencies between feature maps from different pyramid levels. By enabling cross-layer feature modulation, CSAM facilitates the propagation of high-level semantic information to lower-level features and, conversely, allows fine-grained spatial details to refine higher-level representations. The overall framework of AFPN is shown in Figure 4.2. The workflows of SimAM and CSAM are presented in Figure 4.3 and 4.4, respectively.

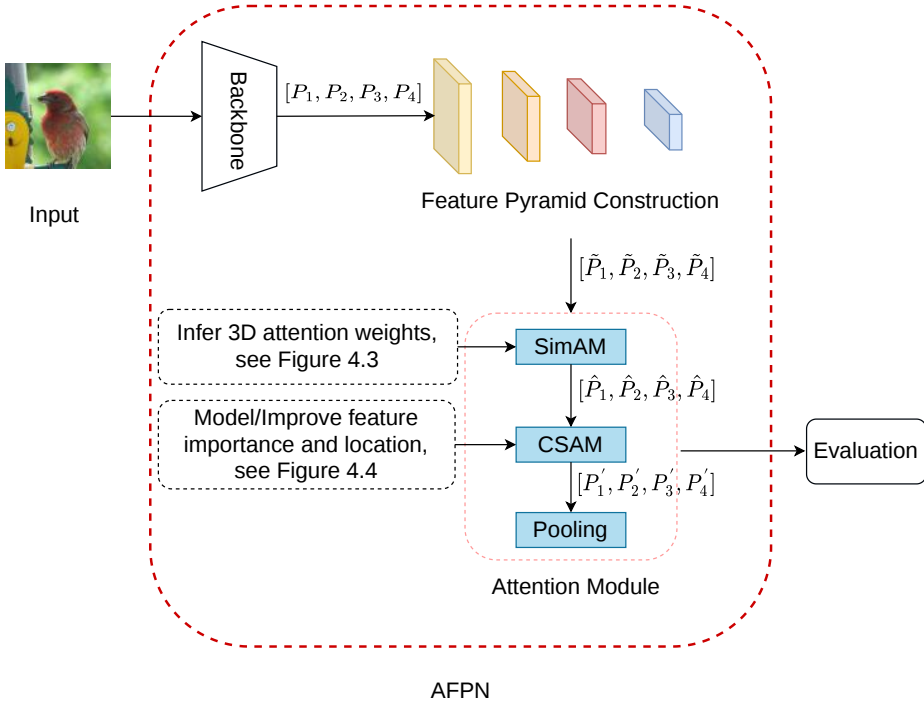


Figure 4.2: The overall framework of AFPN. The detailed implementations of SimAM and CSAM are presented in **Figure 4.3** and **Figure 4.4**

By combining these two attention mechanisms, AFPN achieves effective multi-scale feature interaction while maintaining computational efficiency. The joint operation

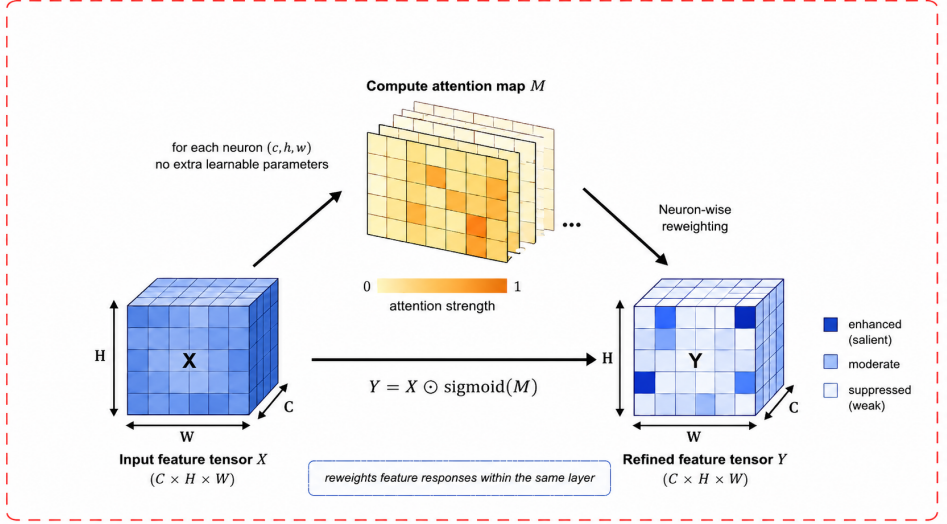


Figure 4.3: The workflow of SimAM

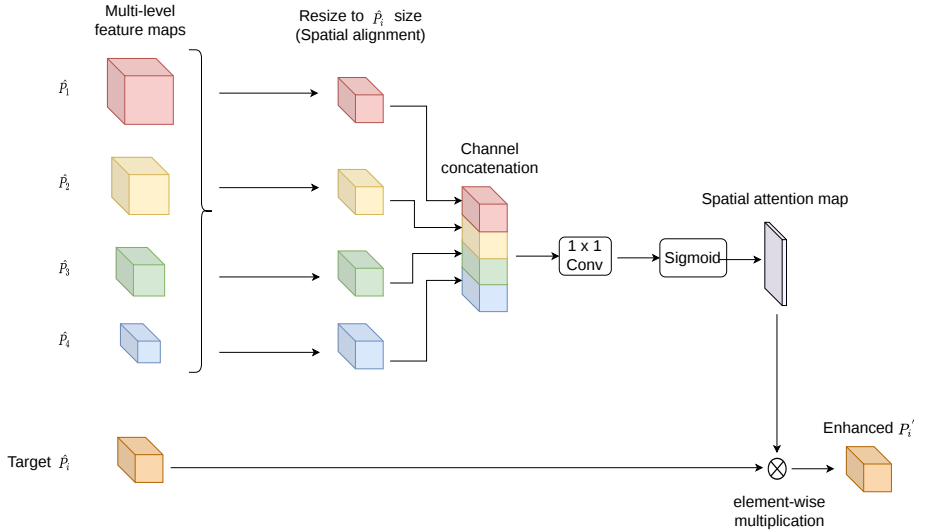


Figure 4.4: The workflow of CSAM

of SimAM and CSAM allows AFPN to retain richer high-level semantic information and simultaneously refine spatial representations crucial for few-shot recognition. Furthermore, this hierarchical attention fusion design enables the model to

generalize better across diverse visual domains, particularly under limited training samples.

In summary, our contributions are:

- We exploit a new feature pyramid architecture inherently captures multi-scale contextual information, which is crucial for recognizing objects from limited examples.
- We propose a novel dual-attention mechanism (SimAM and CSAM) enables dynamic and efficient feature refinement and interaction across scales, ensuring that semantically rich information is preserved and enhanced.
- We design a robust and efficient framework that significantly boosts few-shot classification accuracy by learning more discriminative and generalizable features from limited data, as validated by our extensive experiments.

4.2 Related Work

4.2.1 Few-Shot Learning Paradigm

Few-shot learning (FSL) has emerged as a fundamental paradigm for enabling deep models to recognize new classes with very limited annotated samples. The main challenge in FSL is how to bridge the gap between base classes and novel classes, requiring models to learn transferable representations that generalize to different tasks. Early works such as Matching Network [14] and Relation Network [24] introduced episodic training strategies to mimic the few-shot scenario during meta-learning. Later, optimization-based methods like MAML [2] and its variants [28, 191] focused on rapid adaptation through gradient-based fine-tuning. Recently, researchers have turned to embedding enhancement and regularization strategies to improve the quality of learned representations. For example, self-distillation [269], manifold alignment [117], and meta-transfer learning [198] have been used to preserve feature consistency across tasks. Furthermore, several studies explored leveraging structural priors or hierarchical semantics [276] to enhance the discriminative power of few-shot models. Despite these advances, most existing frameworks rely on single-scale or task-specific embeddings, which limits their ability to capture fine-grained details and contextual information crucial for few-shot generalization.

4.2.2 Attention Mechanisms for Feature Enrichment

Different from traditional convolutional paradigms in computer vision, attention mechanisms enable dynamic feature refinement through content-aware processing, defining a new way for neural networks to capture data dependencies. Vaswani *et al.* [277] proposed a transformer model in natural language processing, where attention

allows a network to assign dynamic importance to different features, enhancing interpretability and adaptability. This achievement was quickly extended to computer vision. Hu *et al.* [241] introduced the Squeeze-and-Excitation (SE) block, which learned to emphasize informative feature channels through global context modeling, while Woo *et al.* [278] proposed the Convolutional Block Attention Module (CBAM), which integrates both spatial and channel attention for refined feature selection. Subsequently, Wang *et al.* [279] developed Non-local Neural Networks to capture long-range dependencies beyond local convolutional receptive fields, inspiring later attention-driven architectures such as Vision Transformers (ViT) [280] and Swin Transformers [281], which demonstrated that pure attention mechanisms can outperform convolution-based models in large-scale vision tasks.

In FSL, attention mechanisms have been widely used to address intra-class variation and inter-class confusion. Hou *et al.* [22] introduced a cross-attention module to align query and support embeddings, while Ye *et al.* [239] extended this idea with self-attention for adaptive prototype learning. Other works, such as Yang *et al.* [275] and Wang *et al.* [282], explored lightweight and efficient attention variants—like SimAM and ECA-Net—that retain discriminative power with minimal computational overhead. Moreover, recent studies have investigated cross-scale and multi-head attention formulations [34, 35] to enhance spatial consistency and semantic richness across hierarchical layers, which is particularly beneficial for FSL tasks. In general, these advances demonstrate that attention mechanisms not only enhance visual discrimination but also provide a flexible way to integrate contextual and hierarchical knowledge, making them an important component for improving few-shot learning performance.

4.3 Proposed Methods

To enhance the model’s generalization ability in FSL scenarios, we propose an Attention-Guided Feature Pyramid Network (AFPN), an effective framework that integrates multi-scale representation learning with dual attention mechanisms. In FSL settings, models often fail to capture discriminative high-level semantics while maintaining fine-grained local features. To mitigate this, AFPN introduces a hierarchical feature fusion module guided by a Simple Parameter-Free Attention Module (SimAM) and a Cross-Layer Spatial Attention Module (CSAM). The proposed architecture balances local and global contextual representations, enabling the model to produce robust embeddings for unseen classes with minimal data.

AFPN mainly consists of five main components: (1) a feature extraction backbone; (2) a feature pyramid network (FPN) for top-down semantic aggregation; (3) a SimAM module that adaptively refines intra-layer neuron importance; (4) a CSAM module that models inter-layer spatial interactions; and (5) a classification head for final prediction.

4.3.1 Multi-Scale Feature Extraction

Given an input image $x \in R^{3 \times H \times W}$ (where H and W denote its height and width, respectively), AFPN employs a stack of convolutional layers to progressively extract hierarchical features from shallow to deep levels. These extracted feature maps are defined as:

$$\begin{aligned} P_1 &= f_1(x), \\ P_2 &= f_2(x), \\ P_3 &= f_3(x), \\ P_4 &= f_4(x), \end{aligned} \tag{4.1}$$

where f_i represents the convolutional transformation at the i -th layer. Each feature map $P_i \in R^{C_i \times H_i \times W_i}$ captures higher semantic information but with decreasing spatial resolution.

4.3.2 Feature Pyramid Construction

Inspired by the Feature Pyramid Network (FPN) [17], we construct a top-down fusion pathway to integrate features with different scales. The FPN module enhances low-level representations by upsampling and merging higher-level semantic information while maintaining computational efficiency:

$$\begin{aligned} \tilde{P}_4 &= P_4, \\ \tilde{P}_i &= P_i + U_p(\tilde{P}_{i+1}), \quad i = 1, 2, 3. \end{aligned} \tag{4.2}$$

where U_p denotes bilinear interpolation. The output of this module is a set of multi-scale features $\{\tilde{P}_1, \tilde{P}_2, \tilde{P}_3, \tilde{P}_4\}$, which contain richer contextual information for further attention refinement.

4.3.3 Simple Parameter-Free Attention Module (SimAM)

To refine the importance of individual neurons within each feature map, AFPN employs the SimAM module, which computes neuron-level attention weights based on energy minimization without introducing additional learnable parameters. For any feature map $\tilde{P}_i \in \{\tilde{P}_1, \tilde{P}_2, \tilde{P}_3, \tilde{P}_4\}$, we use $M_i = H_i \times W_i$ to denote the number of neurons on i -th channel, SimAM computes the energy function for each neuron z as follows:

$$e_t = \frac{(t - \tilde{\mu})^2 + 2(\tilde{\sigma})^2 + 2\lambda}{4(\tilde{\sigma}^2 + \lambda)} \tag{4.3}$$

where t represents the target neuron, $\tilde{\mu} = \frac{1}{M_i} \sum_{j=1}^{M_i} z_j$ and $\tilde{\sigma}^2 = \frac{1}{M_i} \sum_{j=1}^{M_i} (z_j - \tilde{\mu})^2$, which denote the mean and variance of all neurons in the same channel excluding t , and λ is a regularization hyperparameter. The attention weights can be obtained

through a sigmoid activation:

$$A_i = \text{sigmoid}(e_i). \quad (4.4)$$

Then the refined feature map can be calculated as:

$$\hat{P}_i = A_i \odot \tilde{P}_i, \quad (4.5)$$

where $\odot$ denotes element-wise multiplication. This formulation allows the module to emphasize informative spatial locations without introducing additional parameters, making it particularly suitable for few-shot learning scenarios.

4.3.4 Cross-Layer Spatial Attention Module (CSAM)

While the SimAM module efficiently enhances salient neurons within each feature level, its capacity to model dependencies across different scales still remains limited. To address this, we design a Cross-Layer Spatial Attention Module (CSAM) that performs spatial reasoning between multi-scale features. For a target feature map P_i' , all refined feature maps are first resized to a common resolution and concatenated along the channel dimension:

$$F_i = \text{Concat} \left(\left\{ \text{Resize} \left(\hat{P}_j; H_i, W_i \right) \right\}_{j=1}^4 \right). \quad (4.6)$$

Then, we apply a 1×1 convolutional layer $g(\cdot)$ to generate a spatial attention map, followed by a sigmoid activation:

$$F_i' = \text{sigmoid}(g(F_i)). \quad (4.7)$$

Then the output of CSAM is computed as follows:

$$P_i' = \hat{P}_i \odot F_i' \quad (4.8)$$

4.3.5 Classification

For the enhanced feature maps $\{P_i'\}_{i=1}^4$ output from CSAM, we first apply a pooling layer on each layer to get a scale-specific descriptor::

$$v_i = \text{GAP}(P_i'), \quad (4.9)$$

where $\text{GAP}(\cdot)$ denotes global average pooling. Then we concatenate these descriptors to form a multi-scale embedding:

$$v = [v_1; v_2; v_3; v_4]. \quad (4.10)$$

Finally, we pass the embedding to a classifier $h(\cdot)$ to obtain the prediction:

$$y' = h(v). \quad (4.11)$$

4.4 Experiments

4.4.1 Benchmarks and Implementation Details

Benchmarks: Our experiments are carried out on three widely adopted FSL benchmarks: miniImageNet [14], tieredImageNet [62] and CIFAR-FS [112]. The miniImageNet dataset is a subset of ImageNet, including 100 classes with 600 images per class. Following the partitioning protocol in [115], we employ 64 classes for training, 16 classes for validation, and the remaining 20 classes for testing. The tieredImageNet dataset contains 608 classes organized into 34 coarse-grained categories. We select 20 categories (351 classes) for training, 6 categories (97 classes) for validation, and 8 categories (160 classes) for testing. CIFAR-FS is derived from the CIFAR100 dataset and is split into 64 training classes, 16 validation classes, and 20 testing classes, with each class containing 600 images.

Implementation Details: For miniImageNet and tieredImageNet, all images are resized to 84×84 , while images for CIFAR-FS are resized to 32×32 . We use a simple convolutional neural network consisting of four blocks (Conv4) as the backbone. Each block comprises the following components: a 64-filter 3×3 convolution layer, a batch normalization layer, a ReLU activation function and a 2×2 max-pooling layer. We use SGD with a momentum of 0.9 and a weight decay of $5e-4$. The initial learning rate is set as 0.05 and decayed with a factor of 0.1. Each epoch includes 64 samples, we train 150 epochs and decay three times for miniImageNet and tieredImageNet, 120 epochs and decay twice for CIFAR-FS. The evaluation protocol is the accuracy with the 95% confidence interval of the test set when 600 random tasks are selected, with 15 query samples each task. We train an N -way logistic regression classifier for the evaluation.

4.4.2 Experimental results

miniImageNet and tieredImageNet. Table 4.1 presents a comparison between our approach (AFPN) and the state-of-the-art methods on two popular FSL benchmarks. Across both datasets and under both 1-shot and 5-shot settings, AFPN obtains state-of-the-art performance. On miniImageNet, AFPN achieves 54.52% in the 5-way 1-shot task and 72.87% in the 5-way 5-shot task, surpassing the previously best-reported results by 1.65% and 1.85%, respectively. On tieredImageNet, AFPN achieves 57.62% and 74.69% in the 1-shot and 5-shot tasks, respectively. AFPN achieves improvements over the best-reported numbers by 2.48% under 5-way 1-shot and 0.46% under 5-way 5-shot settings.

Table 4.1: Comparison of AFPN (our approach) to prior works on miniImageNet and tieredImageNet datasets, with mean accuracy (%) and 95% confidence interval.

Model	Backbone	miniImageNet 5-way		tieredImageNet 5-way	
		1-shot	5-shot	1-shot	5-shot
ProtoNet [1]	Conv4-64	49.42 $\pm$ 0.78	68.20 $\pm$ 0.66	53.31 $\pm$ 0.89	72.69 $\pm$ 0.74
MAML [100]	Conv4-64	48.07 $\pm$ 1.75	63.15 $\pm$ 0.91	51.67 $\pm$ 1.81	70.30 $\pm$ 1.75
RelationNet [24]	Conv4-64	50.44 $\pm$ 0.82	65.32 $\pm$ 0.70	54.48 $\pm$ 0.93	71.31 $\pm$ 0.78
DeepEMD [37]	Conv4-64	51.72 $\pm$ 0.20	65.10 $\pm$ 0.39	51.22 $\pm$ 0.14	65.81 $\pm$ 0.68
CGRN [283]	Conv4-64	50.85 $\pm$ 0.86	64.13 $\pm$ 0.70	55.07 $\pm$ 0.20	71.34 $\pm$ 0.30
CovaMNet [151]	Conv4-64	51.19 $\pm$ 0.76	67.65 $\pm$ 0.63	54.98 $\pm$ 0.90	71.51 $\pm$ 0.75
MetaOptNet [237]	Conv4-64	52.87 $\pm$ 0.57	68.76 $\pm$ 0.48	54.71 $\pm$ 0.67	71.79 $\pm$ 0.59
DN4 [116]	Conv4-64	51.24 $\pm$ 0.74	71.02 $\pm$ 0.64	52.89 $\pm$ 0.23	73.36 $\pm$ 0.73
SSL-ProtoNet [284]	Conv4-64	52.58 $\pm$ 0.45	70.87 $\pm$ 0.36	55.14 $\pm$ 0.49	74.23 $\pm$ 0.40
AFPN	Conv4-64	54.52 $\pm$ 0.79	72.87 $\pm$ 0.60	57.62 $\pm$ 0.92	74.69 $\pm$ 0.70

CIFAR-FS. Table 4.2 illustrates similar comparisons on CIFAR-FS dataset. From Table 4.2, we can find that our method AFPN outperforms the other methods in comparison with the highest accuracy of 62.60% and 79.51% in 1-shot and 5-shot scenarios, respectively. In particular, AFPN achieves improvements over the best-reported numbers by 2.19% under 5-way 1-shot and 2.99% under 5-way 5-shot settings.

Table 4.2: Comparison of AFPN (our approach) to prior works on CIFAR-FS dataset, with mean accuracy (%) and 95% confidence interval.

Model	Backbone	CIFAR-FS 5-way	
		1-shot	5-shot
ProtoNet [1]	Conv4-64	55.50 $\pm$ 0.70	72.00 $\pm$ 0.69
MAML [100]	Conv4-64	58.90 $\pm$ 1.90	71.50 $\pm$ 1.00
CSS [249]	Conv4-64	56.49 $\pm$ 0.93	74.59 $\pm$ 0.72
SSL-ProtoNet [284]	Conv4-64	60.41 $\pm$ 0.52	76.52 $\pm$ 0.38
AFPN	Conv4-64	62.60 $\pm$ 0.91	79.51 $\pm$ 0.70

4.4.3 Ablation Study

Benefits of Feature Pyramid Module. To study the impact of feature pyramid module, we evaluate our approach with and without feature pyramid module on miniImageNet. As shown in Table 4.3, for the model trained without feature pyramid module, it achieves 54.12% and 72.45.% in 1-shot and 5-shot tasks, respectively. For the model trained with feature pyramid module, it achieves 0.4% and 0.42% improvements in 1-shot and 5-shot tasks, respectively. These results confirm the effectiveness of feature pyramid module.

Contributions of Each Layer in the Backbone. To more clearly assess the contributions of each layer in the backbone, we conduct an ablation study on mini-

4. ATTENTION-GUIDED FEATURE PYRAMID NETWORK FOR FEW-SHOT LEARNING

Table 4.3: FSL results on miniImageNet, with and without feature pyramid module.

Model	miniImageNet 5-way	
	1-shot	5-shot
AFPN (without feature pyramid module)	54.12 ± 0.79	72.45 ± 0.61
AFPN (with feature pyramid module)	54.52 ± 0.79	72.87 ± 0.60

ImageNet by training the model with different selection of layers. As reported in Table 4.4, AFPN with P_4 layer obtains accuracies of 53.91% and 72.14% for the 1-shot and 5-shot settings, respectively. AFPN gets better performance with the addition of other layers and gains 0.55% and 0.73% improvements after adding all the layers in the backbone. Overall, these results highlight the key role of each layer in the backbone.

Table 4.4: FSL results on miniImageNet, with different layers.

Selected Layer	miniImageNet 5-way	
	1-shot	5-shot
P_4	53.91 ± 0.81	72.14 ± 0.62
P_3, P_4	54.06 ± 0.79	72.22 ± 0.60
P_2, P_3, P_4	54.11 ± 0.80	72.46 ± 0.60
P_1, P_2, P_3, P_4	54.52 ± 0.79	72.87 ± 0.60

Benefits of Attention Module. To quantify the effect of attention module, we perform a comparison on miniImageNet by training the same model with different attention module. Baseline denotes our AFPN without any attention module, which only contains a backbone and a feature pyramid network module. Specifically, the baseline follows the standard AFPN pipeline, where the backbone extracts multi-level feature representations and the AFPN performs feature fusion across different scales. No additional attention mechanism is introduced in the baseline, so it can serve as a fair reference for evaluating the effectiveness of different attention modules. As shown in Table 4.5, The baseline reaches 52.73% in the 1-shot setting and 70.62% in the 5-shot setting. Incorporating SimAM module boosts the results, bringing 0.74% improvement for 1-shot and 1.02% for 5-shot compared to the baseline. AFPN further gets 1.05% and 1.23% improvements after introducing CSAM module. This consistent gain demonstrates that attention module plays a crucial role in improving performance.

4.5 Chapter Conclusions

In this chapter, we focus on a fundamental, yet underexplored challenge in few-shot image classification — how to effectively exploit multi-scale visual information

Table 4.5: FSL results on miniImageNet, with different attention module.

Model	miniImageNet 5-way	
	1-shot	5-shot
Baseline	52.73 ± 0.79	70.62 ± 0.64
AFPN with SimAM	53.47 ± 0.78	71.64 ± 0.62
AFPN with SimAM + CSAM	54.52 ± 0.79	72.87 ± 0.60

under limited supervision. To this end, we propose a novel Attention-Guided Feature Pyramid Network (AFPN), which is specifically designed to facilitate effective feature interaction across different network layers. In order to get discriminative representations, AFPN first introduces a Simple Parameter-Free Attention Module (SimAM) that efficiently highlights crucial spatial locations without adding learnable parameters, thereby preventing overfitting in data-scarce scenarios. Then, we design a Cross-Layer Spatial Attention Module (CSAM) which explicitly addresses the semantic gaps between adjacent pyramid layers through a structured fusion mechanism. Extensive experiments on standard benchmarks demonstrate that our AFPN achieves competitive performance compared to other state-of-the-art few-shot learning methods. The key insights derived from our experimental analysis can be summarized as follows:

- (1) As visualized in Table 4.5, The parameter-free attention design in SimAM proves its effectiveness in few-shot scenarios (e.g., 1-shot and 5-shot settings). The module successfully enhances model generalization while mitigating the overfitting risk in few-shot learning.
- (2) The CSAM mechanism shows complementary benefits with SimAM, as demonstrated by the performance gains observed in Table 4.5. Specifically, the integration module enables more effective propagation of high-level semantic information to lower layers while preserving fine-grained spatial details from earlier layers. This bidirectional information flow creates more discriminative representations for novel categories, with the most significant improvements observed in more challenging fine-grained classification tasks where subtle visual differences are critical for correct categorization.
- (3) AFPN keeps representational diversity, and ablation studies reveal that the combination of our proposed modules achieves better performance compared to using either module in isolation, showing their synergistic relationship. This finding emphasizes the importance of addressing both intra-layer feature refinement and inter-layer feature fusion in a coordinated framework for few-shot learning problems.