# Multi-directional Knowledge Transfer for Few-Shot Learning

Shuo Wang
University of Science and Technology of China, Hefei, China
shuowangcv@ustc.edu.cn

Xinyu Zhang
University of Science and Technology of China, Hefei, China
zhangxy21@mail.ustc.edu.cn

Yanbin Hao
University of Science and Technology of China, Hefei, China
haoyanbin@hotmail.com

Chengbing Wang
University of Science and Technology of China, Hefei, China
wwq197297@mail.ustc.edu.cn

Xiangnan He*
University of Science and Technology of China, Hefei, China
xiangnanhe@gmail.com

## ABSTRACT

Knowledge transfer-based few-shot learning (FSL) aims at improving the recognition ability of a novel object under limited training samples by transferring relevant potential knowledge from other data. Most related methods calculate such knowledge to refine the representation of a novel sample or enrich the supervision to a classifier during a transfer procedure. However, it is easy to introduce new noise during the transfer calculations since: (1) the unbalanced quantity of samples between the known (base) and the novel categories biases the contents capturing of the novel objects, and (2) the semantic gaps existing in different modalities weakens the knowledge interaction during the training.

To reduce the influences of these issues in knowledge transfer-based FSL, this paper proposes a multi-directional knowledge transfer (MDKT). Specifically, (1) we use two independent unidirectional knowledge self-transfer strategies to calibrate the distributions of the novel categories from base categories in the visual and the textual space. It aims to yield transferable knowledge of the base categories to describe a novel category. (2) To reduce the inferences of semantic gaps, we first use a bidirectional knowledge connection to exchange the knowledge between the visual and the textual space. Then we adopt an online fusion strategy to enhance the expressions of the textual knowledge and improve the prediction accuracy of the novel categories by combining the knowledge from different modalities. Empirical studies on three FSL benchmark datasets demonstrate the effectiveness of MDKT, which improves the recognition accuracy on novel categories under limited samples, especially on 1-shot and 2-shot training tasks.

## CCS CONCEPTS

• **Computing methodologies** → *Transfer learning*;

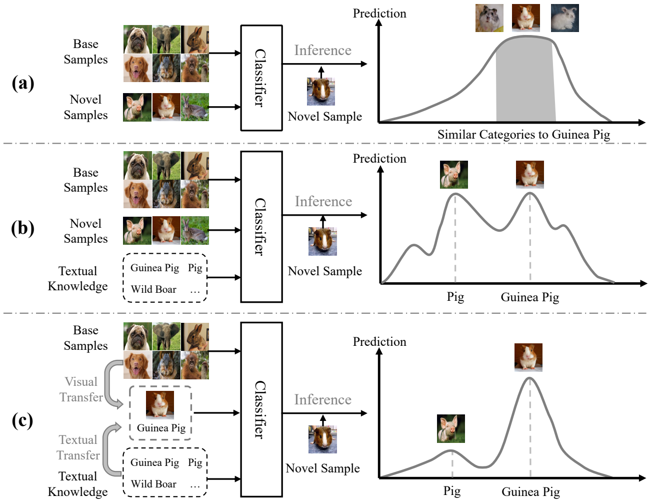


**Figure 1: The predictions of (a) classical classification method, (b) common knowledge transfer-based method, and (c) multi-directional knowledge transfer method, where (c) benefits from both (a) and (b), and achieves a more accurate prediction on a novel sample.**

## KEYWORDS

Few-Shot Learning, Knowledge Transfer

**ACM Reference Format:**
Shuo Wang, Xinyu Zhang, Yanbin Hao, Chengbing Wang, and Xiangnan He. 2022. Multi-directional Knowledge Transfer for Few-Shot Learning. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547837

## 1 INTRODUCTION

In the past few years, convolutional neural networks (CNNs) have been proved the powerful ability on several visual tasks, such as classification [14, 15, 17, 20, 40], and information transformation [4, 6, 12, 13, 41, 50, 51]. Although CNNs have strong robustness to the content of objects, they can hardly show a good performance without large amounts of training data. Conversely, humans can recognize a new category with a few samples of it because they have seen many other related objects or learned them from other semantic knowledge, and thus are already familiar with their salient

[*] Corresponding author is Xiangnan He, who is also affiliated with the Institute of Dataspace, Hefei Comprehensive National Science Center.

*MM '22, October 10–14, 2022, Lisboa, Portugal*

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547837

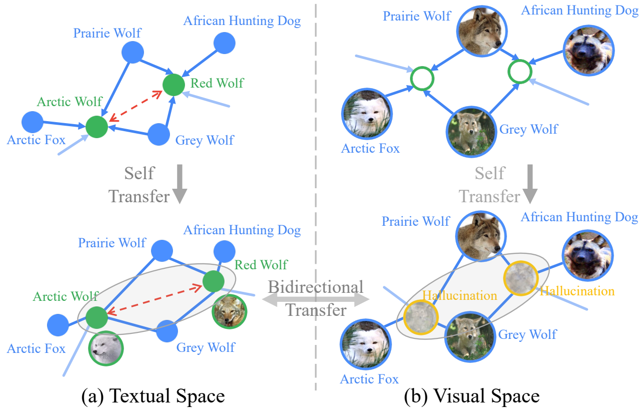


**Figure 2: A procedure of our multi-directional knowledge transfer (MDKT). (a) Given a series of textual features of novel categories (points in green) and its related base categories (blue points), the self-transfer in textual space adjusts the distance (dotted line in red) of the novel categories by relations between the features. (b) Given a series of samples of the base categories (circles in blue), the self-transfer in visual space hallucinates the potential prototypes of related novel samples (circles in yellow) by analyzing the relations of different categories. Finally, a bidirectional transfer strategy connects the different modalities, which aims to close the novel samples (in green circles) with their textual labels.**

features. Therefore, knowledge transfer-based few-shot learning (FSL) has been proposed recently to imitate this human ability.

Most recent knowledge transfer-based methods [16, 23, 42, 43] use knowledge to intervene in training procedures of the representation learning or classifier optimization stage. Specifically, [16] and [43] use a CNN trained on the base categories to extract the global features of novel objects directly. They aim to transfer the textures from the base categories to help describe a novel category. However, this operation is insufficient to represent the novel samples since the number of samples of base categories is much larger than that of novel categories. As shown in Figure 1(a), a CNN trained on the base categories is more inclined to focus on the textures and structures of the objects it learns and ignores the details of a novel sample [42]. [23] extracts the knowledge from semantics and uses it as extra supervision for transfer training. Such extra supervision alleviates the recognition bias of the classifier trained only under the supervision of hard labels. However, it ignores the semantic gaps in different modalities and introduces task-independent noise from the external knowledge, which may mislead the recognition of the classifier. As depicted in Figure 1(b), although the introduced semantic reduces the predicted probability of categories with similar textures, it increases the predictions of categories with similar semantics during the inference. Based on the analysis above, we set our goal in two aspects: reducing the influences of unbalanced training data and connecting the knowledge between the different modalities. Thus, as shown in Figure 1(c), to suppress predictions for irrelevant novel categories, we propose a multi-directional knowledge transfer (MDKT) model which fuses the different transfer procedures: unidirectional knowledge transfer in a single modality and bidirectional knowledge transfer between the different modalities.

For reducing the influences of unbalanced training data, inspired by the humans' ability that describes a new category by many other familiar related knowledge (textual and visual space), we employ two unidirectional knowledge self-transfer strategies in the visual and the textual space independently to refine the descriptions of the novel categories from the base categories, called intra-modality calculation. In textual space, we focus on the relations between the base and the novel features and use them to adjust the distance in the novel categories. These relations help the semantic knowledge of novel categories can be discriminative with others and such discriminative features are used to stabilize the transfer procedure by fusing visual knowledge. As shown in Figure 2(a), given a series of samples (from both base and novel categories) and their descriptions, we first compare the distance between the base and the novel features in textual space, and then adjust the relations of novel features. For example, the refined features of the novel categories "Arctic Wolf" and "Red Wolf" are farther away than that in the original textual space, since the base categories "Arctic Fox" and "African Hunting Dog" help distinguish these novel categories. In visual space, we combine the relations calculated by the textual knowledge to hallucinate the potential prototypes of the novel samples and use these prototypes to help train the classifier. As shown in Figure 2(b), the novel samples of the "Arctic Wolf" and "Red Wolf" categories are related to the base categories by referring to the textual relations. Thus, the potential prototypes are similar to that in textual space. They provide prior knowledge for the classifier.

For connecting the knowledge between the different modalities, deep mutual learning (DML) provides a training strategy that rather than a one-way transfer between a static pre-defined model and a dynamic model, an ensemble of models learns collaboratively and teaches each other throughout the training process [53]. Inspired by DML, we design a bidirectional transfer strategy that combines the refined features in the textual space and the hallucinated prototypes in the visual space to exchange knowledge from different modalities, called inter-modality calculation. This strategy minimizes the consistency between different modalities in both the base and the novel categories to improve the capacity of refined features in unidirectional knowledge transfer procedures. As shown in the second row of Figure 2, similar to the recognition procedure of humans that they can correlate the knowledge from different modalities to comprehend a novel object, the connections between the refined samples are used to explore the relations between the textual features and the visual features of novel categories. For example, the visual sample of "Arctic Wolf" is related to its textual label. It provides extra supervision for the training and reduces the difficulty of optimization. In addition, we employ an online fusion strategy to capture the knowledge from different modalities calculations and improve the prediction accuracy of the novel category. The main contributions of our method are twofold.

(1) We introduce two unidirectional knowledge self-transfer strategies independently in the visual and the textual space to discover the potential knowledge of the novel categories.

(2) We design a bidirectional knowledge connection scheme to exchange knowledge in different modalities. It helps the classifier focus on the relations between the visual and the textual space. Meanwhile, we employ an online fusion strategy to better the relations between the different categories.

## 2 RELATED WORK

In this section, we will briefly introduce the traditional FSL methods and the related knowledge transfer-based FSL methods, and then enumerate the differences between ours and the related methods.

### 2.1 Traditional Few-Shot Learning

Previous methods generally design a classifier with different structures or optimization strategies to predict novel categories. The representative methods are meta-learning, metric-learning, nearest neighbor (NN), and so on. Specifically, the methods [7, 21, 29] based on the meta-learning train a meta-learner from many FSL tasks (with base categories) without relying on ad hoc knowledge to suit for new FSL tasks (with novel categories). Metric-learning [34, 35, 39] attempts to train a network that can make samples of the same category closer and samples of different categories farther in the feature space. NN-based methods [22, 23] use the learned features to search the novel labels in a given support set that is closest (or most similar) to a given feature.

### 2.2 Transfer-based Few-Shot Learning

Recently, the knowledge transfer-based FSL methods have been noticed by researchers. It is because the other knowledge can be introduced to enrich the supervision or enlarge the training information for the classifier during the transferring stage. To analyze the existing related work, we briefly group knowledge transfer-based methods into three categories by different transferring directions: vision-based, semantic-based, and fusion-based transfer.

*2.2.1 Vision-based Transfer.*
The vision transfer-based methods [10, 11, 16, 43, 49] attempt to find the relations between the novel and the base categories in visual space. For example, Gidaris *et al.* design a sample classification weight generator with an attention mechanism and modify the weights of the classifier with the cosine similarity [10]. The work in [11] combines meta-learning with a graph neural network (GNN) to model the relationships of different categories and predicts the parameters of novel classes. Yang *et al.* first “calibrate” the data distribution of the novel categories by calculating the statistical distribution of the base categories. Then they sample the novel features from such “corrected” distribution of novel categories to enrich training samples [49]. There are also many data augmentation methods existing in the FSL task. Such as Hariharan *et al.* hallucinate new samples in the feature space by using a separate Multilayer Perceptron (MLP) to model the relationships between the foregrounds and the backgrounds of images [16]. Wang *et al.* train a meta-learner with hallucination to expand the training set and to classify the samples simultaneously [43]. Similar to the feature hallucination methods, many traditional generation networks, *e.g.*, Generative Adversarial Networks (GAN) [9, 24, 55, 56], Autoencoder (AE) [32], and Variational AE (VAE) [28], have been applied to generate the original image for training. It also improves the performance of the classification.

*2.2.2 Semantic-based Transfer.*
The semantic transfer-based methods [22, 23, 42] aim at using the semantic knowledge from other modalities to refine the representations of the visual samples or enrich the supervision for the classifier. Such as the work in [23] clusters hierarchical textual labels both from the base and the novel categories to train a feature extractor. It helps learn a more transferable feature embedding for recognizing the novel samples. Wang *et at.* introduce semantic soft labels generated from textual knowledge to help the network learn a more powerful classifier [42]. These soft labels are used as supervision to help the classifier find hyperplane in classification space easily. Li *et at.* develop a class-relevant additive margin loss with the semantic similarity between each pair of classes to separate samples in the feature embedding space from similar classes [22].

*2.2.3 Fusion-based Transfer.*
The fusion transfer-based methods [1, 26, 47] capture the knowledge from different modalities and balance its influences to improve the capacity of the classifier. For example, Xing *et al.* first use two prototype networks to model the visual and the textual features and achieve the prototypes of different modalities. Then, a multi-modal fusion strategy with a self-attention mechanism is designed to calibrate the prototypes in the visual space [47]. Peng *et at.* propose a knowledge transfer network, named KTN, to explore the prior knowledge from the semantic knowledge and develop a semantic-visual mapping network to infer the category of novel sample [26]. The work in [1] designs a graph convolutional transfer network to introduce similar visual concepts captured by semantic correlations. Then it associates the classifier weights with graphs construction to update the parameters iteratively. Similar to the vision transfer-based methods, there are also existing many data augmentation methods [31, 45, 46] which fuse the semantic as prior knowledge to generate samples, such as conditional-GAN [45], conditional-VAE [31], and so on. They aim to achieve more samples for the training.

Based on the analysis of the related work, our method belongs to the fusion transfer-based method. The method most related to ours is recently proposed KTN in [26]. The key differences between ours are fourfold. First of all, in addition to the inter-modality transferring in the KTN, we propose the intra-modality unidirectional knowledge transfer. Second, we add interaction to novel categories between the visual and the textual space during the inter-modality bidirectional transferring. Third, we use the correlations of the available labels to construct an adjacency matrix rather than introduce a large-scale WordNet graph [25]. Finally, we replace the offline fusion strategy with an online fusion strategy to improve the recognition performance flexibly.

## 3 APPROACH

In this section, we elaborate on our proposed multi-directional knowledge transfer (MDKT). Firstly, we briefly revisit the preliminaries of the few-shot learning tasks and overview of our framework. Secondly, we illustrate our unidirectional knowledge transfer and bidirectional knowledge connection in detail. Finally, we describe the training and inference procedures of our method.

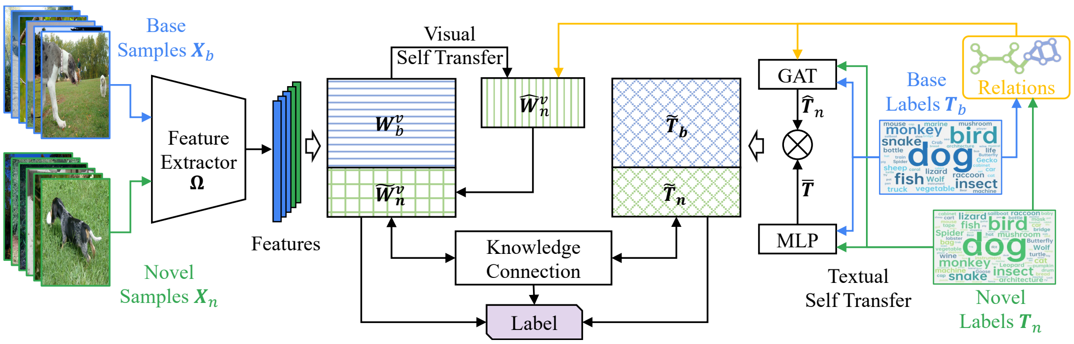


**Figure 3: An overview of our multi-directional knowledge transfer, where $W_*$, $\hat{W}_*$, and $\tilde{W}_*$ is the original weights, potential weights, and transferred weights, respectively, of the classifier in the visual space. And $\bar{T}_*$, $\hat{T}_*$, and $\tilde{T}_*$ is the features, potential features, and transferred features, respectively, of the semantic knowledge in the textual space.**

## 3.1 Preliminaries

The data of few-shot learning tasks can be split into three parts: training set $\mathcal{D}_{\text{train}}$, support set $\mathcal{D}_{\text{support}}$, and testing set $\mathcal{D}_{\text{test}}$. Specifically, $\mathcal{D}_{\text{train}}$ has large-scale training samples (*e.g.*, about hundreds of samples in one category), and the categories of these samples are denoted as $C_{\text{base}}$. It provides a large amount of prior knowledge as known contents to help describe other samples. Conversely, support set $\mathcal{D}_{\text{support}}$ and testing set $\mathcal{D}_{\text{test}}$ have the same category, called $C_{\text{novel}}$, which are disjoint with that in the training set $C_{\text{base}}$. The goal of few-shot learning is to learn an image classification model by using the training set and the support set that can accurately classify images in the testing set from novel categories, where the training samples of novel categories are sampled from $\mathcal{D}_{\text{support}}$ and the testing samples belong to $\mathcal{D}_{\text{test}}$. It usually focuses on the $N$-way-$K$-shot recognition problem that identifies $N$ novel categories and each category has $K$ support samples.

The overview of our multi-directional knowledge transfer is depicted in Figure 3. Before the transfer stage, we represent the visual samples ($\boldsymbol{X}_b$, $\boldsymbol{X}_n$) and their semantic labels ($\boldsymbol{T}_b$, $\boldsymbol{T}_n$) from the base and the novel categories into the features by a visual CNN and a word embedding method, respectively. Denoted the textual features of the base and the novel categories as $\boldsymbol{T}_b = \{\boldsymbol{t}_j \in \mathbb{R}^{d_t}\}_{j=1}^{|C_{\text{base}}|}$ and $\boldsymbol{T}_n = \{\boldsymbol{t}'_i \in \mathbb{R}^{d_t}\}_{i=1}^{|C_{\text{novel}}|}$, respectively, where $d_t$ is the dimension of the textual feature. Then, we employ two different format unidirectional knowledge self-transfer strategies both in the visual and the textual space to learn the potential novel knowledge from the base categories at first. Second, we connect such potential knowledge as bidirectional transfer to exchange information between the different modalities. Finally, the hard labels are used to optimize the whole parameters, which contain the weights of the classifier and the parameters of the transfer network. For the inference stage, we fuse the predictions from different modalities to improve the recognition accuracy of the novel samples. We introduce the details of our method as follows.

## 3.2 Unidirectional Knowledge Self-transfer

### 3.2.1 *Textual Self-Transfer.*

To transfer the knowledge from the base categories to the novel categories, we combine the relations from the textual space with a graph attention network (GAT) [38]. Compared with the traditional graph neural network (GNN) [30], GAT assigns the weights of each node by capturing the characteristics of its neighbors rather than calculating an entire graph. Thus, each node is only related to adjacent nodes and the shared edges. It is suitable for a unidirectional transfer calculation. To achieve this target, we first calculates the correlations between these features by a variant of Euclidean distance function. The correlation of $i^{\text{th}}$ and $j^{\text{th}}$ textual features can be defined as: $d(\boldsymbol{t}_i, \boldsymbol{t}_j) = ||\boldsymbol{t}_i - \boldsymbol{t}_j||_2^{-1}$. Then, we construct the adjacency matrix $\boldsymbol{A}$ of GAT by exploring the category correlation in two stages: (1) Given a textual feature of $k$ novel category $\boldsymbol{t}'_k$, we select the top $M$ base categories with the closest distance (largest similarity) to $\boldsymbol{t}'_k$ as $\mathcal{M}^k_{\text{base}}$. (2) We fuse the correlations of $\boldsymbol{t}_k$ and its related base categories $\mathcal{M}^k_{\text{base}}$ to fill the elements of the adjacency matrix. The $k^{\text{th}}$ row and $m^{\text{th}}$ column of $\boldsymbol{A}$ can be calculated as:

$$a_{k,m} = \frac{d(\boldsymbol{t}'_k, \boldsymbol{t}_m)}{\sum_{i\in\mathcal{M}^k_{\text{base}}}(d(\boldsymbol{t}'_k, \boldsymbol{t}_i))}, \tag{1}$$

where $a_{k,m}$ represents the adjacency relations between the $k^{\text{th}}$ novel category and $m^{\text{th}}$ base category. Thus $\boldsymbol{A} \in \mathbb{R}^{|C_{\text{novel}}|\times|C_{\text{base}}|}$ is an asymmetric matrix which focuses on the transfer knowledge from base categories to novel categories.

Then, we introduce the calculations of GAT. A GAT with $H$-head attentions first mapping the features into different $H$ hidden spaces, and then use the attention mechanism to measure the importance between the adjacency nodes of $h^{\text{th}}$ space:

$$\sigma^h_{k,m} = \frac{\exp(\text{LeakyRelu}([\boldsymbol{t}'_k\boldsymbol{W}^h||\boldsymbol{t}_m\boldsymbol{W}^h]\boldsymbol{W}^t_h))}{\sum_{\boldsymbol{t}_i\in\mathcal{M}^k_{\text{base}}}\exp(\text{LeakyRelu}([\boldsymbol{t}'_k\boldsymbol{W}^h||\boldsymbol{t}_i\boldsymbol{W}^h]\boldsymbol{W}^t_h))}, \tag{2}$$

where $\boldsymbol{W}^h \in \mathbb{R}^{d_t\times d'_t}$ is the parameter of $h^{\text{th}}$-head mapping calculation, $d'_t$ is the size of the transfer space, $\boldsymbol{W}^t_h \in \mathbb{R}^{2d'_t\times 1}$ is the parameter of $h^{\text{th}}$-head attention, and $[\cdot||\cdot]$ is concatenation operation, $\sigma^h_{k,m}$ is the element of $\boldsymbol{\sigma}^h \in \mathbb{R}^{|C_{\text{novel}}|\times|C_{\text{base}}|}$ indicates the importance of $m^{\text{th}}$ feature to $k^{\text{th}}$ feature. Combining with the adjacency matrix $\boldsymbol{A}$, the $h^{\text{th}}$ transferred textual features $\hat{\boldsymbol{T}}^h_n \in \mathbb{R}^{|C_{\text{novel}}|\times d'_t}$ of novel categories can be calculated by:

$$\hat{\boldsymbol{T}}^h_n = (\boldsymbol{A}\odot\boldsymbol{\sigma}^h)\boldsymbol{T}^h_b\boldsymbol{W}^h, \tag{3}$$

where ⊙ is hadamard product. Finally, we average the multi-head outputs to refine the representations of novel categories:

$$\hat{T}_n = \frac{1}{H}\sum \hat{T}_n^h = \frac{1}{H}\sum (\boldsymbol{A} \odot \boldsymbol{\sigma}^{\boldsymbol{h}}) T_b^h \boldsymbol{W}^h, \tag{4}$$

Meanwhile, we use a multi-layer perceptron (MLP) $\Phi_\theta$ to model the channel representations of the features independently:

$$\bar{T} = \Phi_\theta([T_b||T_n]) = \delta(([T_b||T_n])\boldsymbol{W}_\theta^t + \boldsymbol{b}_\theta^t), \tag{5}$$

where $\boldsymbol{W}_\theta^t \in \mathbb{R}^{d_t \times d_t'}$ and $\boldsymbol{b}_\theta^t \in \mathbb{R}^{d_t'}$ are the parameters of $\Phi_\theta$.

Finally, we first concatenate the relations and the contents of the textural features to further fuse two aspects knowledge by using a one-dimensional convolution:

$$\tilde{T} = \text{Conv1D}([\bar{T}||\hat{T}_n]), \tag{6}$$

and then combine the visual features with fused textual features $\tilde{\boldsymbol{T}} \in \mathbb{R}^{(|C_{\text{base}}|+|C_{\text{novel}}|)\times d_t'}$ to optimize the parameters of the transfer network by cross-entropy (CE) loss. Given a batch $B$ of visual samples and its labels as $\mathcal{S} = \{\boldsymbol{x}_i \in \mathbb{R}^{d_v}, l_i\}_{i=1}^B$, where $d_v$ is the dimension of the visual feature, the loss of textual space $\mathcal{L}^t$ can be calculated as:

$$\mathcal{L}^t = \frac{1}{B}\sum_{i=1}^{B} \text{CE}(\text{softmax}(\tilde{\boldsymbol{T}} \cdot \boldsymbol{x}_i^\top), l_i), \tag{7}$$

where minimizing $\mathcal{L}^t$ can be loosely considered as maximizing the association between the visual and the textual knowledge. And in our experiments, we set $d_t' = d_v$ to simplify the optimization stage.

### 3.2.2 *Visual Self-Transfer.*

For visual space, given a batch training set sampled from both the base categories and the novel categories $\mathcal{S} = \{\boldsymbol{x}_i, l_i\}_{i=1}^B$, the traditional classifier aims to fit these features to predict the category of the $n^{\text{th}}$ testing sample $\boldsymbol{x}_n$:

$$p_n = \text{Classifier}(\boldsymbol{x}_n) = \boldsymbol{W}^v \cdot \boldsymbol{x}_n^\top, \tag{8}$$

where $\boldsymbol{W}^v \in \mathbb{R}^{(|C_{\text{base}}|+|C_{\text{novel}}|)\times d_v}$ and $p_n \in \mathbb{R}^{|C_{\text{base}}|+|C_{\text{novel}}|}$ is the parameters and the prediction of the traditional classifier, respectively. However, the identifying process is susceptible to over-fitting on the novel categories with limited training samples. And [16] shows that a classifier trained under the supervision of hard labels without other assistant strategies biases the recognition. To alleviate these issues, we use the relations calculated from the textual space to adjust the weights of the classifier

$$\hat{\boldsymbol{W}}_n^v = \boldsymbol{A} \cdot \boldsymbol{W}_b^v, \tag{9}$$

where $\boldsymbol{A}$ is adjacency matrix, $\boldsymbol{W}_b^v \in \mathbb{R}^{|C_{\text{base}}|\times d_v}$ is the weight of the base categories in traditional classifier $\boldsymbol{W}^v$, $\hat{\boldsymbol{W}}_n^v \in \mathbb{R}^{|C_{\text{novel}}|\times d_v}$ represents the transferred weights of novel categories. $\hat{\boldsymbol{W}}_n^v$ provides extra potential knowledge from relations to improve the represented ability of novel categories.

In few shot learning task, the essential purpose of a classifier is to learn from the training samples and classify the testing samples. Thus, similar to the calculations in textual space, we combine the transferred weights with the original weights to fit the visual samples by cross-entropy loss

$$\mathcal{L}^v = \frac{1}{B}\sum_{i=1}^{B} \text{CE}(\text{softmax}(\tilde{\boldsymbol{W}^v} \cdot \boldsymbol{x}_i^\top), l_i), \tag{10}$$

where $\tilde{\boldsymbol{W}}^v = \boldsymbol{W}^v \oplus \hat{\boldsymbol{W}}_n^v$, $\tilde{\boldsymbol{W}}^v \in \mathbb{R}^{(|C_{\text{base}}|+|C_{\text{novel}}|)\times d_v}$, and ⊕ is elements summation.

## 3.3 Bidirectional Knowledge Transfer

For intra-modality knowledge transfer, it could be insufficient to infer novel categories by exploring only the visual or the textual information. Thus, KTN [26] maximize the consistency between the vision-based classifiers of base categories and external knowledge to help predict the category of a visual sample, where the consistency $C$ is simplified as

$$C = \sum_{c\in C_{\text{base}}} ||\boldsymbol{W}_c^v - \boldsymbol{T}_c||_2. \tag{11}$$

However, it only focuses on the relationship in the base categories and ignores that between the base and the novel categories. It performs good results since KTN chooses a sub-graph of WordNet [25] as a knowledge graph to model the semantic information, where WordNet contains all categories in the 21K ImageNet data [8]. Meanwhile, it needs that the semantic features are identical distribution and related to the validation data. Thus, the distribution of the novel categories can be inferred directly from the base categories.

In our method, we use the pre-trained word2vec method [23] to represent the features of semantics and calculate the relations between these features to construct the knowledge graph. It is easy to introduce bias from the original semantic knowledge. Thus, inspired by DML [53], we design a bidirectional connection strategy that calculates the consistency between all categories to alleviate the influences of the bias. Benefiting from the intra-modality knowledge transfer, the connections capture the correlation accurately from the transferred knowledge and provide the reversed supervision for the self-transfer process. The optimization of this connection can be calculated by using mean squared error (MSE) loss

$$\mathcal{L}^c = \frac{1}{|C_{\text{base}}| + |C_{\text{novel}}|}\Big(\sum_{c\in\{C_{\text{base}}+C_{\text{novel}}\}} ||\tilde{\boldsymbol{W}}_c^v - \tilde{\boldsymbol{T}}_c||_2\Big). \tag{12}$$

In our transfer method, the semantic-based knowledge transfer and the vision-based knowledge transfer are complementary to each other. Therefore, we propose an online fusion strategy to integrate them during the training and the inference stage, and optimize the parameters by the hard labels. Similar to the calculations in intra-modality transfer stage, given a batch training set $\mathcal{S} = \{\boldsymbol{x}_i, l_i\}_{i=1}^B$, the fusion strategy models the distribution over the all categories:

$$\mathcal{L}^m = \frac{1}{B}\sum_{i=1}^{B} \text{CE}(\text{softmax}((\tilde{\boldsymbol{W}}^v + \lambda\tilde{\boldsymbol{T}})) \cdot \boldsymbol{x}_i^\top), l_i), \tag{13}$$

where $\lambda$ is the hyper-parameter to control the weight of the fusion.

## 3.4 Training and Inference

For the training stage, we utilize hard labels to simultaneously optimize the parameters of different parts. Thus, the total loss $\mathcal{L}$ for a training batch is defined as

$$\mathcal{L} = \mathcal{L}^v + \mathcal{L}^t + \mathcal{L}^m + \mu\mathcal{L}^c, \tag{14}$$

where $\mu$ is weighting factor and $\mu$ is set to 100 in our experiments to balance the different losses.

In our method, we believe that the trained network can better express the distribution of the current training dataset. Thus we retrain our network with new semantic features under $\mathcal{L}$ after the first training. Specifically, we replace $\boldsymbol{T}$ of the second training with $\tilde{\boldsymbol{W}}^v + \lambda\tilde{\boldsymbol{T}}$ calculated from the first training to construct a new adjacency matrix $\tilde{\boldsymbol{A}}$, where $\tilde{\boldsymbol{W}}^v + \lambda\tilde{\boldsymbol{T}}$ is related to the training dataset and can be used to describe more accurate relations. The effectiveness of this strategy can be found in Section 4.3.3.

For the inference stage, we extract the predictions after second training from the connection module which benefits from both the visual and the textual knowledge, and classify the novel sample into a specific category by using the *argmax* function.

## 4 EXPERIMENTS

In this section, we conduct experiments to evaluate the performance of our proposed MDKT. Our experiments are intended to address the following research questions (RQ):
**RQ1:** What are the benefits of unidirectional knowledge transfer?
**RQ2:** How do the influences of bidirectional knowledge connection and fusion strategy?
**RQ3:** What are the effects of different training strategies of the proposed MDKT?
**RQ4:** How does MDKT perform top-$K$ accuracy as compared with the state-of-the-art FSL methods?

### 4.1 Datasets and Evaluations

**ImageNet-FS** contains 1000 categories. It is divided into 389 base categories $S_{\text{base}}$ and 611 novel categories $S_{\text{novel}}$, where 300 novel categories $S^1_{\text{novel}}$ are used for validating the hyper-parameters, and the remaining 311 novel categories $S^2_{\text{novel}}$ are used for classifier learning and testing. There are about 1300 samples in a base category $S_{\text{base}}$. For novel categories, there are 5 settings with $K = 1$, 2, 5, 10, and 20 support samples per category. The evaluation of this benchmark is to recognize the samples from these 311 novel categories $S^2_{\text{novel}}$. More details of the settings can be found in [16].
**ImNet** contains 1000 base categories and 360 novel categories. For novel categories, there are 5 settings with $K = 1, 2, 3, 4$, and 5 support samples per category. The evaluation of this benchmark is to recognize the samples from these 360 novel categories. More details are described in [19].
**Mini-ImageNet** consists of 100 categories and each category has 600 images. It is divided into three parts: 64 base categories, 16 novel categories for validation, and the remaining 20 novel categories for testing. This dataset is evaluated on several 5-way-$K$-shot classification tasks. In each task, 5 novel categories are sampled first, then $K$ samples in each of the 5 categories are sampled for training, and finally 15 samples (different from the previous $K$ samples) in each of the 5 categories are sampled for testing. To report the results, we sample 800 such tasks and average accuracies over all the tasks.

### 4.2 Experimental Setting

For fair comparisons with other methods, we use ResNet [17] as the feature extractor for the ImageNet-FS and ImNet, and use the recent popular backbone ResNet-12 for the Mini-ImageNet. We follow the training strategies in [42] to optimize the parameters of such feature extractors with the $\mathcal{D}_{\text{train}}$ under the supervision of hard labels in different datasets. The embedding sizes $d_v$ of visual feature represented by ResNet-10, ResNet-50, and ResNet-12 are 512, 2048, and 512, respectively. Meanwhile, we use the pre-trained word2vec [23] to represent the labels with vectors.

**Table 1: Top-5 accuracies (%) in the evaluation of the textual self-transfer on $S^1_{\text{novel}}$, where "w/o" means "without".**

| Training under * | $K = 1$ | $K = 2$ | $K = 5$ | $K = 10$ | $K = 20$ |
|---|---|---|---|---|---|
| $\mathcal{L}^t$ w/o Transfer | 60.8 | 68.5 | 75.5 | 77.9 | 80.2 |
| $\mathcal{L}^t$ | **60.9** | **68.6** | **76.0** | **79.6** | **81.5** |

**Table 2: Top-5 accuracies (%) in the evaluation of the visual self-transfer on $S^1_{\text{novel}}$.**

| Training under * | $K = 1$ | $K = 2$ | $K = 5$ | $K = 10$ | $K = 20$ |
|---|---|---|---|---|---|
| $\mathcal{L}^v$ w/o Transfer | 52.2 | 64.6 | 75.6 | **80.2** | **82.9** |
| $\mathcal{L}^v$ | **56.1** | **66.4** | **75.8** | 80.1 | 82.7 |

After the feature extraction stage, we first calculate the relationships between the textual features and choose the nearest $|\mathcal{M}_{\text{base}}| = 5$ base categories of each novel category to construct the adjacency matrix $\boldsymbol{A}$ in a compromised calculation and cost. $H$ is set to 8 by following in [37]. Then, we train the transfer parameters by the Adam optimization [18] with the starting learning rate of 0.001 and the weight decay of 0.001. For inference, followed by the strategy in [26] that gradually decreases $\lambda$ with increasing novel support samples, $\lambda$ is set to 2, 1, 1/2 when novel training shot $K = 1$, $1 < K < 10$, and $K \geq 10$, respectively.

### 4.3 Ablation Study

In the ablation study, we use the validation set $S^1_{\text{novel}}$ of ImageNet-FS with ResNet-50 to evaluate the effectiveness of the different parts of our method.

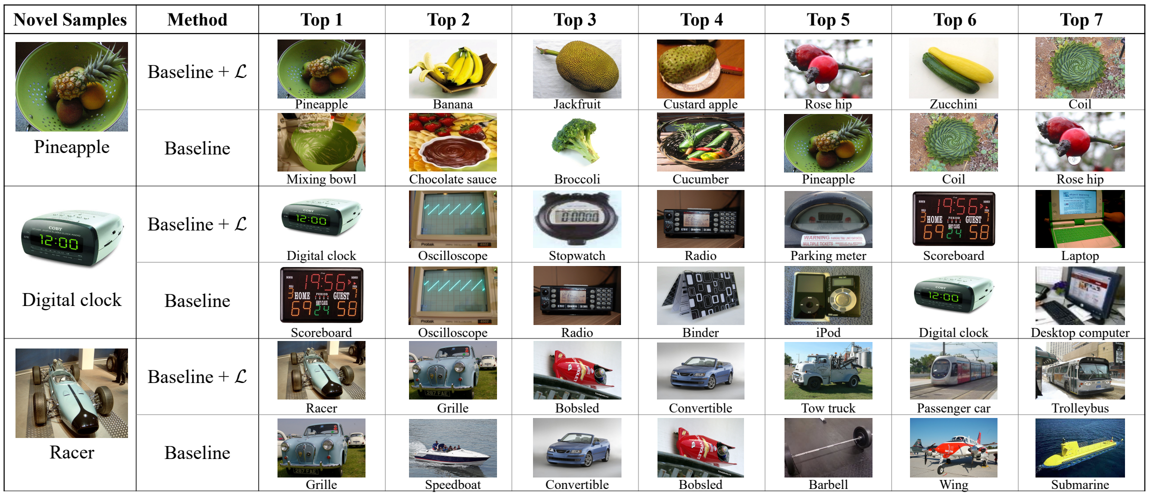
#### 4.3.1 *Unidirectional Knowledge Transfer.* (RQ1)

**Textual Self-Transfer.** In this ablation study, we combine the visual features $\boldsymbol{x}$ with different textual features, *i.e.*, $\bar{\boldsymbol{T}}$ and $\tilde{\boldsymbol{T}}$ to construct the textual-based classifier, and train the classifier under $\mathcal{L}^t$ in Eq. (7). As shown in Table 1, the results indicate that the textual knowledge ("Training under $\mathcal{L}^t$") can help the classifier recognize the novel samples. Specifically, compared with the training without transfer strategy, the introduced relations in $\tilde{\boldsymbol{T}}$ improve about 1% accuracy on 5-shot, 10-shot, and 20-shot tasks.

**Visual Self-Transfer.** In this ablation study, similar to the experiments in textual self-transfer, we use the original weights $\boldsymbol{W}^v$ and the transferred weights $\tilde{\boldsymbol{W}}^v$ of different classifiers to model the visual features as in Eq. (10). The results are shown in Table 2. It indicates that the introduced relations can help the basic network classify the novel samples, especially on 1-shot and 2-shot tasks. Specifically, It achieve 3.9% and 1.8% accuracy improvements for $K = 1$ and $K = 2$. Meanwhile, there are almost no losses of accuracy when large support samples, which validates the effectiveness of our visual self-transfer strategy.

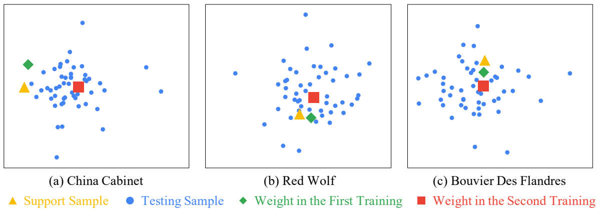
#### 4.3.2 *Bidirectional Knowledge Connection.* (RQ2)

Compared with the results on 1-shot and 2-shot tasks in Table 1, we can find that the classifier trained with only textual knowledge is



| Novel Samples | Method | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 | Top 6 | Top 7 |
|---|---|---|---|---|---|---|---|---|
| Pineapple | Baseline + $\mathcal{L}$ | Pineapple | Banana | Jackfruit | Custard apple | Rose hip | Zucchini | Coil |
| | Baseline | Mixing bowl | Chocolate sauce | Broccoli | Cucumber | Pineapple | Coil | Rose hip |
| Digital clock | Baseline + $\mathcal{L}$ | Digital clock | Oscilloscope | Stopwatch | Radio | Parking meter | Scoreboard | Laptop |
| | Baseline | Scoreboard | Oscilloscope | Radio | Binder | iPod | Digital clock | Desktop computer |
| Racer | Baseline + $\mathcal{L}$ | Racer | Grille | Bobsled | Convertible | Tow truck | Passenger car | Trolleybus |
| | Baseline | Grille | Speedboat | Convertible | Bobsled | Barbell | Wing | Submarine |

**Figure 4: The recognition results of several novel samples by the classifier with and without bidirectional knowledge connection, denoted as “Baseline + $\mathcal{L}$” and “Baseline”, respectively. In this experiment, $K = 1$.**



(a) China Cabinet (b) Red Wolf (c) Bouvier Des Flandres

**Figure 5: The visualizations of our retraining strategy. We use the T-SNE method [36] to visualize the distributions of support samples, testing samples, and the classifier weights of different training stages. In this experiment, $K = 1$.**

**Table 3: Top-5 accuracies (%) in the evaluation of the bidirectional knowledge connection on $S^1_{\text{novel}}$.**

| Training under * | | $K = 1$ | $K = 2$ | $K = 5$ | $K = 10$ | $K = 20$ |
|---|---|---|---|---|---|---|
| (1) | $\mathcal{L}^t + \mathcal{L}^v$ | 61.3 | 69.3 | 77.0 | 80.5 | 82.8 |
| (2) | $\mathcal{L}^m$ | 61.4 | 68.9 | 76.9 | 80.5 | 82.9 |
| (3) | $\mathcal{L}$ | **61.8** | **69.6** | **77.1** | **80.7** | **83.5** |

insufficient to achieve a better result when limited samples. Meanwhile, the results of “Training under $\mathcal{L}^v$” in Table 2 can hardly improve the performances on large training samples $K \geq 5$. Thus, it provides the possibility of knowledge fusion in different modalities. In this ablation study, we design three comparisons to evaluate the effectiveness of the bidirectional knowledge connection: (1) an offline fusion that fuses the results of two sub-models trained under the visual and the textual loss, *i.e.*, $\mathcal{L}^v$ and $\mathcal{L}^t$, independently. (2) an online fusion that is trained only under the connection loss $\mathcal{L}^m$. (3) an online fusion which is trained under the whole loss $\mathcal{L}$. As depicted in Table 3, the fusion results of (2) and (3) are better than that by separate training strategy with $\mathcal{L}^t$ or $\mathcal{L}^v$ in Table 1 and 2. Meanwhile, the performance of offline fusion is slightly better than that of “Training under $\mathcal{L}^m$” since there is more supervision ($\mathcal{L}^v$ and $\mathcal{L}^t$) in the separate training process. Furthermore, our bidirectional knowledge fusion “Training under $\mathcal{L}$” shows the best performance. Specifically, it achieves 0.5%-1% accuracy improvements for $S^1_{\text{novel}}$. It also indicates that our bidirectional connection provides a communication channel for capturing knowledge from different modalities.

**Table 4: Top-5 accuracies (%) in the evaluation of the retraining strategy on $S^1_{\text{novel}}$.**

| Method | $K = 1$ | $K = 2$ | $K = 5$ | $K = 10$ | $K = 20$ |
|---|---|---|---|---|---|
| First Training | 61.8 | 69.6 | 77.1 | 80.7 | 83.4 |
| Second Training | **62.5** | **70.0** | **77.2** | **80.8** | 83.4 |

In Figure 4, we show several examples of the results by the network with the bidirectional knowledge connection (denoted as “Baseline + $\mathcal{L}$”) and the network without it (denoted as “Baseline”). It is easy to see that the method of “Baseline + $\mathcal{L}$” obtains the top-ranked results that are more relevant to the input objects. For example, when the input novel image is a kind of fruit (“Pineapple” here), all the top 5 results of “Baseline + $\mathcal{L}$” are fruit labels, but the first and the second results (“Mixing bowl” and “Chocolate sauce”) of “Baseline” are mislead by the textures of the bowl. This figure shows the effectiveness of the bidirectional knowledge connection.

### 4.3.3 *Retraining Strategy.* (**RQ3**)

In this ablation, we retrain our network with a new adjacency matrix $\tilde{A}$ calculated from $\tilde{W}^v + \lambda\tilde{T}$ of the first training stage and compare the performances of different training stages in Table 4. We can find the performances of the second training stage are better than that in the first training stage, especially on 1-shot and 2-shot tasks, which are significant for FSL tasks.

In Figure 5, we show the distribution of several novel features and the weights from different training stages. It is easy to see that the weights of the first training stage (◆) are closed to the support

Table 5: Top-5 accuracies (%) by different methods on $S^2_{\text{novel}}$ of ImageNet-FS.

| | Method with ResNet-10 | | | | | Method with ResNet-50 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $K = 1$ | $K = 2$ | $K = 5$ | $K = 10$ | $K = 20$ | $K = 1$ | $K = 2$ | $K = 5$ | $K = 10$ | $K = 20$ |
| Prototypical Nets [34] | 39.3 | 54.4 | 66.3 | 71.2 | 73.9 | 49.5 | 59.9 | 70.1 | 75.1 | 77.6 |
| Matching Networks [39] | 43.6 | 54.0 | 66.0 | 72.5 | 76.9 | 49.6 | 64.0 | 74.4 | 78.1 | 80.0 |
| SGM + Hallucination[16] | 44.3 | 56.0 | 69.7 | **75.3** | **78.6** | 52.8 | 64.4 | 77.3 | **82.0** | **84.9** |
| wDAE-GNN [11] | 48.0 | 59.7 | 70.3 | 75.0 | 77.8 | —— | —— | —— | —— | —— |
| KTCH [23] | —— | —— | —— | —— | —— | 58.1 | 67.3 | **77.6** | 81.8 | 84.2 |
| IDeMe-Net [3] | 51.0 | 60.9 | 70.4 | 73.4 | 75.1 | 60.1 | 69.6 | 77.4 | 80.2 | —— |
| KTN [26] | 54.7 | 61.7 | 70.4 | 75.0 | 77.9 | 61.9 | 68.7 | 76.4 | 80.1 | 82.4 |
| Our MDKT | **55.2** | **63.2** | **70.8** | 75.0 | 78.2 | **62.6** | **70.1** | **77.6** | 81.5 | 83.7 |

Table 6: Top-5 accuracies (%) by different methods on the novel categories from ImNet.

| Method | Novel Categories | | | | |
|---|---|---|---|---|---|
| | $K = 1$ | $K = 2$ | $K = 3$ | $K = 4$ | $K = 5$ |
| NN (from [23]) | 34.2 | 43.6 | 48.7 | 52.3 | 54.0 |
| PPA [27] | 33.0 | 43.1 | 48.5 | 52.5 | 55.4 |
| LSD [5] | 33.2 | 44.7 | 50.2 | 53.4 | 57.6 |
| KTCH [23] | 39.0 | 48.9 | 54.9 | 58.7 | 60.5 |
| KGTN [1] | 42.5 | 50.3 | 55.4 | 58.4 | 60.7 |
| Our Baseline | 36.1 | 47.9 | 54.0 | 58.1 | 60.8 |
| Our MDKT | **44.4** | **53.3** | **58.1** | **61.7** | **63.8** |

Table 7: Top-1 accuracies (%) by different methods on the testing novel categories of Mini-ImageNet.

| Method | $K = 1$ | $K = 5$ |
|---|---|---|
| Meta-Baseline [2] | 63.17 ± 0.23% | 79.26 ± 0.17% |
| MetaFun [48] | 64.13 ± 0.13% | 80.82 ± 0.17% |
| P-Transfer [33] | 64.21 ± 0.77% | 80.38 ± 0.59% |
| MMKD [42] | 64.40 ± 0.43% | 83.05 ± 0.28% |
| IEPT [52] | 67.05 ± 0.44% | 82.90 ± 0.30% |
| FRN [44] | 66.45 ± 0.19% | 82.83 ± 0.13% |
| BML [54] | 67.04 ± 0.63% | **83.63** ± 0.29% |
| Our MDKT | **67.39** ± 0.76% | 82.25 ± 0.53% |

samples (▲) and the weights of the second training stage (■) tend to the center of the testing samples (●). This figure clearly shows the effectiveness of the retraining strategy.

## 4.4 Comparisons with Other Methods (RQ4)

**ImageNet-FS**. The compared methods include Prototypical Nets (PN) [34], Matching Networks (MN) [39], SGM [16], wDAE-GNN [11], KTCH [23], IDeMe-Net [3], and KTN [26]. All the results with ResNet-10 and ResNet-50 on $S^2_{\text{novel}}$ are listed in Table 5. Our method outperforms others in the cases of small training samples ($K < 10$) and achieves comparable results with SGM [16] on $K \geq 10$ tasks. Note that SGM is a hallucination method which expand the training set for classifier training.

**ImNet**. The compared methods include Nearest Neighbor (NN) [23], PPA [27], LSD [5], KTCH [23], and KGTN [1]. The Top-5 accuracies by our and these methods on the novel categories are listed in Table 6. We can see that our method performs best on this dataset in all the cases. Compared with the "Baseline" which uses one-layer perceptron as visual classifier, our method achieves the huge improvements on all tasks, especially on 1-shot. Compared with the previous best model KGTN, our improvements for $K = 1$, 2, 3, 4, and 5 are 1.9%, 3.0%, 2.7%, 3.3%, and 3.1%, respectively.

**Mini-ImageNet**. The compared methods include Meta-Baseline [2], MetaFun [48], P-Transfer [33], MMKD [42], IEPT [52], FRN [44], and BML [54]. As shown in Table 7, our MDKT achieves nearly 3-4% improvements for the baseline method Meta-Baseline (we use its extracted features as our model input). Particularly, MDKT that uses pre-trained features can even achieve competitive performance to BML which is an end-to-end learning method.

## 5 CONCLUSION

In this paper, we have proposed multi-directional knowledge transfer to tackle the problem of few-shot learning. Specifically, (1) two intra-modality unidirectional knowledge transfer strategies in the visual and the textual space calibrate the distributions of the novel categories from the base categories. (2) the inter-modality bidirectional connection reduces the influences of semantic gaps between the visual and the textual knowledge. (3) an online fusion and re-training strategy both improve the prediction accuracy of the novel categories. The extensive experiments have demonstrated the effectiveness of our proposed method on three benchmarks, especially on 1-shot and 2-shot training tasks.

Note that the improvements of our MDKT gradually decreased when the number of support samples gradually increased. It is because the visual sample brings rich descriptions for training, which leads the classifier to ignore the relatively weak knowledge of semantics. In our future work, we will focus on boosting the performances on tasks with a large number of support samples by using such semantics.

## ACKNOWLEDGMENTS

The work was supported in part by the National Key Research and Development Program of China (2021ZD0111802), by the National Natural Science Foundation of China (No. U21B2026), and by The University Synergy Innovation Program of Anhui Province (No. GXXT-2021-009).

## REFERENCES

[1] Riquan Chen, Tianshui Chen, Xiaolu Hui, Hefeng Wu, Guanbin Li, and Liang Lin. 2020. Knowledge graph transfer network for few-shot recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[2] 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*.

[3] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. 2019. Image deformation meta-networks for one-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8680–8689.

[4] Lechao Cheng, Zunlei Feng, Xinchao Wang, Ya Jie Liu, Jie Lei, and Mingli Song. [n. d.]. Boundary Knowledge Translation based Reference Semantic Segmentation. ([n. d.]).

[5] Matthijs Douze, Arthur Szlam, Bharath Hariharan, and Hervé Jégou. 2018. Low-shot learning with large-scale diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[6] Zunlei Feng, Lechao Cheng, Xinchao Wang, Xiang Wang, Ya Jie Liu, Xiangtong Du, and Mingli Song. 2021. Visual Boundary Knowledge Translation for Foreground Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1334–1342.

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*.

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems* (2013).

[9] Hang Gao, Zheng Shou, Alireza Zareian, Hanwang Zhang, and Shih-Fu Chang. 2018. Low-shot learning via covariance-preserving adversarial augmentation networks. *Advances in Neural Information Processing Systems* (2018).

[10] Spyros Gidaris and Nikos Komodakis. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[11] Spyros Gidaris and Nikos Komodakis. 2019. Generating Classification Weights With GNN Denoising Autoencoders for Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[12] Dan Guo, Shuo Wang, Qi Tian, and Meng Wang. 2019. Dense Temporal Convolution Network for Sign Language Translation.. In *International Joint Conference on Artificial Intelligence*. 744–750.

[13] Yanbin Hao, Chong-Wah Ngo, and Bin Zhu. 2021. Learning to match anchor-target video pairs with dual attentional holographic networks. *IEEE Transactions on Image Processing* 30 (2021), 8130–8143.

[14] Yanbin Hao, Shuo Wang, Pei Cao, Xinjian Gao, Tong Xu, Jinmeng Wu, and Xiangnan He. 2022. Attention in Attention: Modeling Context Correlation for Efficient Video Classification. *IEEE Transactions on Circuits and Systems for Video Technology* (2022).

[15] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. 2022. Group Contextualization for Video Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 928–938.

[16] Bharath Hariharan and Ross Girshick. 2017. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision*. 3018–3027.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[18] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

[19] Elyor Kodirov, Tao Xiang, and Shaogang Gong. 2017. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.

[21] Kwonjoon Lee, Subhransu 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.

[22] 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.

[23] Aoxue Li, Tiange Luo, Zhiwu Lu, Tao Xiang, and Liwei Wang. 2019. Large-scale few-shot learning: Knowledge transfer with class hierarchy. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 7212–7220.

[24] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. 2020. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[25] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[26] 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.

[27] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. 2018. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[28] Danilo Rezende, Ivo Danihelka, Karol Gregor, Daan Wierstra, et al. 2016. One-shot generalization in deep generative models. In *International conference on machine learning*. PMLR.

[29] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. 2019. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*.

[30] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks* 20, 1 (2008), 61–80.

[31] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[32] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. 2018. Delta-encoder: an effective sample synthesis method for few-shot object recognition. *Advances in Neural Information Processing Systems* (2018).

[33] Zhiqiang Shen, Zechun Liu, Jie Qin, Marios Savvides, and Kwang-Ting Cheng. 2021. Partial is better than all: Revisiting fine-tuning strategy for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[34] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*.

[35] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[36] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[37] 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).

[38] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. *International Conference on Learning Representations* (2018).

[39] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*.

[40] Shuo Wang, Huixia Ben, Yanbin Hao, Xiangnan He, and Meng Wang. 2022. Boosting Hyperspectral Image Classification with Dual Hierarchical Learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).

[41] Shuo Wang, Dan Guo, Wen-gang Zhou, Zheng-Jun Zha, and Meng Wang. 2018. Connectionist temporal fusion for sign language translation. In *Proceedings of the 26th ACM international conference on Multimedia*. 1483–1491.

[42] Shuo Wang, Jun Yue, Jianzhuang Liu, Qi Tian, and Meng Wang. 2020. Large-scale few-shot learning via multi-modal knowledge discovery. In *European Conference on Computer Vision*. Springer, 718–734.

[43] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. 2018. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7278–7286.

[44] Davis Wertheimer, Luming Tang, and Bharath Hariharan. 2021. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8012–8021.

[45] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. 2018. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[46] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. 2019. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10275–10284.

[47] 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).

[48] Jin Xu, Jean-Francois Ton, Hyunjik Kim, Adam Kosiorek, and Yee Whye Teh. 2020. Metafun: Meta-learning with iterative functional updates. In *International Conference on Machine Learning*. PMLR, 10617–10627.

[49] Shuo Yang, Lu Liu, and Min Xu. 2020. Free lunch for few-shot learning: Distribution calibration. *International Conference on Learning Representations* (2020).

[50] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[51] Dingwen Zhang, Wenyuan Zeng, Guangyu Guo, Chaowei Fang, Lechao Cheng, and Junwei Han. 2021. Weakly Supervised Semantic Segmentation via Alternative Self-Dual Teaching. *arXiv preprint arXiv:2112.09459* (2021).

[52] Manli Zhang, Jianhong Zhang, Zhiwu Lu, Tao Xiang, Mingyu Ding, and Songfang Huang. 2020. IEPT: Instance-level and episode-level pretext tasks for few-shot learning. In *International Conference on Learning Representations*.

[53] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4320–4328.

[54] Ziqi Zhou, Xi Qiu, Jiangtao Xie, Jianan Wu, and Chi Zhang. 2021. Binocular mutual learning for improving few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8402–8411.

[55] Bin Zhu and Chong-Wah Ngo. 2020. CookGAN: Causality based text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5519–5527.

[56] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11477–11486.