Connecting Giants: Synergistic Knowledge Transfer of Large Multimodal Models for Few-Shot Learning

Hao Tang[1], **Shengfeng He**[2*] and **Jing Qin**[1]
[1]Centre for Smart Health, The Hong Kong Polytechnic University
[2]School of Computing and Information Systems, Singapore Management University
{howard-hao.tang, harry.qin}@polyu.edu.hk, shengfenghe@smu.edu.sg

Abstract

Few-shot learning (FSL) addresses the challenge of classifying novel classes with limited training samples. While some methods leverage semantic knowledge from smaller-scale models to mitigate data scarcity, these approaches often introduce noise and bias due to the data's inherent simplicity. In this paper, we propose a novel framework, Synergistic Knowledge Transfer (SYNTRANS), which effectively transfers diverse and complementary knowledge from large multimodal models to empower the off-the-shelf few-shot learner. Specifically, SYNTRANS employs CLIP as a robust teacher and uses a few-shot vision encoder as a weak student, distilling semantic-aligned visual knowledge via an unsupervised proxy task. Subsequently, a training-free synergistic knowledge mining module facilitates collaboration among large multimodal models to extract high-quality semantic knowledge. Building upon this, a visual-semantic bridging module enables bi-directional knowledge transfer between visual and semantic spaces, transforming explicit visual and implicit semantic knowledge into category-specific classifier weights. Finally, SYNTRANS introduces a visual weight generator and a semantic weight reconstructor to adaptively construct optimal multimodal FSL classifiers. Experimental results on four FSL datasets demonstrate that SYNTRANS, even when paired with a simple few-shot vision encoder, significantly outperforms current state-of-the-art methods.

1 Introduction

Deep learning models have achieved remarkable success in numerous computer vision tasks [Li *et al.*, 2019]. However, their effectiveness typically relies on deep neural architectures [He *et al.*, 2016] and large-scale training datasets [Russakovsky *et al.*, 2015], which hinders their applicability in real-world scenarios where annotated data are scarce. In contrast, humans exhibit an exceptional ability to acquire new concepts and recognize categories from only a handful of samples, aided by extensive prior knowledge and contextual understanding [Ralph *et al.*, 2017]. This gap has motivated researchers to investigate few-shot learning (FSL) [Tang *et al.*, 2020; Wu *et al.*, 2022], where a model classifies query samples into one of N novel classes, each provided with only K labeled examples.

The effectiveness of FSL heavily relies on leveraging prior knowledge to address data scarcity. Existing methods commonly transfer knowledge [Tang *et al.*, 2022; Zha *et al.*, 2023] from a disjoint base dataset to novel classes. Early works primarily focused on efficiently exploiting visual prior knowledge, including metric-based [Snell *et al.*, 2017] and optimization-based paradigms [Ravi and Larochelle, 2016], both striving to train a base learner capable of rapid adaptation to novel classes with limited training data. While these methods have achieved promising results, there remains a huge gap in comparison to how humans utilize accumulated knowledge and experiences. As a result, semantic-based methods have emerged to explore various types of semantic knowledge as auxiliary information to improve FSL performance. This semantic knowledge can be obtained either manually (*i.e.,* attribute annotations) or automatically (*i.e.,* word vectors). Unfortunately, acquiring attribute annotations requires substantial human effort and may be infeasible for large-scale datasets, while word vectors derived from a single class name tend to be noisy or lack contextual richness. Hence, how to effectively collect and utilize high-quality prior knowledge in FSL is worthy of further investigation.

Perceptual filling-in [Neumann *et al.*, 2001] is a fundamental characteristic of the human visual system, in which the brain employs prior knowledge and contextual cues to intuitively "*fill in*" missing information, resulting in a coherent and comprehensive perception. This phenomenon becomes particularly apparent in scenarios where visual stimuli are limited or partially obscured, enabling a seamless visual experience despite incomplete data. Inspired by this phenomenon, we hypothesize that transferring rich external knowledge to the off-the-shelf few-shot learner can further improve performance. Recently, Large Multimodal Models (LMMs) [Ouyang *et al.*, 2022; Radford *et al.*, 2021] containing abundant implicit knowledge have emerged as powerful repositories of external knowledge in various domains. These models encompass extensive understandings of the vi-

* Corresponding author.

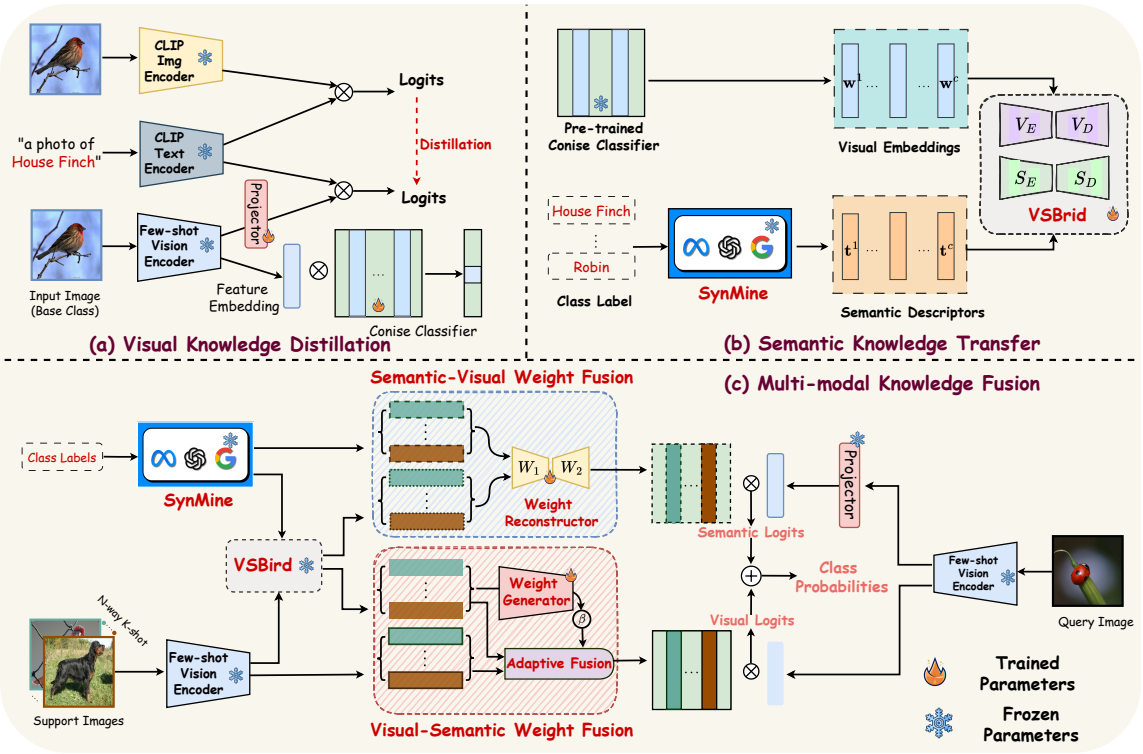


Figure 1: The pipeline of the proposed **Syn**ergistic Knowledge **Trans**fer (SYNTRANS) framework.

sual world, language, and inter-entity relationships, encapsulating diverse knowledge and information. A natural question thus arises: **can we replicate this human-like cognitive process by connecting the rich, multimodal knowledge of these "giants" to compensate for incomplete visual data?**

In this paper, we propose a Synergistic Knowledge Transfer (SYNTRANS) framework to harness the extensive knowledge embedded in large multimodal models to empower the small few-shot learner. Three key challenges arise in achieving this goal: ❶ *effectively distilling desired visual and semantic knowledge from these models*, ❷ *transforming explicit or implicit knowledge into a usable form*, and ❸ *adaptively integrating them with limited visual data to improve FSL performance*. As illustrated in Figure 1, SYNTRANS addresses these challenges in three stages. First, we introduce a vast CLIP model as a strong teacher, adding a linear projection layer after the frozen few-shot vision encoder as a weak student to distill semantic-aligned visual knowledge via an unsupervised proxy task. Next, our Synergistic Knowledge Mining (SynMine) module exploits a large language model to generate comprehensive text descriptions by tapping into the implicit knowledge through chain-of-thought prompting. These descriptions are then refined by a visual-language model into rich semantic descriptors, producing deeper, context-aware understanding of class characteristics. Central to the semantic transfer stage is the Visual-Semantic Bridging (VSBrid) module, which leverages a dual encoder-decoder design to facilitate bi-directional knowledge transfer between the visual and semantic spaces, ultimately mapping these high-quality visual embeddings and semantic descriptors to practical class-specific classifier weights. Finally, a visual weight generator and a semantic weight reconstructor are incorporated for dynamic visual-semantic knowledge fusion, constructing robust and adaptable multimodal classifiers.

We evaluate SYNTRANS on four benchmark datasets, demonstrating its state-of-the-art performance even when equipped with a simple few-shot vision encoder. To the best of our knowledge, SYNTRANS is the first framework that systematically integrates knowledge from large multimodal models to empower small few-shot learners, opening new avenues for bridging the gap between human-like intuition and machine learning in FSL.

2 Related Works

Visual-based FSL Methods. Visual-based FSL methods [Tang *et al.*, 2023; Fu *et al.*, 2023; Fu *et al.*, 2024] transfer prior visual knowledge from base classes to novel classes. Broadly, these approaches can be categorized into two branches: optimization-based methods and metric-based methods. Their core distinction lies in how they leverage the support set, either by fine-tuning an end-to-end network or by directly generating classifiers for novel classes. Optimization-based methods [Finn *et al.*, 2017; Ravi and Larochelle, 2016] focus on learning an effective initialization or optimization strategy, enabling rapid model adaptation to novel tasks with only a few fine-tuning steps. Despite their flexibility, such methods can face meta-overfitting issues when limited labeled data are available. In contrast, metric-based methods [Vinyals *et al.*, 2016; Snell *et al.*, 2017; Sung *et al.*, 2018] learn a metric space where samples

from the same category lie close together while those from different categories are farther apart. Another emerging paradigm [Chen *et al.*, 2019] involves pre-training a powerful feature extractor on base data and directly generating classifier weights for new classes using, for instance, a cosine classifier [Luo *et al.*, 2018]. However, purely visual approaches may struggle under limited training samples, as real-world data often contain background noise and significant intraclass variation. Consequently, relying solely on visual cues may be insufficient for robust recognition of novel categories. This limitation naturally leads to the question of how high-quality semantic knowledge can be integrated to complement visually dominated FSL methods.

Semantic-based FSL Methods. To overcome the shortcomings of purely visual approaches, semantic-based FSL methods [Li *et al.*, 2023; Lu *et al.*, 2023] leverage auxiliary semantic information such as attributes [Lampert *et al.*, 2009], word embeddings [Xian *et al.*, 2019], or even knowledge graphs [Miller, 1995]. For example, AM3 [Xing *et al.*, 2019] combines label embeddings derived from class names with visual prototypes to generate semantic prototypes, which are then adaptively fused. Knowledge graphs also provide valuable correlation cues: KTN [Peng *et al.*, 2019] builds a graph convolutional network (GCN) with node representations and edges derived from label embeddings and semantic relationships, allowing knowledge transfer from base to novel categories. Nevertheless, these methods often rely on either sparse attribute annotations or word vectors obtained from limited textual sources (*e.g.*, Glove [Pennington *et al.*, 2014], Word2vec [Mikolov *et al.*, 2013]), which may lack sufficient contextual richness. As a result, the potential noise in these external semantics can hinder performance. Encouraged by the rise of large multimodal models [Ouyang *et al.*, 2022; Radford *et al.*, 2021] that encapsulate abundant implicit knowledge, we investigate how to extract and distill higher-quality prior knowledge to complement visual-based FSL. Our experiments show that the desired knowledge distilled from these large multimodal models can significantly boost FSL performance, even when used with a relatively simple pre-trained backbone.

3 Method

3.1 Problem Formulation

We consider the standard FSL setting [Vinyals *et al.*, 2016], where two disjoint sets of classes are given: a *base* set $\mathcal{C}_{base}$ and a *novel* set $\mathcal{C}_{novel}$, such that $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$. The model first trains on samples from $\mathcal{C}_{base}$, denoted by $\mathcal{D}_{base} = \{(x, y)\}$, where each pair (x, y) corresponds to an image x and its one-hot label y drawn from $\mathcal{C}_{base}$. For semantic reference, each label y can be mapped to a specific class name c, such as "`House Finch`" or "`Robin`". During evaluation, we adopt the "N-way K-shot" protocol, which randomly selects N categories from $\mathcal{C}_{novel}$ to construct two subsets: a support set and a query set. The support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N\times K}$ contains K examples for each of the N novel classes, while the query set $\mathcal{Q} = \{(x_i, y_i)\}_{i=1}^{N\times Q}$ includes Q samples from the same N categories. Here, x_i denotes the i-th image, and y_i is one of the novel class labels in $\mathcal{C}_{novel}$. The main objective of FSL is to leverage the base knowledge (learned from $\mathcal{D}_{base}$) along with the limited support samples in $\mathcal{S}$ to classify new query images in $\mathcal{Q}$.

3.2 Overall Framework

As shown in Figure 1, the SYNTRANS framework consists of three stages: visual knowledge distillation, semantic knowledge transfer, and multi-modal knowledge fusion. These stages work synergistically to enhance a few-shot learner by integrating visual and semantic knowledge. **Notably, unlike traditional FSL methods, our method does not require fine-tuning the vision encoder at any stages**. This unique characteristic highlights the efficiency of SYNTRANS, making it readily applicable on top of existing few-shot learners.

Visual Knowledge Distillation. In this stage, we train a lightweight projector f_φ and a cosine classifier f_Φ. The projector f_φ enables SYNTRANS to perform CLIP-like vision-semantic alignment, while the classifier f_Φ distills more task-relevant visual knowledge from the few-shot vision encoder for subsequent knowledge transfer. Specifically, the parameters of f_Φ are optimized using cross-entropy loss, while the parameters of f_φ are optimized via the teacher-student distillation paradigm [Wu *et al.*, 2024].

Semantic Knowledge Transfer. This stage consists of two phases: semantic knowledge mining and visual-semantic knowledge bridging. First, we introduce the train-free Synergistic Knowledge Mining (SynMine) module, which efficiently extracts implicit knowledge from visual- and language models to generate high-quality semantic descriptors. Next, the Visual-Semantic Bridging (VSBird) module facilitates bidirectional knowledge transfer between the visual and semantic spaces, mapping the visual embeddings and semantic descriptors to category-specific classifier weights.

Multi-modal Knowledge Fusion. In this stage, the frozen few-shot learner computes visual prototypes for all classes in the support set. Simultaneously, the VSBird module generates classifier weights based on these visual prototypes and the semantic descriptors. To integrate these weights into robust multimodal few-shot classifiers, we introduce a visual weight generator and a semantic weight reconstructor, which function as meta-learners, adaptively combining both types of classifier weights for current FSL task.

3.3 Visual Knowledge Distillation

In this stage, we introduce a large CLIP teacher model to distill semantic-aligned visual knowledge, empowering the frozen few-shot learner with the ability to perform CLIP-like vision-semantic alignment. Unlike the heavy CLIP vision encoder, the frozen few-shot vision encoder is lightweight and compatible with existing FSL methods, such as IER [Rizve *et al.*, 2021] and SMKD [Lin *et al.*, 2023], enabling training from scratch for simplicity.

As shown in Figure 1(a), we align the few-shot vision encoder with the CLIP vision encoder by learning a linear projector. To achieve this, we treat unsupervised vision-semantic alignment as a proxy task, inspired by [Li *et al.*,

2024], and use knowledge distillation to align the output distributions of both models. For a given FSL task with images $\{x_i\}^{N\times(K+Q)}$ and their class names $\{c_j\}^N$ from $\mathcal{C}_{base}$, the CLIP teacher model first processes the category names using a fixed prompt template (e.g., "a photo of a {CLASS}"), then passing images and category names through the image encoder f_I^t and text encoder f_T^t to obtain normalized teacher image features u_i^t and text features w_j^t. We then input the same images into the frozen few-shot vision encoder f_I^s to obtain the normalized student image features u_i^s. The learnable projector $f_\varphi(\cdot)$ is introduced to match the feature dimensions at minimal computational cost while ensuring alignment quality. The teacher and student image features, along with the generated teacher text features, are used to compute the output logits q_i^t and q_i^s for the teacher and student models, respectively. The knowledge distillation loss is formulated using Kullback-Leibler divergence:

$$\mathcal{L}_{kd}\left(q^t, q^s, \tau\right) = \tau^2 KL\left(\sigma\left(q^t/\tau\right), \sigma\left(q^s/\tau\right)\right). \quad (1)$$

where $\sigma(\cdot)$ is the softmax function and τ is the temperature.

Additionally, the cosine classifier f_Φ is trained to acquire transferable visual knowledge from the base dataset $\mathcal{D}_{base}$ using the pre-trained few-shot vision encoder. Let $W_{base} = \{\mathbf{w}^c\}_{c\in\mathcal{C}_{base}}$ denotes the weight vectors of classifier f_Φ. The classification loss, $\mathcal{L}_{ce}$, is computed with cross-entropy over the base classes $\mathcal{C}_{base}$:

$$\mathcal{L}_{ce} = -\log \frac{\exp\left(\cos(f_I^s(x_i), \mathbf{w}^c)\right)}{\sum_{c'=1}^{|\mathcal{C}_{base}|} \exp\left(\cos f_I^s(x_i), \mathbf{w}^{c'}\right)}. \quad (2)$$

This allows the learned classifier can provide more meaningful and task-relevant visual knowledge for subsequent knowledge transfer and fusion within our SYNTRANS. Finally, we combine the knowledge distillation and classification losses into a multi-task objective to ensure the projector f_φ and classifier f_Φ perform their respective tasks effectively within the same FSL task as $\mathcal{L}_{vis} = \mathcal{L}_{ce} + \mathcal{L}_{kd}$.

3.4 Synergistic Knowledge Mining

In leveraging the capabilities of large multimodal models, our proposed SynMine first utilizes the rich common-sense knowledge embedded in large language model (LLM) to generate detailed class descriptions. Additionally, SynMine takes advantage of the advanced image-text alignment capabilities of pre-trained visual-language model (VLM) to refine these descriptions into high-quality semantic descriptors, enhancing their relevance for FSL.

As illustrated in Figure 2, SynMine follows a multi-step process to enhance the comprehensive understanding of class characteristics. Inspired by the "chain-of-thought" prompting technique [Wei *et al.*, 2022], which improves LLM performance through intermediate reasoning, we guide LLM using this technique. Initially, a concise definition matching the visual content is provided to the LLM to eliminate ambiguity and help it generate class descriptions focused on visual features. The first prompt is as follows:

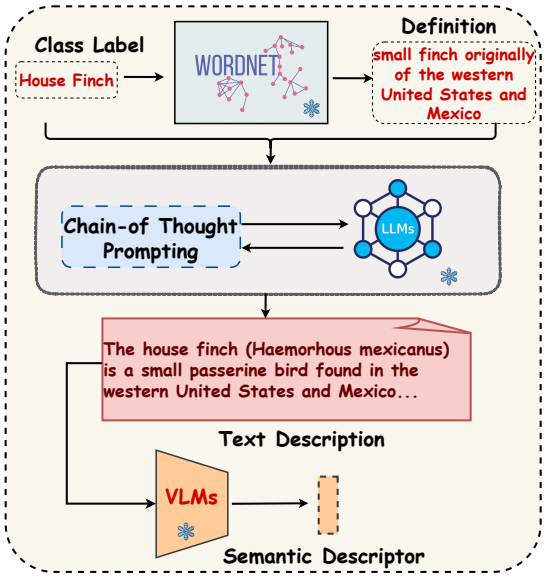


Figure 2: The pipeline of how the proposed SynMine module generates high-quality semantic descriptors.

> **Prompt 1:** [DEFINITION] is the definition of the [CLASS]. Can you describe the visual features associated with this category?

Here, [DEFINITION] is a brief class definition obtained from WordNet [Miller, 1995], and [CLASS] refers to the class name. Next, we refine these descriptions by focusing on distinctive visual attributes based on the initial responses. The second prompt is as follows:

> **Prompt 2:** Please describe the [CLASS] in a maximum of five sentences, focusing on discriminative visual features. Make the description more detailed and aligned with scientific facts, avoiding general summaries and subjective interpretations.

The generated descriptions are then processed by the text encoder of a pre-trained VLM, producing high-quality semantic descriptors. This approach is preferred over traditional word embedding models, such as Word2Vec [Mikolov *et al.*, 2013], as it captures deeper contextual meanings and nuances, enhancing the discrimination of visual features across classes.

3.5 Bidirectional Visual-Semantic Bridging

After distilling high-quality semantic descriptors through the SynMine module, we introduce the Visual-Semantic Bridging (VSBird) module, as shown in Figure 3. VSBird employs a dual autoencoder architecture to establish bidirectional mappings within and between the visual and semantic spaces, aiming to reconstruct multimodal classifier weights based on visual embeddings and semantic descriptors. This step is crucial for mapping visual and semantic knowledge into actionable classifier weights for FSL.

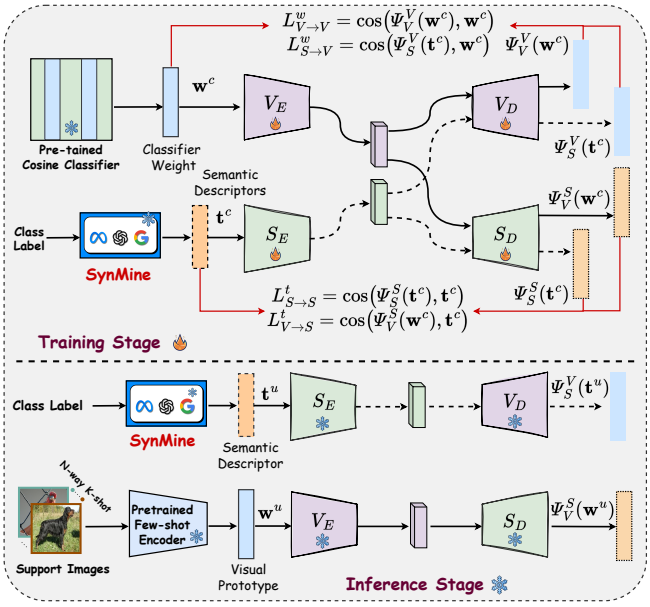


Figure 3: The pipeline of the proposed **V**isual-**S**emantic **Brid**ging (VSBird) module.

In this module, the classifier weights f_Φ are treated as visual embeddings encoding distinctive visual knowledge. We then use the corresponding semantic descriptors to facilitate the visual-semantic bridging. Specifically, let $W_{base} = \{\mathbf{w}^c\}_{c\in\mathcal{C}_{base}}$ denote the classifier weights and $T_{base} = \{\mathbf{t}^c\}_{c\in\mathcal{C}_{base}}$ the semantic descriptors. Our goal is to learn a mapping $\Psi_S^V : \mathcal{S} \to \mathcal{V}$, where $\mathcal{S}$ and $\mathcal{V}$ represent the semantic and visual spaces, respectively. The VSBird architecture consists of two encoder-decoder subnetworks: the visual encoder $V_E : \mathcal{V} \to \mathcal{Z}$, the semantic encoder $S_E : \mathcal{S} \to \mathcal{Z}$, the visual decoder $V_D : \mathcal{Z} \to \mathcal{V}$, and the semantic decoder $S_D : \mathcal{Z} \to \mathcal{S}$, with $\mathcal{Z}$ as the latent space. These components define the desired semantics-to-weights mapping as $\Psi_S^V(\mathbf{t}^c) = V_D(S_E(\mathbf{t}^c))$. To avoid bias towards base classes, we introduce self- and cross-reconstruction objectives. These ensure both modalities preserve their structures within their respective spaces while also aligning the latent spaces. The self-reconstruction objectives minimize the following terms:

$$\begin{aligned}\mathcal{L}^w_{V\to V} &= \cos\left(\Psi_V^V\left(\mathbf{w}^c\right), \mathbf{w}^c\right) = \cos\left(V_D\left(V_E(\mathbf{w}^c)\right), \mathbf{w}^c\right),\\ \mathcal{L}^t_{S\to S} &= \cos\left(\Psi_S^S\left(\mathbf{t}^c\right), \mathbf{t}^c\right) = \cos\left(S_D\left(S_E(\mathbf{t}^c)\right), \mathbf{t}^c\right).\end{aligned} \tag{3}$$

Here, $\cos(\cdot,\cdot)$ denotes the cosine distance function. While the two autoencoders preserve structure within their respective spaces, they do not ensure alignment between the two latent spaces. To address this, we introduce two additional cross-reconstruction objectives in a symmetric manner:

$$\begin{aligned}\mathcal{L}^w_{S\to V} &= \cos\left(\Psi_S^V\left(\mathbf{t}^c\right), \mathbf{w}^c\right) = \cos\left(V_D\left(S_E(\mathbf{t}^c)\right), \mathbf{w}^c\right),\\ \mathcal{L}^t_{V\to S} &= \cos\left(\Psi_V^S\left(\mathbf{w}^c\right), \mathbf{t}^c\right) = \cos\left(S_D\left(V_E(\mathbf{w}^c)\right), \mathbf{t}^c\right).\end{aligned} \tag{4}$$

The final objective of VSBird is a balanced combination of these reconstruction terms:

$$\mathcal{L}^w_t = \alpha*(\mathcal{L}^w_{V\to V}+\mathcal{L}^t_{S\to S})+(1-\alpha)*(\mathcal{L}^w_{S\to V}+\mathcal{L}^t_{V\to S}), \tag{5}$$

where α is a weight coefficient used to control the balance between self-reconstruction and cross-reconstruction.

During inference, for a novel class $u \in \mathcal{C}_{novel}$, the semantic encoder S_E and visual decoder V_D infer the semantic-derived classifier weight:

$$\mathbf{w}^u_{s'} = \Psi_S^V(\mathbf{t}^u) = V_D(S_E(\mathbf{t}^u)), \tag{6}$$

and for a unseen visual prototype $\mathbf{w}^u$, the visual encoder V_E and semantic decoder S_D infer the visual-derived classifier weight:

$$\mathbf{w}^u_{v'} = \Psi_V^S(\mathbf{w}^u) = S_D(V_E(\mathbf{w}^u)). \tag{7}$$

3.6 Multi-modal Knowledge Fusion

In the knowledge fusion stage, we create N-way K-shot meta-tasks from the base training set $\mathcal{D}_{base}$ to mimic the few-shot scenario during testing. The main goal is to develop a visual weight generator and a semantic weight reconstructor to combine visual-based and semantic-based classifier weights, forming robust multimodal classifiers for FSL tasks.

Given the pre-trained few-shot vision encoder f_I^s, we calculate the visual-based classifier weight $\mathbf{w}_v^m$ for each class m in the support set $\mathcal{S}$ as: $\mathbf{w}_v^m = \frac{1}{\left\|\sum_{i=1}^K f_I^s(x_i)\right\|_2}\sum_{i=1}^K f_I^s(x_i)$, where K is the number of samples per class m. Next, similar to Equations (6) and (7), we transfer visual and semantic knowledge from SynMine to generate visual-derived classifier weight $\mathbf{w}_{v'}^m$ and semantic-derived classifier weight $\mathbf{w}_{s'}^m$ for each class $m \in \mathcal{S}$. These weights complement each other: the visual-based weight $\mathbf{w}_v^m$ and semantic-derived classifier weight $\mathbf{w}_{s'}^m$ complement one another, and similarly, the semantic-based weight $\mathbf{w}_s^m$ and visual-derived weight $\mathbf{w}_{v'}^m$ complement each other.

To facilitate visual-semantic weight fusion, we introduce the visual weight generator G and semantic weight reconstructor R. The generator G, consisting of a fully connected layer followed by a sigmoid function, adaptively produces a weight coefficient as $\beta = \frac{1}{1+\exp\left(-G\left(\mathbf{w}_{s'}^m\right)\right)}$, with values restricted to the range $[0, 1]$. This coefficient β is used to balance the contributions of $\mathbf{w}_{s'}^m$ and $\mathbf{w}_v^m$ in the visual-dominated classifier as $\mathbf{w}_V^m = \beta \cdot \mathbf{w}_{s'}^m + (1-\beta)\cdot\mathbf{w}_v^m$. Similarly, the semantic-based weight $\mathbf{w}_s^m$ and visual-derived weight $\mathbf{w}_{v'}^m$ are concatenated and passed through the reconstructor R to form the semantic-dominated classifier as $\mathbf{w}_S^m = R\left(\mathbf{w}_s^m, \mathbf{w}_{v'}^m\right) = \sigma\left(\left[\mathbf{w}_s^m \cdot \mathbf{w}_{v'}^m\right]^\top W_1\right)W_2$, where σ is an activation function, and W_1 and W_2 are learnable weights. Afterward, we generate multimodal classifiers for the N-way K-shot meta-task: $\mathbf{W}_V = \{\mathbf{w}_V^m\}_{m=1}^N$ and $\mathbf{W}_S = \{\mathbf{w}_S^m\}_{m=1}^N$. Given an image q from the query set, we compute the probabilities P_v and P_s using cosine similarity as $P_v = \frac{\exp(\cos(f_I^s(q), \mathbf{w}_V^m))}{\sum_{m'=1}^N \exp\left(\cos(f_I^s(q), \mathbf{w}_V^{m'})\right)}$ and $P_s = \frac{\exp(\cos(f_\varphi(f_I^s(q)), \mathbf{w}_S^m))}{\sum_{m'=1}^N \exp\left(\cos(f_\varphi(f_I^s(q)), \mathbf{w}_S^{m'})\right)}$. We use the softmax cross-entropy loss function to train the generator G and reconstructor R.

	Method	Backbone	MiniImageNet		TieredImageNet	
			5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
Visual-Based	MatchNet [Vinyals *et al.*, 2016]	ResNet-12	65.64 ± 0.20	78.72 ± 0.15	68.50 ± 0.92	80.60 ± 0.71
	ProtoNet [Snell *et al.*, 2017]	ResNet-12	62.29 ± 0.33	79.46 ± 0.48	68.25 ± 0.23	84.01 ± 0.56
	MAML [Finn *et al.*, 2017]	ResNet-12	58.05 ± 0.21	58.05 ± 0.10	67.92 ± 0.17	72.41 ± 0.20
	MetaOptNet [Lee *et al.*, 2019]	ResNet-18	62.64 ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
	FEAT [Ye *et al.*, 2020]	ResNet-12	66.78 ± 0.20	82.05 ± 0.14	70.80 ± 0.23	84.79 ± 0.16
	Meta-Baseline [Chen *et al.*, 2021]	ResNet-12	63.17 ± 0.23	79.26 ± 0.17	68.62 ± 0.27	83.29 ± 0.18
	CVET [Yang *et al.*, 2022b]	ResNet-12	70.19 ± 0.46	84.66 ± 0.29	72.62 ± 0.51	86.62 ± 0.33
	FGFL [Cheng *et al.*, 2023]	ResNet-12	69.14 ± 0.80	86.01 ± 0.62	73.21 ± 0.88	87.21 ± 0.61
	SUN [Dong *et al.*, 2022]	ViT-S	67.80 ± 0.45	83.25 ± 0.30	72.99 ± 0.50	86.74 ± 0.33
	SMKD [Lin *et al.*, 2023]	ViT-S	74.28 ± 0.18	88.82 ± 0.09	78.83 ± 0.20	91.02 ± 0.12
	FewTURE [Hiller *et al.*, 2022]	Swin-T	72.40 ± 0.78	86.38 ± 0.49	76.32 ± 0.87	89.96 ± 0.55
Semantic-Based	KTN [Peng *et al.*, 2019]	ResNet-12	61.42 ± 0.72	70.19 ± 0.62	68.01 ± 0.73	79.06 ± 0.70
	AM3 [Xing *et al.*, 2019]	ResNet-12	65.30 ± 0.49	78.10 ± 0.36	69.08 ± 0.47	82.58 ± 0.31
	PC-FSL [Zhang *et al.*, 2021]	ResNet-12	69.68 ± 0.76	81.65 ± 0.54	74.19 ± 0.90	86.09 ± 0.60
	SEGA [Yang *et al.*, 2022a]	ResNet-12	69.04 ± 0.26	79.03 ± 0.18	72.18 ± 0.30	84.28 ± 0.21
	LPE-CLIP [Yang *et al.*, 2023]	ResNet-12	71.64 ± 0.40	79.67 ± 0.32	73.88 ± 0.48	84.88 ± 0.36
	KSTNET [Li *et al.*, 2023]	ResNet-12	71.51 ± 0.73	82.61 ± 0.48	75.52 ± 0.77	85.85 ± 0.59
	4S-FSL [Lu *et al.*, 2023]	ResNet-12	72.64 ± 0.70	84.73 ± 0.50	-	-
	SP-CLIP [Chen *et al.*, 2023]	Visformer-T	72.31 ± 0.40	83.42 ± 0.30	78.03 ± 0.46	88.55 ± 0.32
	SemFew [Zhang *et al.*, 2024]	Swin-T	78.94 ± 0.66	86.49 ± 0.50	82.37 ± 0.77	89.89 ± 0.52
	SYNTRANS	ResNet-12	76.20 ± 0.69	86.12 ± 0.54	79.69 ± 0.81	87.78 ± 0.60
	SYNTRANS	ViT-S	**81.30 ± 0.61**	**89.96 ± 0.42**	**84.31 ± 0.54**	**91.73 ± 0.44**

Table 1: Comparison with state-of-the-art methods on MiniImageNet and TieredImageNet.

Inference. During inference, the visual-dominated classifier and the semantic-dominated classifier are complementary to each other. Therefore, we propose a fusion mechanism to obtain the final class logits. Given a test image x_t, the prediction is made as follows:

$$y^* = \arg\max\left(\langle [f_I^s(x_t), f_\varphi(f_I^s(x_t))], [\mathbf{W}_V, \lambda \mathbf{W}_S] \rangle\right), \quad (8)$$

where λ is a positive balancing coefficient, empirically set to $\frac{1}{K}$ in our SYNTRANS.

4 Experiments

4.1 Datasets

Following the settings in [Zhang *et al.*, 2024], we evaluate the performance of the proposed SYNTRANS framework on four widely used benchmarks in FSL. Two of these datasets are derived from the ImageNet dataset [Russakovsky *et al.*, 2015]: MiniImageNet [Vinyals *et al.*, 2016] and TieredImageNet [Ren *et al.*, 2018]. Another two dataset are CIFAR-FS [Lee *et al.*, 2019] and FC100 [Oreshkin *et al.*, 2018].

4.2 Implementation Details

Architecture. In all experimental setups, we utilize ResNet-12 and ViT-Small (ViT-S) as the few-shot vision encoders. Specifically, the ResNet-12 encoder is pretrained using the training strategy described in IER [Rizve *et al.*, 2021], while the ViT-S encoder follows the strategy reported in SMKD [Lin *et al.*, 2023]. For the ResNet-12 encoder, visual features are obtained by averaging the outputs from the final residual block, resulting in a feature dimension of 640. For the ViT-S encoder, visual embeddings are computed by averaging the hidden states from the last transformer block, yielding a feature dimension of 384. During the visual knowledge distillation stage, we use both the vision and text encoders from Res50x4 CLIP [Radford *et al.*, 2021] as a strong teacher and employ two linear layers to construct the linear projector f_φ. In the SynMine module, we leverage the GPT-3.5-turbo model as the large language model and Res50x4 CLIP [Radford *et al.*, 2021] as the visual-language model. The SynMine module facilitates the alignment of features across vision and language modalities. The encoders and decoders of the VSBird module consist of single-layer linear mappings, with ReLU activation following the encoder mappings. A simple fully connected layer serves as the learnable weight generator G. The weight reconstructor R combines visual and textual features using two fully connected layers followed by a LeakyReLU activation function. The hidden layer has a dimension of 2048.

Method	Backbone	CIFAR-FS		FC100	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
ProtoNet [Snell *et al.*, 2017]	ResNet-12	72.20 ± 0.73	83.50 ± 0.50	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
RFS [Tian *et al.*, 2020]	ResNet-12	71.50 ± 0.80	86.90 ± 0.50	42.60 ± 0.70	59.10 ± 0.60
SUN [Dong *et al.*, 2022]	ViT-S	78.37 ± 0.46	88.84 ± 0.32	-	-
SMKD [Lin *et al.*, 2023]	ViT-S	80.08 ± 0.18	90.63 ± 0.13	50.38 ± 0.16	68.37 ± 0.16
FewTURE [Hiller *et al.*, 2022]	Swin-T	77.76 ± 0.81	88.90 ± 0.59	47.68 ± 0.78	63.81 ± 0.75
SEGA [Yang *et al.*, 2022a]	ResNet-12	78.45 ± 0.24	86.00 ± 0.20	-	-
LPE-CLIP [Yang *et al.*, 2022a]	ResNet-12	80.62 ± 0.41	86.22 ± 0.33	-	-
4S-FSL [Lu *et al.*, 2023]	ResNet-12	74.50 ± 0.84	88.76 ± 0.53	-	-
SP-CLIP [Chen *et al.*, 2023]	Visformer-T	82.18 ± 0.40	88.24 ± 0.32	48.53 ± 0.38	61.55 ± 0.41
SemFew [Zhang *et al.*, 2024]	Swin-T	84.34 ± 0.67	89.11 ± 0.54	54.27 ± 0.77	65.02 ± 0.72
SYNTRANS	ResNet-12	82.58 ± 0.75	89.42 ± 0.56	52.30 ± 0.75	64.91 ± 0.59
SYNTRANS	ViT-S	**84.64 ± 0.65**	**90.81± 0.41**	**56.38 ± 0.69**	**69.45± 0.54**

Table 2: Comparison with state-of-the-art methods on CIFAR-FS and FC100.

Training Details. During the visual knowledge distillation stage, we freeze both the few-shot vision encoder and CLIP's vision encoder, focusing on optimizing the linear projector and cosine classifier. For ResNet-12, we follow the methods in [Li *et al.*, 2023] and resize the input images to 84×84. For ViT-S, we resize the input image to 320×320 for MiniImageNet and TieredImageNet, and to 224×224 for CIFAR-FS and FC100, maintaining consistency with SMKD [Lin *et al.*, 2023]. In the knowledge transfer stage, we freeze all parameters in the SynMine module and only train the parameters of the VSBird module. In the meta-training stage, we train only the parameters of the weight generator and weight reconstructor. For both stages, we employ the Adam optimizer [Kingma and Ba, 2015] with an initial learning rate of 0.0001 and weight decay of 5×10^{-4}. In particular, the VSBird module is trained for 50 epochs with the hyperparameter α set to 0.7, while the weight generator and weight reconstructor are trained for 10 epochs.

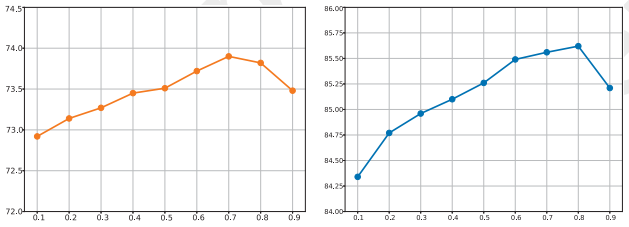
Evaluation protocol. The proposed method is evaluated under 5-way 1/5-shot settings on the novel dataset, with 600 few-shot tasks randomly sampled from it. Each task consists of 15 query samples per class. We report the average accuracy (%) with 95% confidence intervals.

4.3 Benchmark Comparisons and Evaluations

Tables 1 and 2 summarize the performance of recent state-of-the-art FSL methods on the MiniImageNet, TieredImageNet, CIFAR-FS, and FC100 datasets, focusing on the 5-way 1/5-shot tasks. The experimental results demonstrate that the SYNTRANS framework achieves outstanding performance across all datasets, particularly when visual information is limited. In the 5-way 1-shot scenario, SYNTRANS outperforms the most relevant semantic-based method, SemFew [Zhang *et al.*, 2024], by a margin of 2.98% due to its flexible knowledge transfer framework that mines rich knowledge from diverse large models. Notably, SYNTRANS shows greater improvement in the 1-shot setting compared to the 5-shot setting. In the 5-way 5-shot scenario, SYNTRANS still maintains an advantage over all state-of-the-art methods, though the improvements are less pronounced than in the 1-

Knowledge Source	Knowledge Encoder	ResNet-12	
		5-way 1-shot	5-way 5-shot
Class Names	Word2vec	72.66 ± 0.70	84.68 ± 0.50
Class Names	VLMs	72.87 ± 0.71	84.67 ± 0.50
Short Definitions (WordNet)	Word2vec	72.03 ± 0.73	84.18 ± 0.52
Short Definitions (WordNet)	VLM	73.52 ± 0.73	85.26 ± 0.51
Rich Descriptions (LLM)	Word2vec	74.41 ± 0.70	85.76 ± 0.51
Rich Descriptions (LLM)	VLM	**76.20 ± 0.69**	**86.12 ± 0.54**

Table 3: Ablation study about knowledge quality on MiniImageNet.



(a) 5-way 1-shot (b) 5-way 5-shot

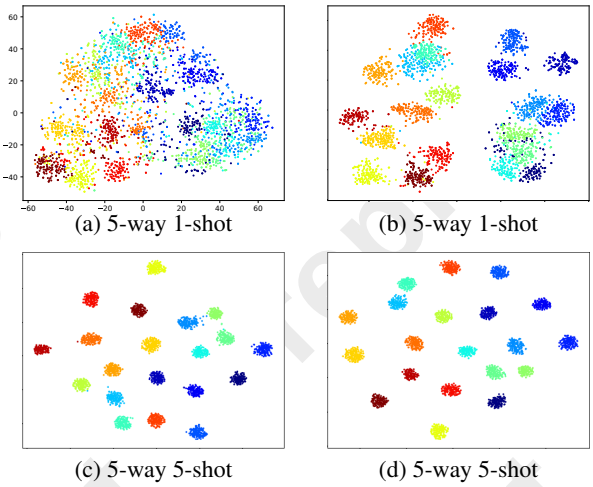
Figure 4: Influence of weight coefficient α on MiniImageNet.

shot setting. This suggests that even with more visual data, high-quality semantic knowledge from large models can still enhance performance. Overall, the results highlight the effectiveness of SYNTRANS in various scenarios, demonstrating its ability to leverage both semantic and visual knowledge.

4.4 Ablation Studies

Influence of Semantic Knowledge Quality. As shown in Table 3, we evaluate the impact of knowledge quality using different sources and encoders. First, we compare common semantics (class names) with generic descriptions from WordNet (*i.e.,* "Short Definitions") and richer descriptions from LLMs via SynMine (*i.e.,* "Rich Descriptions"). Results show that LLM-generated descriptions yield the best performance. This suggests that rich descriptions provide deeper semantic understanding, adding nuanced attributes to the classifier weights that simple class names cannot. Next, we compare Word2Vec models with VLMs for encoding semantic knowledge. VLMs outperform Word2Vec in both 1-shot and 5-shot settings, benefiting from their multi-modal training, which enhances semantic understanding. The strong performance of VLMs, particularly when paired with LLM-generated descriptions, highlights the effectiveness of combining large models for optimal knowledge transfer.

Influence of Hyper-parameter α. The VSBird module is essential for bridging visual and semantic spaces, where the dual autoencoder architecture encourages both self-reconstruction within each space and cross-reconstruction between spaces. Figure 4 shows the performance of the visual-dominated classifier $\mathbf{W}_V$ on MiniImageNet with varying weight coefficients α in Equation (5). A larger α enhances the significance of self-reconstruction loss during VSKB module training. When α is too small, self-reconstruction loss is suppressed by cross-reconstruction loss, leading to lower accuracy. This suggests that while cross-reconstruction is crucial for aligning latent spaces, preserving the individual structures of visual and semantic spaces is equally important. However, beyond $\alpha = 0.7$ for the 1-shot setting and $\alpha = 0.8$ for the 5-shot setting, accuracy slightly decreases, indicating



(a) 5-way 1-shot (b) 5-way 1-shot

(c) 5-way 5-shot (d) 5-way 5-shot

Figure 5: t-SNE visualization of the classification weights for all novel categories in Mini-ImageNet. (a) 1-shot visual-based classifier. (b) 1-shot multi-modal based classifier. (c) 5-shot visual-based classifier. (d) 5-shot multi-modal based classifier.

that excessive self-reconstruction diminishes the benefits of cross-reconstruction. Thus, balancing self-reconstruction and cross-reconstruction is crucial for optimal performance, with $\alpha = 0.7$ achieving the best trade-off across settings.

Effect of Multi-modal Knowledge Fusion. As shown in Figure 5, we utilize t-SNE visualization to present classifier weights for all novel categories of MiniImageNet. Figure 5(a) illustrates classifier weights derived solely from visual data in the 1-shot setting, revealing loosely defined clusters with significant category overlap. Conversely, Figure 5(b) shows the 1-shot results with fused visual and semantic knowledge, where clusters are more compact and distinct. This highlights the substantial benefit of semantic knowledge in scenarios with limited samples. Figure 5(c) shows the 5-shot classifier weights based solely on visual data, showing more tightly grouped clusters with less overlap. Figure 5(d) presents the 5-shot results with multi-modal knowledge fusion, where clusters are even more distinct and compact. This demonstrates that rich knowledge still improves FSL performance even with a higher number of samples.

5 Conclusion

In this paper, we delve into the previously unexplored potential of harnessing the extensive knowledge available in large multimodal models to empower the pre-trained few-shot learner. As a result, we propose a Synergistic Knowledge Transfer (SYNTRANS) framework that effectively transfers diverse and complementary knowledge from both visual- and large-language models to address FSL tasks. The essence of SYNTRANS lies in its proficient capability to distill and transform explict visual knowledge and implicit semantic knowledge from these large models into practical classifier weights, thereby significantly improving FSL performance through a multimodal knowledge fusion manner. Experimental results on four benchmark datasets demonstrate the superior efficacy of SYNTRANS compared to the state-of-the-art methods.

Acknowledgments

This work is supported by the Shenzhen-Hong Kong-Macao Science and Technology Plan Project (Category C) under the Shenzhen Municipal Science and Technology Innovation Commission (Project No. SGDX20230821092359002), and a grant for Collaborative Research with World-leading Research Groups of The Hong Kong Polytechnic University (Project No. G-SACF). Additional support is provided by the Guangdong Natural Science Funds for Distinguished Young Scholars (Grant No. 2023B1515020097), the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-GV-2023-011), and the Lee Kong Chian Fellowships.

References

[Chen *et al.*, 2019] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019.

[Chen *et al.*, 2021] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *ICCV*, pages 9062–9071, 2021.

[Chen *et al.*, 2023] Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, and Tieniu Tan. Semantic prompt for few-shot image recognition. In *CVPR*, pages 23581–23591, 2023.

[Cheng *et al.*, 2023] Hao Cheng, Siyuan Yang, Joey Tianyi Zhou, Lanqing Guo, and Bihan Wen. Frequency guidance matters in few-shot learning. In *ICCV*, pages 11814–11824, 2023.

[Dong *et al.*, 2022] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Self-promoted supervision for few-shot transformer. In *ECCV*, pages 329–347, 2022.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017.

[Fu *et al.*, 2023] Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *CVPR*, pages 24575–24584, 2023.

[Fu *et al.*, 2024] Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In *ECCV*, pages 247–264, 2024.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hiller *et al.*, 2022] Markus Hiller, Rongkai Ma, Mehrtash Harandi, and Tom Drummond. Rethinking generalization in few-shot classification. In *NeurIPS*, 2022.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *ICLR*, 2015.

[Lampert *et al.*, 2009] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[Lee *et al.*, 2019] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, pages 10657–10665, 2019.

[Li *et al.*, 2019] Zechao Li, Jinhui Tang, and Tao Mei. Deep collaborative embedding for social image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2070–2083, 2019.

[Li *et al.*, 2023] Zechao Li, Hao Tang, Zhimao Peng, Guo-Jun Qi, and Jinhui Tang. Knowledge-guided semantic transfer network for few-shot image recognition. *IEEE Trans. Neural Networks Learn. Syst.*, 2023.

[Li *et al.*, 2024] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *CVPR*, pages 26607–26616, 2024.

[Lin *et al.*, 2023] Han Lin, Guangxing Han, Jiawei Ma, Shiyuan Huang, Xudong Lin, and Shih-Fu Chang. Supervised masked knowledge distillation for few-shot transformers. In *CVPR*, pages 19649–19659. IEEE, 2023.

[Lu *et al.*, 2023] Jinda Lu, Shuo Wang, Xinyu Zhang, Yanbin Hao, and Xiangnan He. Semantic-based selection, synthesis, and supervision for few-shot learning. In *ACM MM*, pages 3569–3578, 2023.

[Luo *et al.*, 2018] Chunjie Luo, Jianfeng Zhan, Xiaohe Xue, Lei Wang, Rui Ren, and Qiang Yang. Cosine normalization: Using cosine similarity instead of dot product in neural networks. In *ICANN*, pages 382–391, 2018.

[Mikolov *et al.*, 2013] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[Miller, 1995] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[Neumann *et al.*, 2001] Heiko Neumann, Luiz Pessoa, and Thorsten Hansen. Visual filling-in for computing perceptual surface properties. *Biol. Cybern.*, 85(5):355–369, 2001.

[Oreshkin *et al.*, 2018] Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, pages 721–731, 2018.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, pages 27730–27744, 2022.

[Peng *et al.*, 2019] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image

recognition with knowledge transfer. In *ICCV*, pages 441–449, 2019.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[Ralph *et al.*, 2017] Matthew A Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T Rogers. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42–55, 2017.

[Ravi and Larochelle, 2016] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2016.

[Ren *et al.*, 2018] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018.

[Rizve *et al.*, 2021] Mamshad Nayeem Rizve, Salman H. Khan, Fahad Shahbaz Khan, and Mubarak Shah. Exploring complementary strengths of invariant and equivariant representations for few-shot learning. In *CVPR*, pages 10836–10846, 2021.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087, 2017.

[Sung *et al.*, 2018] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208, 2018.

[Tang *et al.*, 2020] Hao Tang, Zechao Li, Zhimao Peng, and Jinhui Tang. Blockmix: Meta regularization and self-calibrated inference for metric-based meta-learning. In *ACM MM*, pages 610–618, 2020.

[Tang *et al.*, 2022] Hao Tang, Chengcheng Yuan, Zechao Li, and Jinhui Tang. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognit.*, 130:108792, 2022.

[Tang *et al.*, 2023] Hao Tang, Jun Liu, Shuanglin Yan, Rui Yan, Zechao Li, and Jinhui Tang. M3net: Multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In *ACM Multimedia*, pages 1719–1728, 2023.

[Tian *et al.*, 2020] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *ECCV*, pages 266–282, 2020.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NeurIPS*, pages 3630–3638, 2016.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, pages 24824–24837, 2022.

[Wu *et al.*, 2022] Zongqian Wu, Peng Zhou, Guoqiu Wen, Yingying Wan, Junbo Ma, Debo Cheng, and Xiaofeng Zhu. Information augmentation for few-shot node classification. In *IJCAI*, pages 3601–3607, 2022.

[Wu *et al.*, 2024] Zongqian Wu, Yujie Mo, Peng Zhou, Shangbo Yuan, and Xiaofeng Zhu. Self-training based few-shot node classification by knowledge distillation. In *AAAI*, volume 38, pages 15988–15995, 2024.

[Xian *et al.*, 2019] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. F-VAEGAN-D2: A feature generating framework for any-shot learning. In *CVPR*, pages 10275–10284, 2019.

[Xing *et al.*, 2019] Chen Xing, Negar Rostamzadeh, Boris N. Oreshkin, and Pedro O. Pinheiro. Adaptive cross-modal few-shot learning. In *NeurIPS*, pages 4847–4857, 2019.

[Yang *et al.*, 2022a] Fengyuan Yang, Ruiping Wang, and Xilin Chen. SEGA: semantic guided attention on visual prototype for few-shot learning. In *WACV*, pages 1586–1596, 2022.

[Yang *et al.*, 2022b] Zhanyuan Yang, Jinghua Wang, and Yingying Zhu. Few-shot classification with contrastive learning. In *ECCV*, pages 293–309, 2022.

[Yang *et al.*, 2023] Fengyuan Yang, Ruiping Wang, and Xilin Chen. Semantic guided latent parts embedding for few-shot learning. In *WACV*, pages 5436–5446, 2023.

[Ye *et al.*, 2020] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *CVPR*, pages 8808–8817, 2020.

[Zha *et al.*, 2023] Zican Zha, Hao Tang, Yunlian Sun, and Jinhui Tang. Boosting few-shot fine-grained recognition with background suppression and foreground alignment. *IEEE Trans. Circuits Syst. Video Technol.*, 33(8):3947–3961, 2023.

[Zhang *et al.*, 2021] Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. Prototype completion with primitive knowledge for few-shot learning. In *CVPR*, pages 3754–3762, 2021.

[Zhang *et al.*, 2024] Hai Zhang, Junzhe Xu, Shanlin Jiang, and Zhenan He. Simple semantic-aided few-shot learning. In *CVPR*, pages 28588–28597, 2024.