

KNN Transformer with Pyramid Prompts for Few-Shot Learning

Anonymous Authors

ABSTRACT

Few-Shot Learning (FSL) aims to recognize new classes with limited labeled data. Recent studies have attempted to address the challenge of rare samples with textual prompts to modulate visual features. However, they usually struggle to capture complex semantic relationships between textual and visual features. Moreover, vanilla self-attention is heavily affected by useless information in images, severely constraining the potential of semantic priors in FSL due to the confusion of numerous irrelevant tokens during interaction. To address these aforementioned issues, a k -NN Transformer with Pyramid Prompts (KTPP) is proposed to select discriminative information with k -NN Context Attention (KCA) and adaptively modulate visual features with Pyramid Cross-modal Prompts (PCP). First, for each token, the KCA only selects the k most relevant tokens to compute the self-attention matrix and incorporates the mean of all tokens as the context prompt to provide the global context in three cascaded stages. As a result, irrelevant tokens can be progressively suppressed. Secondly, pyramid prompts are introduced in the PCP to emphasize visual features via interactions between text-based class-aware prompts and multi-scale visual features. This allows the ViT to dynamically adjust the importance weights of visual features based on rich semantic information at different scales, making models robust to spatial variations. Finally, augmented visual features and class-aware prompts are interacted via the KCA to extract class-specific features. Consequently, our model further enhances noise-free visual representations via deep cross-modal interactions, extracting generalized visual representation in scenarios with few labeled samples. Extensive experiments on four benchmark datasets demonstrate significant gains over the state-of-the-art methods, especially for the 1-shot task with 2.28% improvement on average due to semantically enhanced visual representations.

CCS CONCEPTS

• Computing methodologies → Object recognition.

KEYWORDS

Multi-Modality, Prompt, Transformer, Few-Shot Learning

1 INTRODUCTION

Deep learning has achieved significant success in computer vision. However, it typically relies heavily on large-scale labeled datasets, which can be costly and sometimes impractical to obtain, such as in rare disease diagnosis [33, 41] and industrial anomaly detection

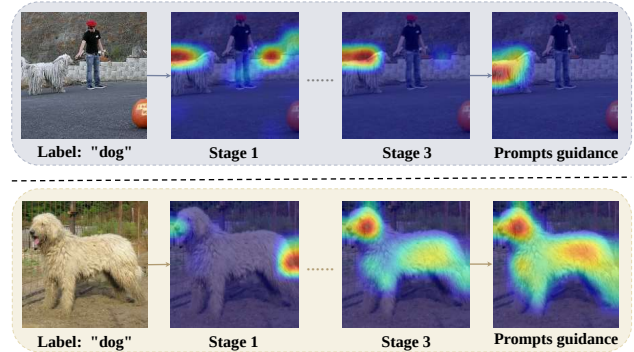


Figure 1: The image annotated as “dog” contains abundant spurious information such as people, walls, etc. Moreover, the scale variations present across different images, thus limiting the performance of ViTs in FSL. Our KTPP achieves coarse-to-fine filtration of noise, adaptation to spatial variations, and prompts-guided class-specific visual extraction.

[15, 25]. Instead, humans can recognize new objects from only a few images. To emulate human learning ability and reduce the dependency on annotated data, Few-Shot Learning (FSL) [50] is proposed to learn from a limited number of data, which has drawn considerable concern in recent years. Despite of remarkable progress [1, 7, 13, 14, 20, 50, 60, 68], achieving human-level performance remains challenging, especially in the 1-shot scenario with only one labeled sample.

Most methods [7, 13, 14, 50] utilize support features to infer class prototypes for classification. However, the limited number of labeled samples makes it challenging to learn discriminative features, leading to semantic biases in the generation of class-specific prototypes. To address this issue, an alternative is to involve integrating additional semantic information (e.g., category descriptions) with visual prototypes. Following this view, some studies [6, 31, 61–63, 73] reveal that integration of semantic priors can enhance prototype representation. Despite their success, most methods integrate semantic embeddings at a higher hierarchical level (e.g., prototypes and classifiers), lacking underlying feature interactions and ignoring distribution differences between modalities. Moreover, a recent study [6] attempted to modulate visual features by semantic embeddings of labels as prompts during the extraction process. However, it struggles to capture complex semantic relationships between textual and visual features by employing linear layers for direct and single-scale fusion.

Moreover, several approaches [6, 13, 20] perform interactions within or across modalities in Vision Transformers (ViTs) via vanilla self-attention. However, for vanilla self-attention where each patch is forced to interact with all tokens for computing the attention matrix, noisy tokens are inevitably introduced into the calculation.

Permission to make digital or hard copies of all or part of this work for personal or commercial use, or to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM MM, 2024, Melbourne, Australia
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
<https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

This incorporation of noisy tokens renders image features vulnerable to spurious correlations, especially in cluttered backgrounds and occlusions. As shown in Fig. 1, the image annotated as “dog” consists of one person, apple, wall, tree, and other class-irrelevant objects, making the ViT falsely associate these objects with the label when only few labeled samples are available. Therefore, the potential of semantic priors is severely constrained in cross-modal interactions via vanilla self-attention due to the confusion of numerous irrelevant tokens during interaction. As a result, designing an effective cross-modal fusion strategy and improving the self-attention mechanism are crucial to achieving deep cross-modal interactions and filtering out class-irrelevant visual features in FSL frameworks.

To address these aforementioned issues, a k -NN Transformer with Pyramid Prompts (KTPP) is proposed to fully leverage semantic information and capture class-specific support features. Specifically, the KTPP mainly consists of two modules: K -NN Context Attention (KCA) and Pyramid Cross-modal Prompts (PCP). Firstly, the KCA selects the k most relevant tokens for each token to compute attention weights instead of all tokens, effectively excluding irrelevant information and reducing the computational burden. As a result, irrelevant tokens can be progressively suppressed in a coarse-fine manner via three cascaded stages, as shown in Fig. 1. To avoid the issue of paying too much attention to the local context and neglecting the discriminative information of global feature [19, 32] in the top- k operation, the mean of all tokens is regarded as context prompts which are incorporated to enhance the global context information of each token. Secondly, the pre-trained CLIP model [46] has an excellent capability of cross-modal feature alignments, extracting rich and unbiased semantic embeddings. Therefore, we exploit the class-aware prompts encoded from class names via the CLIP. the PCP enables these class-aware prompts to interact with multi-scale support features, learning pyramid prompts that capture complex semantic relationships between textual and visual features. The pyramid structure allows the ViT to dynamically focus on the support features most relevant to the semantic information, making it robust to spatial variations. Finally, the augmented support features and the class-aware prompts are then interacted via the KCA, which enables the ViT to further capture discriminative support features due to the semantically guided class-specific features. As shown in Fig. 1, the KTPP progressively filters out noisy tokens at different stages and further enhances noise-free visual representations via deep cross-modal interactions guided by prompts.

Overall, our contributions are summarized as follows:

- A KTPP method is proposed to fully exploit cross-modal information and select discriminative visual features.
- A K -NN Context Attention module is proposed to filter out class-irrelevant features in a coarse-fine manner, capturing class-specific features in cross-modal interactions and improving computational efficiency.
- Pyramid Cross-modal Prompts are proposed to enhance visual features via deep cross-modal interactions, improving the adaptivity ability to spatial variations.
- The proposed method achieves significant performance in four FSL benchmarks, particularly improving 1-shot accuracy by 1%-4% compared to the state-of-the-art approaches.

2 RELATED WORK

2.1 Few-Shot Learning

In general, methods for Few-Shot Learning can be divided into two types. One is optimization-based methods [16, 42, 49, 74], aimed at quickly adapting models to new data. The other type is metric-based methods adopted in this work aims to classify by measuring distances between support and query samples within the feature space. ProtoNet [50] uses Euclidean distance. Deepemd [68] utilizes the Earth Mover’s Distance. others [12, 21, 36, 64, 71] have integrated Transformer layers as classifiers. Moreover, recent works have incorporated textual modality as auxiliary information to boost generalization on novel classes. AM3 [61] merges semantic and visual prototypes through an adaptive fusion mechanism. Multiple-Semantics [66] aims to enhance prototype representation by integrating multiple semantic information TRAML [31] leverages semantic similarities between classes to improve classification accuracy by a task-relevant adaptive margin loss. SVAE [62] generates additional visual features from semantic information through a Conditional Variational Autoencoder (CVAE) [51], resulting in more reliable prototypes. KTN [45] utilizes graph convolutional networks and knowledge graphs to explicitly leverage semantic information for learning classifiers. These methods apply semantic information at a higher level such as prototypes and classifiers, which lack foundational visual feature interactions and ignore distribution differences. In contrast, we utilize ViT as the backbone, leveraging semantic information in the extraction process to modulate visual features adaptively.

2.2 Prompt Learning

To diminish the dependency on extensive training datasets and enhance the flexibility of pre-trained models for downstream tasks, several prompt-based approaches [6, 10, 23, 46, 53, 72, 73] have been proposed. CLIP [46] utilizes constructed prompts like “a photo of a class” to compute similarity scores between image embeddings and textual prompts for classification. CoOp [73] introduces learnable continuous prompt tokens for fine-tuning the pre-trained CLIP. Further building on CoOp, CoCoOp [72] employs a Meta-Net to transform each prompt token into instance-specific contexts, significantly improving model adaptability. MAPLE [23] designs corresponding prompts for each modality to guide the learning of the respective modality. VPT [22] introduces learnable parameters as visual prompts to fine-tune the CLS token of the ViT, enhancing the flexibility and adaptability of the model. Frozen [53] conducts multimodal few-shot learning by using image features paired with text as prompts. SP [6] utilizes textual labeled embedding matching images as prompts to supplement visual features by employing linear layers for direct and single-scale fusion. In this paper, we propose Pyramid Cross-modal Prompts to learn pyramid prompts, which can dynamically adjust the importance weights of visual features based on rich semantic information at different scales, making models robust to spatial variations.

2.3 Vision Transformer

Vision Transformers (ViTs) have shown remarkable success in numerous computer vision tasks [2, 5, 26, 34, 37, 43, 59, 70], benefiting

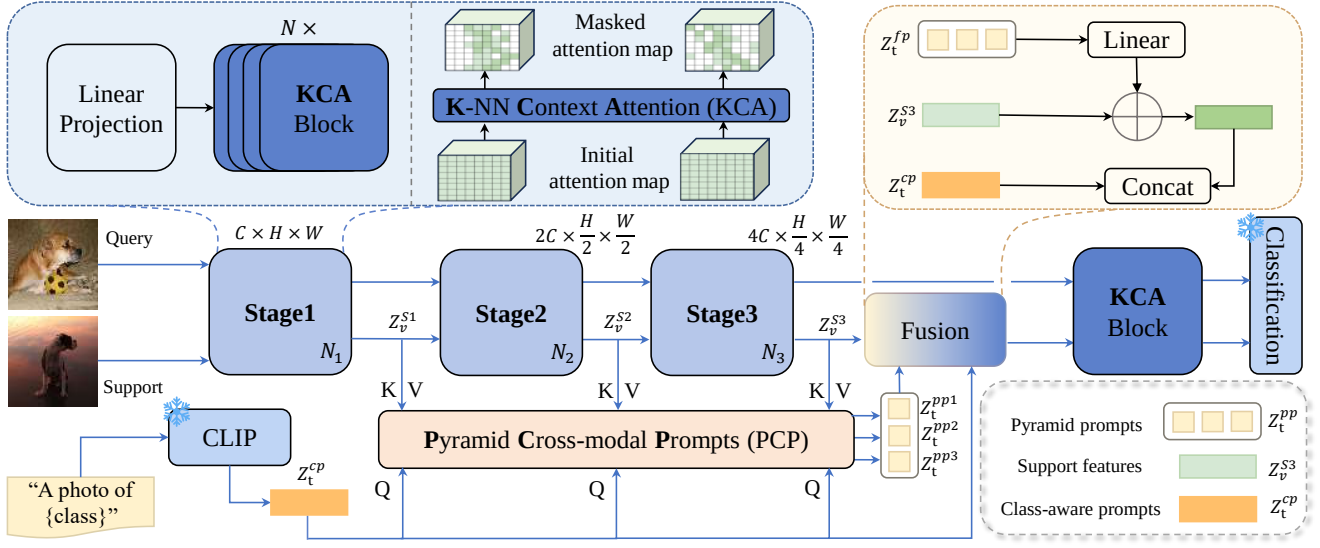


Figure 2: The framework of our proposed KTPP. Image patches are sequentially fed into three cascaded stages where KCA filters out irrelevant tokens and linear projection extracts multi-scale support features Z_v^S . Text-based class-aware prompts Z_t^{cp} are exploited via the CLIP. Pyramid prompts Z_t^{pp} are obtained to enhance support features via the PCA between Z_t^{cp} and Z_v^S . The enhanced support features and Z_t^{cp} are interacted via the KCA to extract class-specific support features.

from their proficiency to capture long-range dependencies among image patches. To match or surpass the performance of CNNs of similar size trained on ImageNet [11], ViTs are usually trained on large-scale datasets. Few recent works [4, 13, 20, 58] have explored the adaptation of ViTs on small datasets. IDMM [4] pretrains ViT using instance discrimination loss, followed by fine-tuning on the target dataset. SUN [13] introduces local supervision to enhance dependency learning of tokens. FewTURE [20] also achieves excellent performance with a fully Transformer-based architecture. SP [6] utilizes semantic prompts to enhance the visual feature by cross-modal interactions based on self-attention. Despite these advances, the fully connected self-attention mechanism in these models usually introduces noisy tokens, which can degrade the quality of the representation. KVT [58] attempts to refine this mechanism, yet the performance is limited in scenarios with few labeled samples due to excessive focus on local feature context. In this paper, we propose a k -NN Context Attention mechanism that progressively filters out irrelevant information in visual modality via three cascaded stages, captures discriminative features in cross-modal interactions guided by semantic information, and enhances global context information in FSL. To the best of our knowledge, we are the first to consider the way of integrating cross-modal data and context prompts for improving k -NN attention and applying it in FSL.

3 METHOD

We first overview the whole pipeline, and then introduce the preliminary of FSL. Next, two novel components of K-NN Context Attention (KCA) and Pyramid Cross-modal Prompts (PCP) are presented.

3.1 Overview

Fig. 2 shows the framework of our proposed KTPP. To suppress irrelevant information, the KCA progressively filters out noisy tokens in three cascaded stages. To exploit rich semantic priors, class-aware prompts are extracted by inputting class names into the CLIP. To capture complex semantic relationships between different modalities, pyramid prompts are obtained to enhance support features by applying the PCP at different scales, enabling the ViT to dynamically adjust the importance weights of support features based on semantic information. Consequently, models are robust to spatial variations. To select discriminative support features, enhanced support features and class-aware prompts are interacted via the KCA, which can extract class-specific support features guided by semantic information. Finally, the ViT can generate generalized class prototypes for classification.

3.2 Preliminary

Few-Shot Learning (FSL) focuses on generalizing the knowledge learned from training classes C_{train} to test classes C_{test} with only a few labeled samples, where the training and test categories are disjoint ($C_{\text{train}} \cap C_{\text{test}} = \emptyset$). This task is typically formalized as an N -way M -shot problem where the training involves N distinct categories and each category has M labeled images. An episodic training strategy [55] is adopted, where each episode consists of two parts: a support set $S = \{(x_i, y_i)\}_{i=1}^{N \times M}$ with $N \times M$ labeled samples, and a query set $Q = \{(x_i, y_i)\}_{i=1}^{N \times H}$ with $N \times H$ test samples to evaluate the performance.

In the FSL approach, prototype-based inference is commonly utilized. Specifically, N class prototypes are obtained by averaging

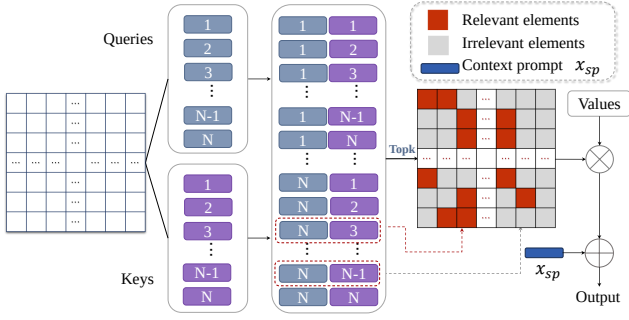


Figure 3: Illustration of K-NN Context Attention (KCA). For each query, the k highest scorers from the N query-key pairs are selected for computing attention weights, and the rest are set to negative infinity. The context prompt is the mean of all tokens weighted by the KCA.

all samples of each class as follows:

$$c_i = \frac{1}{|S_i|} \sum_{j=1}^M f(x_j), \quad (1)$$

where $x_j \in S_i$. S_i denotes the i -th support class. $|S_i|$ represents the number of samples in each support class. f is the feature extractor.

For a query image q_i , the similarity score between q_i and all support classes is determined by calculating the distance function between q_i and N class prototypes. The probability of q_i belonging to the n -th support class is calculated as follows:

$$p_k = \frac{\exp(\gamma \cdot d(f(q_i), c_i))}{\sum_{j=1}^N \exp(\gamma \cdot d(f(q_i), c_j))}, \quad (2)$$

where γ is a temperature parameter. $d()$ denotes the distance function. The support class corresponding to the maximum probability is regarded as the classification result.

3.3 k -NN Context Attention

Following[55], the vanilla self-attention formula is defined as:

$$\hat{V} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \quad (3)$$

The fully connected self-attention mechanism allocates weights to all tokens without distinguishing noise, making ViTs susceptible to noisy tokens when only a limited number of samples are available. To mitigate this issue, we propose a K -NN Context Attention (KCA) mechanism, which aims to progressively filter out irrelevant tokens in a coarse-fine manner via three cascaded stages.

Specifically, the KCA computes the attention matrix $A \in \mathbb{R}^{n \times n}$ by the dot product of all query-key pairs, like vanilla attention:

$$A = \frac{QK^T}{\sqrt{d}}. \quad (4)$$

We further construct a masked attention matrix $\hat{A} \in \mathbb{R}^{n \times n}$, selecting the k most relevant keys for each query, as shown in Fig. 3. This is achieved by performing the top- k operation on each row of the attention matrix to obtain the highest k scores and their indexes.

In the $\hat{A}$, the selected elements remain unchanged and all others are set to negative infinity. The formulation is defined as follows:

$$\hat{A}_{ij} = \text{Mask}_{k-nn}(A_{ij}) = \begin{cases} A_{ij} & \text{if } (i, j) \text{ in } I \in \mathbb{R}^{n \times k} \\ -\infty & \text{otherwise} \end{cases}, \quad (5)$$

where $I \in \mathbb{R}^{n \times k}$ represents the collection of $n \times k$ indexes.

Therefore, the KCA naturally forms a sparse attention matrix, concentrating the focus of the ViT on class-relevant features. In the proposed framework, image patches are sequentially fed into three stages to perform the KCA, thereby progressively filtering out irrelevant tokens by selecting similar tokens and refining features in a coarse-fine manner.

In vanilla self-attention, attention weights are calculated based on the relationships between the position and all other positions, resulting in $O(N^2)$ computational complexity for a sequence of length N . In contrast, the KCA considers only the K nearest neighbors for each position, reducing computational complexity to approximately $O(NK)$, where K is much smaller than N . This approximation significantly reduces computational costs and maintains its relatively high performance.

However, a potential risk is the over-focus on local features to the detriment of global context. This limits the generalization of the ViT to new classes with few samples. To address this problem, an average representation of the entire tokens is used as the context prompt to capture global information:

$$z_{cp} = \text{Average}(\text{softmax}(\hat{A}_{ij})V), \quad (6)$$

where z_{cp} represents context prompts. Incorporating z_{cp} into each token can enhance global context information.

The whole KCA formula can be defined as:

$$\tilde{V} = \text{softmax}(\text{Mask}_{k-nn}(\frac{QK^T}{\sqrt{d}}))V + \alpha z_{cp}, \quad (7)$$

where α is a learnable parameter, allowing the ViT to adaptively adjust the coefficient between local and global context information.

3.4 Pyramid Cross-modal Prompts

The proposed Pyramid Cross-modal Prompts (PCP) leverage semantic embeddings to enhance support representations, allowing the ViT to adaptively modulate support features based on semantic information, capturing complex semantic correlations between textual and visual modalities via deep cross-modal interactions. The key component of PCP is the cross-modal enhancement module which is illustrated in Fig. 4. It illustrates cross-modal attention to leverage semantic information and generate semantic-based visual features.

Specifically, to exploit rich and unbiased semantic information, class-aware prompts $Z_t^{cp} \in \mathbb{R}^{L_t \times d_t}$ from class names are obtained by the CLIP which has an excellent capability of cross-modal feature alignments. Additionally, three support features $Z_v^{Si} \in \mathbb{R}^{L_v \times d_v}$ at different scales are extracted via three stages.

Given weight matrices $W_{Q_t} \in \mathbb{R}^{d_t \times d}$, $W_{K_v} \in \mathbb{R}^{d_v \times d}$, and $W_{V_v} \in \mathbb{R}^{d_v \times d}$, the query Q , key K , and value V in the cross-modal attention

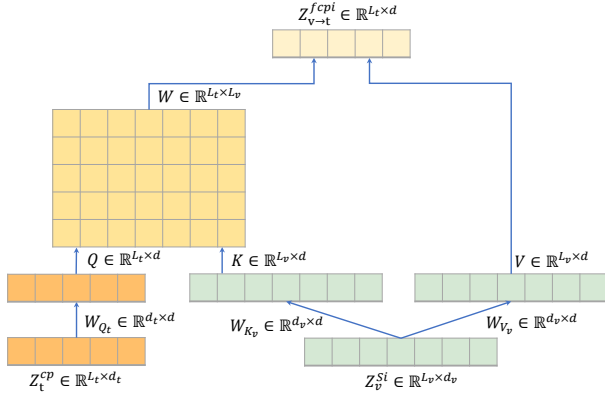


Figure 4: Illustration of cross-modal enhancement module in the PCA. Class-aware prompts Z_t^{cp} provide queries Q , dynamically adjusting the importance weights of visual features.

are computed as follows:

$$\begin{aligned} Q &= Z_t^{cp} W_{Q_t} \in \mathbb{R}^{L_t \times d}, \\ K &= Z_v^{si} W_{K_v} \in \mathbb{R}^{L_v \times d}, \\ V &= Z_v^{si} W_{V_v} \in \mathbb{R}^{L_v \times d}. \end{aligned} \quad (8)$$

Next, the attention weight matrix of textual modality for visual modality is obtained as:

$$W_{v \rightarrow t} = \text{softmax}(QK^T \beta), \quad (9)$$

where $W_{v \rightarrow t}$ is the attention weight matrix and β is a scaling parameter. Due to discrepancies in modality distributions, the initial training phase may yield limited relevance between textual and visual features, resulting in small values within the weight matrix. To enhance the learning of model parameters, we introduce the hyperparameter β to scale the matrix before applying softmax.

After applying the identical operation to the support features of the remaining two scales, we derive the pyramid prompts Z_v^{fpp} , representing the inherent semantic details of visual features chosen by textual features across various levels. Utilizing the multi-scale operation, ViT captures diverse spatial information, thereby enhancing its capability to adapt to spatial variations.

The pyramid prompts are linearly mapped to the dimension of the support feature Z_v^{S3} for fusion, defined as:

$$\widetilde{Z}_v^S = Z_v^{S3} + \lambda \sum_{i=1}^N \text{Linear}(Z_v^{fppi}), \quad (10)$$

where $\widetilde{Z}_v^S$ represents the support features emphasized by the pyramid prompts, N is typically set to 3, and λ as a learnable parameter.

The augmented support features $\widetilde{Z}_v^S$ and class-aware prompts Z_t^{cp} are interacted in the ViT via the KCA. The output y_v^S is calculated as follows:

$$y_v^S = \text{KCA} \left[\left(\widetilde{Z}_v^S \parallel Z_t^{cp} \right) \right], \quad (11)$$

where $\parallel$ represents the concatenation operation along the spatial dimension. The KCA enables the ViT to extract discriminative

support features due to the semantically guided capture of class-specific features, followed by obtaining more generalized class prototypes for classification via Eq. 1.

The similarity scores between the query samples and class prototypes are computed via the cosine distance, defined as follows:

$$p_k = \frac{\exp(\gamma \cdot \cos(f(q_i), c_i))}{\sum_{j=1}^N \exp(\gamma \cdot \cos(f(q_i), c_j))}, \quad (12)$$

where $\cos()$ denotes the cosine distance function. The classification results of the query image q_i is the support class with maximum probability.

The ViT is trained by minimizing the Cross-Entropy (CE) loss function, which can be formulated as follows:

$$\Gamma = \arg \min \sum_{i=1}^M \mathcal{L}^{\text{CE}}(f(q_i), y_i), \quad (13)$$

where $\mathcal{L}^{\text{CE}}$ represents CE loss. $f()$ represents the ViT backbone. q_i and y_i donate a query image and its corresponding ground-truth label.

4 EXPERIMENTS

We first present the four benchmark datasets. Next, we introduce the details of the experiments, followed by the few-shot classification performance and the model analysis.

4.1 Datasets

The effectiveness of the proposed method is evaluated on four few-shot datasets: miniImageNet [56], tieredImageNet [47], CIFAR-FS [30], and FC100 [44]. MiniImageNet and TieredImageNet are both subsets of ImageNet [11]. MiniImageNet has 64 training classes, 16 validation classes, and 20 test classes. TieredImageNet includes 351 training classes, 97 validation classes, and 160 test classes. CIFAR-FS and FC100 are derived from the CIFAR-100 dataset [27]. CIFAR-FS has 64 training classes, 16 validation classes, and 20 test classes employing a random partitioning strategy. FC100 adopts a unique superclass partitioning method, featuring 12 superclasses in the training set (equivalent to 60 classes), and 4 superclasses each in the validation and test sets, totaling 20 classes. Notably, only the training set is utilized for model training, with no overlap between the training, validation, and test sets in few-shot learning settings.

4.2 Implementation Details

Backbone. In all experimental setups, we employ Visformer-T [8] as the feature extractor. The Visformer-T model boasts a more small number of parameters, rendering our model smaller compared to Resnet-12 models [48, 50, 68], or even several times smaller than the ViT-S/16 and WRN-28 models [20, 49, 57]. KCA Block consists of two Batchnorm layers, a two-layer MLP, and Multi-Head KCA. Furthermore, we employ ViT-B/32 CLIP as a text encoder, extracting an output dimension of 512.

Training Details. Our approach follows a two-phase training procedure, similar to traditional frameworks such as Meta-baseline [7], including pre-training and meta-tuning stages. Throughout both phases, KCA serves as the attention mechanism for the ViT. Moreover, PCP is employed in the meta-tuning phase. For input

Table 1: Results (%) for the 5-way 1-shot and 5-way 5-shot settings on MiniImageNet and TieredImageNet. The average accuracy with 95% confidence interval are reported. Bold font indicates the best results. Blue font indicates the suboptimal results.

Methods	Venue	Backbone	Params	MiniImageNet 5-way		TieredImageNet 5-way	
				1-shot	5-shot	1-shot	5-shot
ProtoNet [50]	NeurIPS'2017	ResNet-12	12.4M	62.39 $\pm$ 0.21	80.53 $\pm$ 0.14	68.23 $\pm$ 0.23	84.03 $\pm$ 0.16
KTN [45]	ICCV'2019	ResNet-12	12.4M	61.42 $\pm$ 0.72	74.16 $\pm$ 0.56	-	-
AM3 [61]	NeurIPS'2019	ResNet-12	12.4M	65.30 $\pm$ 0.49	78.10 $\pm$ 0.36	69.08 $\pm$ 0.47	82.58 $\pm$ 0.31
TRAML [31]	CVPR'2020	ResNet-12	12.4M	67.10 $\pm$ 0.52	79.54 $\pm$ 0.60	-	-
DeepEMD [68]	CVPR'2020	ResNet-12	12.4M	65.91 $\pm$ 0.82	82.41 $\pm$ 0.56	71.16 $\pm$ 0.87	86.03 $\pm$ 0.58
Meta-Baseline [7]	ICCV'2021	ResNet-12	12.4M	63.17 $\pm$ 0.23	79.26 $\pm$ 0.17	68.62 $\pm$ 0.27	83.29 $\pm$ 0.18
DeepBDC [60]	CVPR'2022	ResNet-12	12.4M	67.34 $\pm$ 0.43	84.46 $\pm$ 0.28	72.34 $\pm$ 0.49	87.31 $\pm$ 0.32
SVAE [62]	CVPR'2022	ResNet-12	12.4M	74.84 $\pm$ 0.23	83.28 $\pm$ 0.40	76.98 $\pm$ 0.65	85.77 $\pm$ 0.50
Meta-AdaM [52]	NeurIPS'2023	ResNet-12	12.4M	59.89 $\pm$ 0.49	77.92 $\pm$ 0.43	65.31 $\pm$ 0.48	85.24 $\pm$ 0.35
ProtoDiff [14]	NeurIPS'2023	ResNet-12	12.4M	66.63 $\pm$ 0.21	83.48 $\pm$ 0.15	72.95 $\pm$ 0.24	85.15 $\pm$ 0.18
ESPT [48]	AAAI'2023	ResNet-12	12.4M	68.36 $\pm$ 0.19	84.11 $\pm$ 0.12	72.68 $\pm$ 0.22	87.49 $\pm$ 0.14
FGFL [9]	ICCV'2023	ResNet-12	12.4M	69.14 $\pm$ 0.80	86.01 $\pm$ 0.62	73.21 $\pm$ 0.88	87.21 $\pm$ 0.61
BMI [35]	MM'2023	ResNet-12	12.4M	72.96 $\pm$ 0.36	81.94 $\pm$ 0.29	75.45 $\pm$ 0.44	84.77 $\pm$ 0.33
4S [40]	MM'2023	ResNet-12	12.4M	74.53 $\pm$ 0.68	85.78 $\pm$ 0.49	-	-
MetaDiff [67]	AAAI'2024	ResNet-12	12.4M	64.99 $\pm$ 0.77	81.21 $\pm$ 0.56	72.33 $\pm$ 0.92	86.31 $\pm$ 0.62
ALFA [3]	TPAMI'2024	ResNet-12	12.4M	66.61 $\pm$ 0.28	81.43 $\pm$ 0.25	70.29 $\pm$ 0.40	86.17 $\pm$ 0.35
LastShot [65]	TPAMI'2024	ResNet-12	12.4M	67.35 $\pm$ 0.20	82.58 $\pm$ 0.14	72.43 $\pm$ 0.23	85.82 $\pm$ 0.16
LEO [49]	ICLR'2019	WRN-28	36.5M	61.76 $\pm$ 0.08	77.59 $\pm$ 0.12	66.33 $\pm$ 0.05	81.44 $\pm$ 0.09
Align [1]	ECCV'2022	WRN-28	36.5M	65.92 $\pm$ 0.60	82.85 $\pm$ 0.55	74.40 $\pm$ 0.68	86.61 $\pm$ 0.59
AMTNet [28]	MM'2022	WRN-28	36.5M	70.05 $\pm$ 0.46	84.55 $\pm$ 0.29	73.86 $\pm$ 0.50	87.62 $\pm$ 0.33
RankDNN [18]	AAAI'2023	WRN-28	36.5M	66.67 $\pm$ 0.15	84.79 $\pm$ 0.11	74.00 $\pm$ 0.15	88.80 $\pm$ 0.25
STANet [29]	AAAI'2023	WRN-28	36.5M	69.86 $\pm$ 0.46	85.16 $\pm$ 0.29	74.41 $\pm$ 0.50	87.64 $\pm$ 0.33
SUN [13]	ECCV'2022	Visformer-S	12.4M	67.80 $\pm$ 0.45	83.25 $\pm$ 0.30	72.99 $\pm$ 0.50	86.74 $\pm$ 0.33
FewTure [20]	NeurIPS'2022	ViT-S/16	22.0M	68.02 $\pm$ 0.88	84.51 $\pm$ 0.53	72.96 $\pm$ 0.92	86.43 $\pm$ 0.67
SP [6]	CVPR'2023	Visformer-T	10.3M	72.31 $\pm$ 0.40	83.42 $\pm$ 0.30	78.03 $\pm$ 0.46	88.55 $\pm$ 0.32
KTPP	ours	Visformer-T	11.1M	76.71 $\pm$ 0.37	86.46 $\pm$ 0.27	80.80 $\pm$ 0.43	90.01 $\pm$ 0.29

images with a resolution of 224×224 , Visformer-T sequentially extracts feature dimensions of 96, 192, and 384. We utilize the AdamW optimizer [39] with a learning rate of $5e-4$. During the pre-training phase, the number of training epochs is set to 500 for the mini-ImageNet, CIFAR-FS, and FC100 datasets, and 300 epochs for the tieredImageNet dataset. A batch size of 512 is employed, and a cosine learning rate scheduler [38] is utilized for learning rate adjustment. In the meta-tuning phase, the model is trained for 100 epochs using the episodic training strategy. All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU, utilizing PyTorch for code implementations.

Evaluation Protocol. The proposed method is evaluated in 5-way 1-shot and 5-way 5-shot settings. Each class has 15 test instances. We randomly sampled 2,000 episodes from the test class. The average accuracy with 95% confidence interval is reported.

4.3 Comparison with the SOTA Methods

To evaluate the effectiveness of our proposed KTPP, extensive experiments are conducted on four benchmark datasets. Tabs. 1 and 2 present the results of KTPP compared to state-of-the-art methods under 5-way 1-shot and 5-way 5-shot settings.

KTPP demonstrates substantial superiority over the counterparts across all settings. Specifically, in the 1-shot settings, KTPP exhibits

notable advancements, outperforming state-of-the-art results by 1.45% to 3.06% across the four benchmarks. KTPP significantly outperforms previous textual modality-based methods (AM3 [61], KTN [45], TRAML [31], SVAE [62], and SP [6]), improving 1-shot accuracy by 4.4% and 5-shot accuracy by 3.04% on miniImageNet compared to SP. In addition, Compared to previous ViT-based methods (SUN [13], FewTure [20]), KTPP achieves superior results with fewer parameters, improving 1-shot accuracy by 7.81% and 5-shot accuracy by 3.27% on TieredImageNet.

The effectiveness of KTPP is confirmed by the results. This is attributed to the ability of KTPP to progressively filter out class-irrelevant features in a coarse-fine manner and capture discriminative in cross-modal interactions via the KCA. Moreover, pyramid prompts learned in the PCP contribute to enhancing visual features through deep cross-modal interactions, improving adaptability to spatial variations via the pyramid structure.

4.4 Model Analysis

4.4.1 Ablation Study. To fairly prove the effectiveness of our method, we leverage the ViT as the baseline. KCA and PCP are introduced to investigate their impact, as shown in Table 3. First KCA is integrated as the attention mechanism into the baseline, improving the 1-shot and 5-shot accuracy by 3.25% and 2.89% in

Table 2: Results (%) for the 5-way 1-shot and 5-way 5-shot settings on CIFAR-FS and FC100. The average accuracy with 95% confidence interval are reported. Bold font indicates the best results. Blue font indicates the sub-optimal results.

Methods	Venue	Backbone	Params	CIFAR-FS 5-way		FC100 5-shot	
				1-shot	5-shot	1-shot	5-shot
ProtoNet [50]	NeurIPS'2017	ResNet-12	12.4M	72.20 $\pm$ 0.70	83.50 $\pm$ 0.50	41.54 $\pm$ 0.76	57.08 $\pm$ 0.76
TADAM [44]	NeurIPS'2018	ResNet-12	12.4M	-	-	40.10 $\pm$ 0.40	56.10 $\pm$ 0.40
MABAS [24]	ECCV'2020	ResNet-12	12.4M	72.80 $\pm$ 0.70	84.30 $\pm$ 0.50	47.20 $\pm$ 0.60	55.50 $\pm$ 0.60
Meta-NVG [69]	ICCV'2021	ResNet-12	12.4M	73.51 $\pm$ 0.92	85.65 $\pm$ 0.65	42.31 $\pm$ 0.75	58.16 $\pm$ 0.78
Meta-AdaM [52]	NeurIPS'2023	ResNet-12	12.4M	-	-	41.12 $\pm$ 0.49	56.14 $\pm$ 0.49
ALFA [3]	TPAMI'2024	ResNet-12	12.4M	76.32 $\pm$ 0.43	86.73 $\pm$ 0.31	44.54 $\pm$ 0.50	58.44 $\pm$ 0.42
LastShot [65]	TPAMI'2024	ResNet-12	12.4M	76.76 $\pm$ 0.21	87.49 $\pm$ 0.12	44.08 $\pm$ 0.18	59.14 $\pm$ 0.18
PN+rot [17]	ICCV'2019	WRN-28	36.5M	69.55 $\pm$ 0.34	82.34 $\pm$ 0.24	-	-
Align [1]	ECCV'2022	WRN-28	36.5M	-	-	45.83 $\pm$ 0.48	59.74 $\pm$ 0.56
AMTNet [28]	MM'2022	WRN-28	36.5M	80.38 $\pm$ 0.48	89.89 $\pm$ 0.32	-	-
SUN [13]	ECCV'2022	Visformer-S	12.4M	78.37 $\pm$ 0.46	88.84 $\pm$ 0.32	-	-
FewTURE [20]	NeurIPS'2022	ViT-S/16	22.0M	72.80 $\pm$ 0.88	86.14 $\pm$ 0.64	46.20 $\pm$ 0.79	63.14 $\pm$ 0.73
SP [6]	CVPR'2023	Visformer-T	10.3M	82.18 $\pm$ 0.40	88.24 $\pm$ 0.32	48.53 $\pm$ 0.38	61.55 $\pm$ 0.41
KTPP	ours	Visformer-T	11.1M	83.63 $\pm$ 0.57	90.19 $\pm$ 0.30	51.59 $\pm$ 0.40	65.18 $\pm$ 0.40

Table 3: Ablation study of KTPP on miniImageNet. “KCA” means K-NN Context Attention. “PCP” denotes Pyramid Cross-modal Prompts.

Training phase		Module		1-shot	5-shot
Pre	Meta	KCA	PCP		
✓				63.84 $\pm$ 0.38	80.72 $\pm$ 0.31
✓	✓			65.56 $\pm$ 0.37	81.62 $\pm$ 0.31
✓		✓		67.09 $\pm$ 0.37	83.61 $\pm$ 0.30
✓	✓	✓		68.75 $\pm$ 0.38	84.92 $\pm$ 0.29
✓	✓		✓	73.80 $\pm$ 0.38	84.34 $\pm$ 0.28
✓	✓	✓	✓	76.71 $\pm$ 0.37	86.46 $\pm$ 0.27

the pre-training phase. Next, KCA is used in the meta-training phase, increasing the 1-shot and 5-shot accuracy by 3.19% and 3.3%. This is because KCA can select class-specific features, progressively filtering out irrelevant features in a coarse-fine manner in three cascaded stages. Then, PCP is employed in the meta-tuning phase to learn pyramid prompts via deep interactions between modalities, which allows the ViT to dynamically modulate visual features based on semantic information, improving adaptability to spatial variations via the pyramid structure. Therefore, it is observed there are 8.24% and 2.72% improvements in the 1-shot and 5-shot tasks, respectively. Finally, when KCA and PCP together are incorporated into the baseline, a significant improvement of 11.14% and 4.84% can be observed, due to further capture of discriminative features guided by semantic information.

4.4.2 The effect of value k in KCA. Table 4 shows the different values of k in KCA. “ $k = 1$ ” means that each query only selects one most relevant key to compute the attention matrix. smaller values of k make it difficult to distill noisy tokens due to information interactions between irrelevant and relevant tokens. large values of k make KCA tend to be a vanilla self-attention mechanism, both of which lead to performance degradation. For Visformer-T, $k = 10$

Table 4: The effect of value k .

The value of k	1-shot	5-shot
$k = 1$	71.55 $\pm$ 0.35	83.65 $\pm$ 0.29
$k = 5$	75.13 $\pm$ 0.37	84.49 $\pm$ 0.28
$k = 7$	75.50 $\pm$ 0.35	85.69 $\pm$ 0.28
$k = 10$	76.71 $\pm$ 0.37	86.46 $\pm$ 0.27
$k = 15$	76.21 $\pm$ 0.38	85.92 $\pm$ 0.27
$k = 20$	75.81 $\pm$ 0.38	84.81 $\pm$ 0.29
$k = 30$	75.01 $\pm$ 0.37	84.39 $\pm$ 0.30

is optimal on miniImageNet, which enables the ViT to distinguish class-irrelevant tokens more effectively.

Table 5: The performance of different backbone architectures. * denotes results produced by our model.

method	backbone	1-shot	5-shot
ProtoNet [50]	Resnet-12	62.39 $\pm$ 0.21	80.53 $\pm$ 0.14
ProtoNet* [50]	Visformer-T	62.48 $\pm$ 0.45	79.78 $\pm$ 0.30
Meta-Baseline [7]	ResNet-12	63.17 $\pm$ 0.23	79.26 $\pm$ 0.17
Meta-Baseline* [7]	Visformer-T	62.59 $\pm$ 0.40	79.88 $\pm$ 0.35
ours	Visformer-T	76.71 $\pm$ 0.37	86.46 $\pm$ 0.27

4.4.3 Backbone Architectures. Table 5 presents two baseline approaches based on Resnet-12 and corresponding reimplementations based on Visformer-T. Results are reported on miniImageNet. It is evident that directly replacing ResNet-12 with Visformer-T fails to improve results. Instead, our KTPP significantly surpasses baseline approaches with the identical Visformer Architecture due to select discriminative information and adaptively module visual features.

Table 6: Comparison with state-of-the-art prompt and sparse attention methods on miniImageNet.

	Method	1-shot	5-shot
Prompt	VPT [22]	66.89 $\pm$ 0.35	83.42 $\pm$ 0.27
	SP [6]	73.28 $\pm$ 0.40	85.23 $\pm$ 0.28
Sparse attention	KVT [58]	69.57 $\pm$ 0.44	84.78 $\pm$ 0.29
ours	KTPP	76.71 $\pm$ 0.37	86.46 $\pm$ 0.27

4.4.4 Analysis of Prompt and Sparse Attention methods. Table 6 shows the results of different Prompt and Sparse Attention methods. To fairly compare performance, we use the same ViT architectures in the FSL settings. We simply replace PCP with VPT [22] and SP [6] in the meta-tuning phase. However, VPT lacks semantic information by directly introducing some prompt parameters. Moreover, SP struggles to capture complex semantic relationships due to direct fusion. Instead, KTPP employs PCP to learn pyramid prompts, which allow the ViT to adaptively adjust visual features based on semantic information, making it robust to spatial variations via pyramid structure. Next, we directly replace KCA with KVT [58] in the whole training process. However, KVT has a poor generalization to new classes in sparse scenarios with only few labeled samples. Unlike KVT, we utilize semantic information to further extract discriminative visual features via cross-modal interaction in KCA. Meanwhile, the mean of all tokens as context prompts to enhance the global context information of each token. As a result, Our KTPP significantly outperforms the state-of-the-art prompt and sparse attention methods.

Table 7: The effect of different classifiers on the FSL test.

Metrics for classifier	1-shot	5-shot
EU	73.47 $\pm$ 0.35	83.53 $\pm$ 0.30
CO	76.33 $\pm$ 0.37	86.16 $\pm$ 0.30
LR	76.71 $\pm$ 0.37	86.46 $\pm$ 0.27

4.4.5 Classifier Selection Strategy. Table 7 shows three different classifiers and the results on miniImageNet are reported. Classification is achieved based on the distance between the support and query features in the feature space. Cosine distance to achieve nearest neighbor classification gets better results than Euclidean distance. Moreover, the Linear logistic regression classifier gets the best results due to constructing more support samples.

4.4.6 Image Resolution. Table 8 shows three common image resolutions in FSL. The ViT gets similar results at different resolutions, reflecting the adaptability of the proposed method to image resolution due to the full utilization of diverse spatial information.

4.4.7 Visualization. To qualitatively analyze the effectiveness of KTPP, we illustrate the t-SNE [54] results of baseline and KTPP. Fig. 5 shows the visualization of samples with novel classes on three benchmark datasets. In contrast to the spatially cluttered distribution of baseline, our KTPP allows the ViT to capture class-specific features so that intra-class images are closer together and inter-class images are further apart.

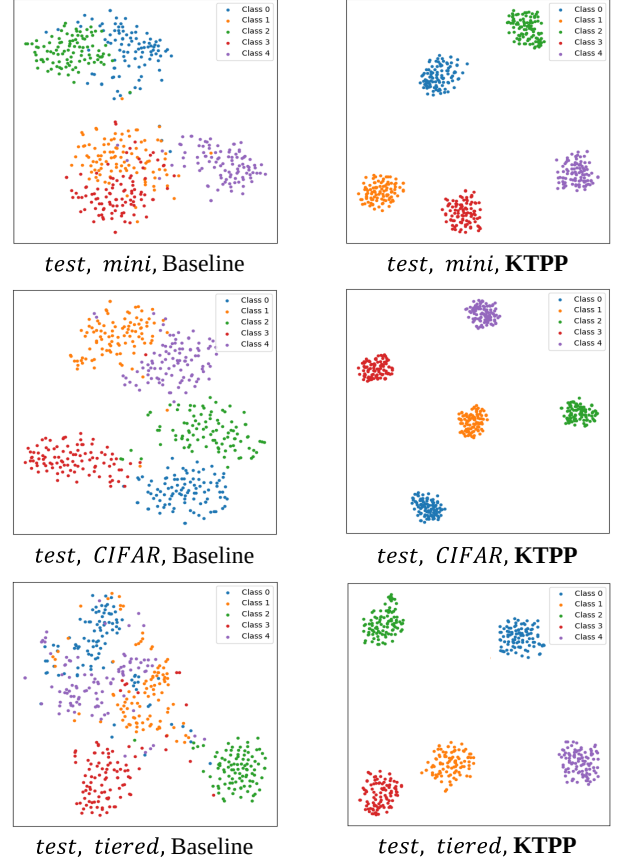
**Figure 5: T-SNE visualization results on novel classes from three benchmark datasets. ViT with KTPP performs better compared to ViT with baseline.****Table 8: The effect of different image resolutions.**

Image resolution	1-shot	5-shot
80 $\times$ 80	76.19 $\pm$ 0.40	86.02 $\pm$ 0.29
84 $\times$ 84	75.86 $\pm$ 0.40	85.86 $\pm$ 0.30
224 $\times$ 224	76.71 $\pm$ 0.37	86.46 $\pm$ 0.27

5 CONCLUSION

In this paper, we propose k -NN Transformer with Pyramid Prompts (KTPP) for Few-Shot Learning. KTPP mainly consists of k -NN Context Attention (KCA) and Pyramid Cross-modal Prompts (PCP). Specifically, KCA achieves progressively noisy filtration via a coarse-to-fine manner in the three cascaded stages and extracts discriminative features guided by semantic information. In PCP, pyramid prompts are generated to enhance visual features via deep cross-modal interactions between textual and multi-scale visual features. This enables the ViT to dynamically module the importance weights of visual features based on semantic information and make it robust to spatial variations via pyramid structure. Experimental results demonstrate the effectiveness of our KTPP model.

REFERENCES

- [1] Arman Afraasiyabi, Jean-François Lalonde, and Christian Gagné. 2020. Associative alignment for few-shot image classification. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V* 16. Springer, 18–35.
- [2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6836–6846.
- [3] Sungyong Baik, Myungsub Choi, Janghoon Choi, Heewon Kim, and Kyoung Mu Lee. 2024. Learning to Learn Task-Adaptive Hyperparameters for Few-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 3 (2024), 1441–1454. <https://doi.org/10.1109/TPAMI.2023.3261387>
- [4] Yun-Hao Cao, Hao Yu, and Jianxin Wu. 2022. Training vision transformers with only 2040 images. In *European Conference on Computer Vision*. Springer, 220–237.
- [5] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 357–366.
- [6] Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, and Tieniu Tan. 2023. Semantic Prompt for Few-Shot Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23581–23591.
- [7] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. 2021. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9062–9071.
- [8] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. 2021. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 589–598.
- [9] Hao Cheng, Siyuan Yang, Joey Tianyi Zhou, Lanqing Guo, and Bihan Wen. 2023. Frequency guidance matters in few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11814–11824.
- [10] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*. PMLR, 1931–1942.
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [12] Carl Doersch, Ankush Gupta, and Andrew Zisserman. 2020. Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems* 33 (2020), 21981–21993.
- [13] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. 2022. Self-promoted supervision for few-shot transformer. In *European Conference on Computer Vision*. Springer, 329–347.
- [14] Yingjun Du, Zehao Xiao, Shengcai Liao, and Cees G. M. Snoek. 2023. ProtoDiff: Learning to Learn Prototypical Networks by Task-Guided Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=lp9GR2t3hn>
- [15] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. 2023. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17481–17490.
- [16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. PMLR, 1126–1135.
- [17] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. 2019. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8059–8068.
- [18] Qianyu Guo, Gong Haotong, Xujun Wei, Yanwei Fu, Yizhou Yu, Wenqiang Zhang, and Weifeng Ge. 2023. Rankdnn: Learning to rank for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 728–736.
- [19] Fusheng Hao, Fengxiang He, Jun Cheng, and Dacheng Tao. 2021. Global-local interplay in semantic alignment for few-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 7 (2021), 4351–4363.
- [20] Markus Hiller, Rongkai Ma, Mehrtash Harandi, and Tom Drummond. 2022. Rethinking generalization in few-shot classification. *Advances in Neural Information Processing Systems* 35 (2022), 3582–3595.
- [21] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. 2019. Cross attention network for few-shot classification. *Advances in neural information processing systems* 32 (2019).
- [22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*. Springer, 709–727.
- [23] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19113–19122.
- [24] Jaekyeom Kim, Hyoungseok Kim, and Gunhee Kim. 2020. Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* 16. Springer, 599–617.
- [25] Soopil Kim, Sion An, Philip Chikontwe, Myeongkyun Kang, Ehsan Adeli, Kilian M Pohl, and Sang Hyun Park. 2024. Few shot part segmentation reveals compositional logic for industrial anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8591–8599.
- [26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).
- [27] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
- [28] Jinxiang Lai, Siqian Yang, Guannan Jiang, Xi Wang, Yuxi Li, Zihui Jia, Xiaochen Chen, Jun Liu, Bin-Bin Gao, Wei Zhang, et al. 2022. Rethinking the Metric in Few-shot Learning: From an Adaptive Multi-Distance Perspective. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4021–4030.
- [29] Jinxiang Lai, Siqian Yang, Wenlong Wu, Tao Wu, Guannan Jiang, Xi Wang, Jun Liu, Bin-Bin Gao, Wei Zhang, Yuan Xie, et al. 2023. SpatialFormer: semantic and target aware attentions for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 8430–8437.
- [30] Kwonjoon Lee, Subhansu Maji, Avinash Ravichandran, and Stefano Soatto. 2019. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10657–10665.
- [31] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. 2020. Boosting few-shot learning with adaptive margin loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12576–12584.
- [32] Aoxue Li, Tiange Luo, Tao Xiang, Weiran Huang, and Liwei Wang. 2019. Few-shot learning with global class representations. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9715–9724.
- [33] Xiaomeng Li, Lequan Yu, Yueming Jin, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. 2020. Difficulty-aware meta-learning for rare disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I* 23. Springer, 357–366.
- [34] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. 2022. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*. Springer, 280–296.
- [35] Zhuoling Li and Yong Wang. 2023. Better Integrating Vision and Semantics for Improving Few-shot Classification. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4737–4746.
- [36] Lu Liu, William L Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle. 2020. A Universal Representation Transformer Layer for Few-Shot Image Classification. In *International Conference on Learning Representations*.
- [37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.
- [38] Ilya Loshchilov and Frank Hutter. 2016. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*.
- [39] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
- [40] Jinda Lu, Shuo Wang, Xinyu Zhang, Yanbin Hao, and Xiangnan He. 2023. Semantic-based Selection, Synthesis, and Supervision for Few-shot Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3569–3578.
- [41] Kushagra Mahajan, Monika Sharma, and Lovekesh Vig. 2020. Meta-DermDiagnosis: Few-Shot Skin Disease Identification Using Meta-Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- [42] Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999* (2018).
- [43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
- [44] Boris Oreshkin, Pau Rodriguez López, and Alexandre Lacoste. 2018. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems* 31 (2018).
- [45] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. 2019. Few-shot image recognition with knowledge transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 441–449.
- [46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [47] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. 2018. Meta-learning for semi-supervised few-shot classification. In *6th International Conference on Learning Representations, ICLR 2018*.
- [48] Yi Rong, Xiongbo Lu, Zhaoyang Sun, Yaxiong Chen, and Shengwu Xiong. 2023. ESPT: a self-supervised episodic spatial pretext task for improving few-shot

- learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 9596–9605.
- [49] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. 2018. Meta-Learning with Latent Embedding Optimization. In *International Conference on Learning Representations*.
- [50] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *NeurIPS* 30 (2017).
- [51] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).
- [52] Siyuan Sun and Hongyang Gao. 2023. Meta-AdaM: An Meta-Learned Adaptive Optimizer with Momentum for Few-Shot Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=d85pPNBHLt>
- [53] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.
- [54] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [56] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems* 29 (2016).
- [57] Heng Wang, Tan Yue, Xiang Ye, Zihang He, Bohan Li, and Yong Li. 2022. Revisit Finetuning strategy for Few-Shot Learning to Transfer the Embeddings. In *The Eleventh International Conference on Learning Representations*.
- [58] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. 2022. Kvt: k-nn attention for boosting vision transformers. In *European conference on computer vision*. Springer, 285–302.
- [59] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. 2021. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8741–8750.
- [60] Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua Li. 2022. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7972–7981.
- [61] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. 2019. Adaptive cross-modal few-shot learning. *Advances in Neural Information Processing Systems* 32 (2019).
- [62] Jingyi Xu and Hieu Le. 2022. Generating representative samples for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9003–9013.
- [63] Kun Yan, Zied Bouraoui, Ping Wang, Shoaib Jameel, and Steven Schockaert. 2021. Aligning visual prototypes with bert embeddings for few-shot learning. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 367–375.
- [64] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. 2020. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8808–8817.
- [65] Han-Jia Ye, Lu Ming, De-Chuan Zhan, and Wei-Lun Chao. 2024. Few-Shot Learning With a Strong Teacher. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 3 (2024), 1425–1440. <https://doi.org/10.1109/TPAMI.2022.3160362>
- [66] Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. 2021. Prototype completion with primitive knowledge for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3754–3762.
- [67] Baoquan Zhang, Chuyao Luo, Demin Yu, Xutao Li, Huiwei Lin, Yunming Ye, and Bowen Zhang. 2024. MetaDiff: Meta-Learning with Conditional Diffusion for Few-Shot Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 16687–16695.
- [68] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. 2020. 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*. 12203–12213.
- [69] Chi Zhang, Henghui Ding, Guosheng Lin, Ruiibo Li, Changhu Wang, and Chunhua Shen. 2021. Meta navigator: Search for a good adaptation policy for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9435–9444.
- [70] Peng Zhao, Qiangchang Wang, and Yilong Yin. 2023. M3R: Masked Token Mixup and Cross-Modal Reconstruction for Zero-Shot Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3161–3171.
- [71] Andrey Zhmoginov, Mark Sandler, and Maksym Vladymyrov. 2022. Hypertransformer: Model generation for supervised and semi-supervised few-shot learning. In *International Conference on Machine Learning*. PMLR, 27075–27098.
- [72] Kaiyang Zhou, Jingkan Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16816–16825.
- [73] Kaiyang Zhou, Jingkan Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.
- [74] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. 2019. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*. PMLR, 7693–7702.