Language-Inspired Relation Transfer for Few-Shot Class-Incremental Learning

Yifan Zhao, *Member, IEEE*, Jia Li, *Senior Member, IEEE*, Zeyin Song, and Yonghong Tian, *Fellow, IEEE*

***Abstract*—Depicting novel classes with language descriptions by observing few-shot samples is inherent in human-learning systems. This lifelong learning capability helps to distinguish new knowledge from old ones through the increase of open-world learning, namely Few-Shot Class-Incremental Learning (FSCIL). Existing works to solve this problem mainly rely on the careful tuning of visual encoders, which shows an evident trade-off between the base knowledge and incremental ones. Motivated by human learning systems, we propose a new Language-inspired Relation Transfer (LRT) paradigm to understand objects by joint visual clues and text depictions, composed of two major steps. We first transfer the pretrained text knowledge to the visual domains by proposing a graph relation transformation module and then fuse the visual and language embedding by a text-vision prototypical fusion module. Second, to mitigate the domain gap caused by visual finetuning, we propose context prompt learning for fast domain alignment and imagined contrastive learning to alleviate the insufficient text data during alignment. With collaborative learning of domain alignments and text-image transfer, our proposed LRT outperforms the state-of-the-art models by over 13% and 7% on the final session of *mini*ImageNet and CIFAR-100 FSCIL benchmarks.**

***Index Terms*—Few-shot learning, class-incremental learning, language-inspired relation transfer.**

I. Introduction

HUMAN brains show their distinctive advantages in recognizing new concepts with only a few limited samples, while not forgetting the old knowledge rapidly. Benefited from the strong perceptual capability of deep neural networks [1], [2], recent advances propose to imitate human learning systems mainly from two aspects, i.e., recognizing new concepts with extremely few samples and learning without forgetting. For the first few-shot learning (FSL) challenge, existing works focus on network learning with fast optimization strategies [3], [4], [5] or measuring with appropriate metrics [6], [7], [8]. And to solve the second challenge as well as alleviate forgetting, class-incremental learning (CIL) methods have made significant progress with mechanisms including rehearsal [9], [10], [11], novel model consolidation [12], [13] and feature space regularization strategies [14]. Nevertheless, when considering these two natural abilities together, unlike human-learning systems, existing methods encounter significant obstacles [15] for generalizing on new concepts or catastrophic forgetting on base knowledge due to the limited new samples for training.

One intuitive idea to solve this problem, i.e., few-shot class-incremental learning (FSCIL), is to adopt knowledge distillation [16], [17] from base classes when gradually learning new concepts. As only a few samples are accessible during incremental phases, naive distillations with these seen samples also lead to severe overfitting. To alleviate this, prevailing works dedicate to decoupling base and incremental learning stages [18] and then fix or slightly tune the backbone representations [19], [20], [21]. Besides these works, other research efforts tend to find generalized feature representations by using sufficient training data from base sessions. Representative works propose to find the flattened region in optimization [22] or construct virtual classes [23] when sufficient training samples are available. Although these works tried to achieve a balanced performance trade-off between the base classes and incremental classes, the dilemma still exists: how to represent the novel incremental classes well without losing distinguishability on the base classes?

When sufficient training samples are available, supervised learning systems present superior performances with state-of-the-art visual encoders. As in Fig. 1(a), visual prototypes of base seen classes filled the embedding space and show clear classification boundaries. However, in Fig. 1(b), during incremental sessions, features of new classes are still represented with the identical encoder that is trained with base classes, which thus leads to prototype confusion or topological damages. In this paper, we argue to solve this dilemma by a new Language-inspired Relation Transfer (LRT) paradigm, which is motivated by the recent advances of Contrastive Language-Image Pretraining (CLIP) [2]. Different from prevailing methods with static visual embedding, when learning a novel category of *Cardinal Bird* in Fig. 1(b), we introduce the language prompt (several words or description sentences) as auxiliary information if there are no sufficient visual samples. Besides, this contrastive learning paradigm constructs a unified feature alignment space of text prompts and image-level features. Hence to transfer

Received 31 March 2023; revised 29 September 2024; accepted 21 October 2024. Date of publication 6 November 2024; date of current version 9 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62132002, Grant 62425101, Grant 62088102, and Grant 62202010, and in part by the Fundamental Research Funds for the Central Universities. Recommended for acceptance by T. Tommasi. *(Corresponding authors: Jia Li; Yonghong Tian.)*

Yifan Zhao and Jia Li are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: zhaoyf@buaa.edu.cn; jiali@buaa.edu.cn).

Zeyin Song is with the School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China.

Yonghong Tian is with the School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China, also with the School of Computer Science, Peking University, Beijing 100871, China, and also with Pengcheng Laboratory, Shenzhen 518055, China (e-mail: yhtian@pku.edu.cn).

Digital Object Identifier 10.1109/TPAMI.2024.3492328

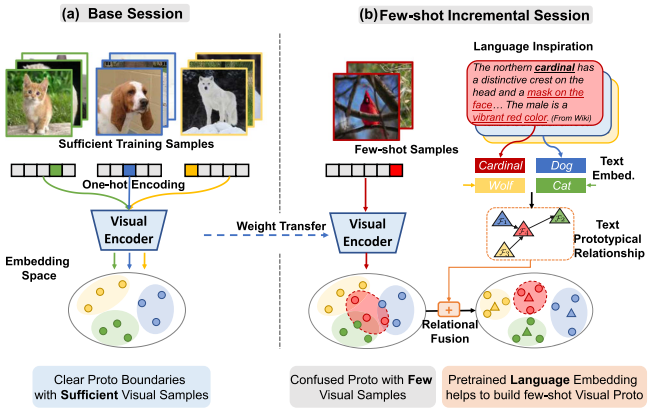

Fig. 1. The motivation of the proposed approach. Visual encoders provide clear boundaries in (a) when learning with base sufficient data, while resulting in confused prototypes with only a few samples of novel classes in (b). Our proposed LRT aims to transfer the pretrained language relationships to help construct a joint feature representation of both base and novel classes.

the text knowledge, our method first builds a graph relation transformation module to transfer the well-embedded language relationships to the inferior visual space, thus the tangled visual features can be reprojected in the correct space driven by the strong language guidance.

Recent trends to predict object classes in CLIP-based models [2], [24], [25] is to calculate the similarity between text and image embedding. Although this zero-shot trend provides preferable generalization capabilities on new classes, it remains a gap [26] compared to the performances using fully supervised visual models. Combining the merits of supervised vision models and language-vision contrastive relations, we consolidate text embedding of one category for prototypical representation and then propose a text-vision prototypical fusion module to incorporate representations from both visual and text domains. In this way, the well-trained language embedding provides strong backing for the standard visual representations, especially in data-deficient few-shot scenarios. However, only finetuning visual data would lead to a misalignment of visual and language domains. Thus we first introduce a context text prompt learning module to depict few-shot visual samples with learnable text prompts instead of the hand-crafted ones in vanilla CLIP [2], which fast mitigates the domain gap with only few incremental samples.

Beyond these improvements in the knowledge transfer module, we also notice that the multi-modality contrastive training would be easy to overfit on specific data domains. It is because although the image visual data are various and sufficient, its corresponding language descriptions (i.e., label texts in our approach) are *monotonous*. Thus to solve this brand new problem, for the multimodal alignment, we randomly mix the input images and also mix their text labels including the learnable prompt tokens as a virtual class. Then the imagined contrastive learning is conducted among these *imagined* prototypes and theoretically N times (N is the number of classes) larger than the vanilla text input space. With the collaboration of text-to-image relation transfer and multi-modal alignment, our proposed LRT is able to achieve a comprehensive understanding of one novel concept without forgetting the old ones. Moreover, LRT does not rely on any auxiliary networks (including the text encoder) during the inference time, making the final model lightweight and implementation-friendly. Experimental evidence demonstrates that LRT outperforms the state-of-the models by 13.3% on *mini*ImageNet [27] and 7.3% on CIFAR-100 [28] benchmarks in the final session.

In summary, our contribution is threefold: 1) We make an attempt to solve the few-shot class-incremental learning with pretrained language understanding and propose a new Language-inspired Relation Transfer (LRT) paradigm. 2) We propose a graph relation transformation module to gradually transfer the text knowledge into few-shot visual prototypes, and introduce a text-vision prototypical fusion strategy for feature representation, which combines the merits of the visual embedding and pretrained language guidance. 3) We propose a context text prompt learning strategy to align the text and image domains with few shots and an imagined contrastive learning strategy to alleviate the *insufficient text* label spaces for generalization representation.

The remainder of this paper is organized as follows: Section II reviews related works and discusses the relations among previous literature. Section III describes the proposed language-inspired relation transfer approach. Qualitative and quantitative experiments with detailed analyses are exhibited in Sections IV and V finally concludes this paper.

II. Related Work

Few-Shot Learning: Inspired by human recognition systems, few-shot learning aims to distinguish conceptually new object categories by inferring from base knowledge. Recent ideas to solve this problem could be roughly divided into two trends: model optimization [3], [4], [5], [29], [30], [31] and metric learning manners [6], [7], [8], [32], [33], [34]. Optimization-based methods focuses on the generalization ability by using meta-learning frameworks. For example, model-agonistic meta-learning [30], [31] aims to learn the fast adaptation ability by learning from the direction of sampled task gradients. While metric-learning-based methods focus on the distance measurement of novel query samples and base knowledge representations. Representative works focuses on the prototype learning [6], local representations [35] and feature space reprojections [36].

Class-Incremental Learning: Class-Incremental Learning (CIL) focuses on one specific direction of the field of continual learning [37], which aims to learn from new classes without forgetting the base knowledge. Prevailing research dedicated to this task focuses on replaying the old memories [9], [10], [11], [38], [39] and regularizing the feature space [14], [40], [41]. Representative methods in the first family including iCaRL [9], CLEAR [11] and A-GEM [10] selectively retain the knowledge from old samples and replay these samples or features when learning the new classes. For example, iCaRL [9] aims to distill the base knowledge when learning samples from new categories,

which greatly alleviates catastrophic forgetting. While the second family of methods [14], [40] tends to build regularized feature space and Besides these with fixed model structures, the other line of works proposes to solve this problem by model ensemble [12] and iterative pruning [13]. This research direction also shares common concerns with few-shot learning to represent new classes. However, when tackling incremental categories with very few samples, rehearsal or distillation-based methods usually face severely catastrophic overfitting and fail to represent the novel categories.

Few-Shot Class-Incremental Learning: As a newly proposed realistic setting, Few-Shot Class-Incremental Learning (FSCIL) proposed by [15] has attracted considerable attention. Inspired by incremental learning methods, several research [16], [17] propose to alleviate the forgetting of base classes by knowledge distillation during few-shot learning. Zhao et al. [42] propose a slow-fast updating framework to achieve a balanced trade-off between the novel updating and old knowledge degradation. As the base samples during incremental learning are infeasible, prevailing methods [22], [23], [43], [44], [45] tend to find the generalized representation during base sessions. For example, Zhou et al. [45] propose to synthesize fake FSCIL tasks from the base dataset with meta-learning strategies. Besides, other works propose to resist the overfitting caused by insufficient training samples by using graph models [18] or selected parameter adjustment [19], [20], [21]. Hersche et al. [21] design a semi-frozen meta-learning framework with rewritable dynamically growing memory. However, although the discovery abilities of novel categories are improved, the frozen visual backbones still restrict their representation abilities to extract sufficient visual cues.

Contrastive Vision-Language Model: Cross-modality pretraining with self-supervised contrastive learning has been widely adopted in various applications. Representative vision-language models including CLIP [2], ALIGN [46] and CyCLIP [47] have shown great success in zero-shot image recognition tasks. Inspired by these works, contrastive pretraining using multi-view [48] or part-level supervision [24] has enlightened many down-stream vision tasks, e.g., zero-shot object detection and visual question answering. Moreover, several very recent works focus on prompt engineering to make a fast adaptation on target domains, including vision prompt [49] and language prompt [25]. Although these aforementioned methods show effectiveness in zero-shot learning, as validated in [25], [26], there is still a huge gap between supervised learning and CLIP-based models. In addition, when jointly optimizing these models, the base classes and novel categories show less distribution gap which cannot be jointly optimized in the few-shot class-incremental setting.

Discussions and Relations: Methods of few-shot learning and class-incremental learning only focus on the single side of the FSCIL problem. Prevailing few-shot class-incremental learning methods achieve preferable performance by alleviating the catastrophic overfitting of base sessions, while the novel discovery capability is still restricted by the inferior representation features trained by limited samples. To overcome this bottleneck, in this paper, we argue that one promising solution to understanding few-shot objects with incremental ability is from the generalized visual-language knowledge: 1) fast adapting the generalized representation to downstream task-specific features, 2) excavating generalized language knowledge to guide the learning of few-shot visual samples, and 3) maintaining the text-image cross-modal alignment with only few samples.

III. Approach

A. Problem Formulations and Baselines

Few-Shot Class-Incremental Learning With Text: FSCIL focuses on the intersection of class-incremental learning and few-shot learning problems, which aims to jointly recognize the incremental classes and base classes with a sequential of given sessions. An FSCIL model sequentially receives S training session $\mathcal{D}^1 \dots \mathcal{D}^S$ with sets of triplets. i.e., $\mathcal{D}^s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s, \mathbf{t}_i^s)\}_{i=1}^{|\mathcal{D}^s|}$, where $\mathbf{x}_i^s \in \mathcal{X}^s, (\mathbf{y}_i^s, \mathbf{t}_i^s) \in \mathcal{C}^s \times \mathcal{E}^s$ denotes the training images, one-hot labels, and text labels with class names respectively. $\mathcal{X}, \mathcal{C}$ and $\mathcal{E}$ are space notations for the visual, label, and text domains. During the training of FSCIL, the base session $\mathcal{D}^1$ contains sufficient training samples of base classes, and the subsequent $2 \sim S$ sessions are defined as typical N-way M-shot few-shot learning problems. During the incremental session, only samples in the current session are visible and label spaces do not contain any overlap. $\mathcal{C}^i \cap \mathcal{C}^j = \varnothing, \forall i, j \in \{1 \dots S\}, i \neq j$, and similarly we have $\mathcal{D}^i \cap \mathcal{D}^j = \varnothing, \mathcal{E}^i \cap \mathcal{E}^j = \varnothing$. With learnable parameters Θ, the overall learning objective is to minimize the measurement ξ across all sessions:

$$\arg\min_{\Theta} \Sigma_{s=1}^{S} \Sigma_{(\mathbf{x},\mathbf{y},\mathbf{t})\sim\mathcal{D}^s} \xi(f_\Theta(\mathbf{x};\mathbf{t}), \mathbf{y}), \tag{1}$$

where ξ are usually set as cosine or Euclidean distances with the one-hot class label $\mathbf{y}$ and the label text $\mathbf{t}$ for each class are introduced as auxiliary input.

Visual Learning Baseline: One intuitive but effective visual learning scheme recently [18], [23] is to use prototypical networks [6] both for base and incremental sessions. We denote $\mathcal{V}_B, \mathcal{V}_I$ for visual encoders of base and incremental sessions respectively. The prototypical networks rely on the slowly updated or fixed visual encoder $\mathcal{V}_B$ that is pretrained on base sessions. During the incremental session, the weight of base visual encoder is transferred to the incremental visual encoder $\mathcal{V}_I(\cdot) \leftarrow \mathcal{V}_B(\cdot)$ (fixed or slightly tuned). The fully connected layers for classifiers are replaced with feature prototypes, as in Fig. 2(a). Thus the visual prototypes $\mathbf{V}^i$ across all classes have the form:

$$\mathbf{V}^i = \frac{1}{WH} \sum_{(\mathbf{x},\mathbf{y})\sim\mathcal{D},\mathbf{y}=\mathcal{C}_i} \sum_{j=1}^{WH} \mathcal{V}_{\{B,I\}}(\mathbf{x}_j; \mathcal{C}_i), \tag{2}$$

where W, H denote the width and height of feature maps respectively. With the averaged prototypes of all classes $\{\mathbf{V}^i\}_{i=1}^{|\mathcal{C}|} \in \mathbb{R}^{1\times1\times D_V}$ and the standard measurements $\xi(\mathbf{x}, \mathbf{V})$, visual models show strong capabilities in alleviating *catastrophic forgetting*. Nevertheless, they are easy to overfit on the limited few-shot incremental data and cannot form the *generalized embedding*.

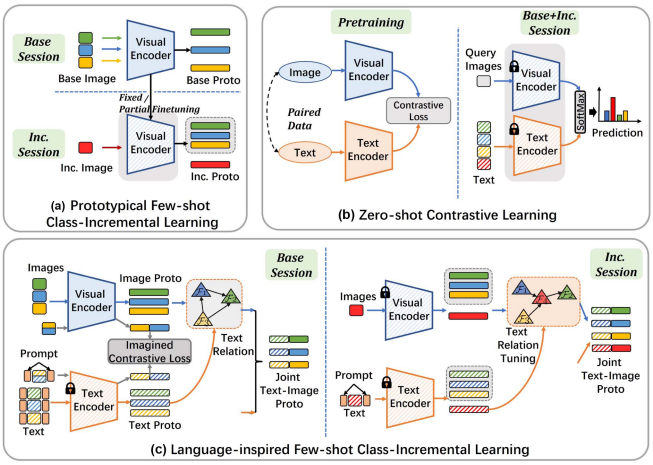

Fig. 2. Illustrations of different learning paradigms. a) Prototypical FSCIL [18], [23]: using visual prototypes for incremental classes. b) Zero-shot CLIP [2]: direct predicting probabilities after image-text contrastive learning. c) Ours: transferring the pretrained text embedding to visual domains meanwhile keeping domain alignment with context prompt and imagined contrastive loss.

B. Connecting Images With Texts in FSCIL

Language-Guided FSCIL Paradigm: Our main motif is to utilize the pretrained knowledge in the text domain to facilitate the learning of few-shot class incremental sessions. To achieve this, we face two major dilemmas beyond the prevailing incremental learning challenges, 1) visual representation scarcity of novel concepts and 2) continual misalignment of multi-modalities caused by imbalanced and downstream learning tasks. Toward these dilemmas, our major pipeline can be simplified as two major steps as in Fig. 4, i.e., *transferring* and *aligning*. For the first dilemma, we advocate transferring the pretrained generalized language concept knowledge to the visual modality by relational knowledge transfer module in Section III-C. Note that this module is constructed for both the base and incremental learning sessions. For the second misalignment dilemma, in Section III-D, we propose the imagined aligning strategy for the base pretraining session and context prompt adaptation only for the incremental session, which jointly alleviate severely misaligned text and visual modality during learning.

Zero-Shot Measurements With Texts: Contrastive pretraining vision-language models including CLIP [2] and ALIGN [46], have offered us a conceptually new solution to solve the few-shot representation predicament. As in Fig. 2(b), taking the advantages of rich language data, this contrastive learning trend shows significant *generalization* ability on extremely few-shot image samples. Given a cluster of N text labels to predict, i.e.,$\{(\mathbf{t},\mathbf{y})|\mathbf{y}=\mathcal{C}_i\}_{i=1}^{N}$, the text encoder $\mathcal{T}(\cdot)$ aligns the text inputs and the image features of query $\mathbf{x}$ in the same space. Hence the zero-shot prediction of each class $\mathcal{C}_i$ is presented as:

$$\mathbf{p}_i = \frac{e^{\xi(\mathcal{V}(\mathbf{x}),\mathcal{T}(\mathbf{t}_i)^\top)}}{\sum_{j=i}^{N} e^{\xi(\mathcal{V}(\mathbf{x}),\mathcal{T}(\mathbf{t}_j)^\top)}}, \tag{3}$$

where $\xi(\mathbf{x},\mathbf{t}) = \mathbf{x}\cdot\mathbf{t}/(||\mathbf{x}||_2^2||\mathbf{t}||_2^2)$ denotes the normalized cosine similarity with omitted scale factors for simplicity. However, this equation only measures the similarity of input images with text prototypes, while omitting the similarity of input to image prototypes.

C. Relational Knowledge Transfer

Although the CLIP-based models show strong generalization capability on unseen categories, FSCIL faces a conceptually different problem, i.e., sufficient visual samples of base classes are available and few-shot incremental samples in the same vein. Omitting these visual clues as well as the class-based prototypes would lead to overfitting due to the target dataset being small compared to the large pretraining domain. Besides, the language-vision pretraining models show a clear performance drop compared with the supervised learning, as demonstrated in [2], [26]. Toward this end, we propose constructing two major modules for knowledge transfer, i.e., the language-guided graph relation transfer (Fig. 4(c)) and text-vision prototypical fusion (Fig. 4(d)). These two modules are consistently constructed for both the base pretraining session and the subsequent incremental learning session.

Language-Guided Graph Relation Transfer: Inspired by the prototypical learning, here we adopt the text encoding features of the same class $\mathbf{T}_i = 1/K\sum_k \mathcal{T}(\mathbf{t}_k), \forall \mathbf{y}_k = \mathcal{C}_i$ as the *text prototypes* to represent the features of class $\mathcal{C}_i$, where K denotes the number of text prompts. This embedding can be formed by using class names or even incorporating the object context features, e.g., *"cardinals are usually in red color"* in Fig. 1, which can provide rich prior knowledge for object recognition, especially in few-shot scenarios. More importantly, the generated text prototypes are naturally distributed in the same space as the visual features and benefited from the contrastive multi-modal pretraining. Considering Fig. 3, the most crucial challenge in incremental sessions is that the new visual prototypes are entangled with the base ones. We therefore decide to disentangle these confused samples by introducing the relationship from pretrained language domain. The pair-wise relationship of text prototypes $\mathcal{T}(\mathbf{t}_i) \in \mathbb{R}^{1\times1\times D_T}$ is:

$$\mathbf{A}_{i,j} = \frac{\mathcal{T}(\mathbf{t}_i)^\top \cdot \mathcal{T}(\mathbf{t}_j)}{\|\mathcal{T}(\mathbf{t}_i)\|\,\|\mathcal{T}(\mathbf{t}_j)\|}. \tag{4}$$

We then construct a relationship transformation graph with the visual prototypes as graph nodes, i.e., $\mathcal{G} = \{\mathbf{V},\mathbf{A}\}, \mathbf{V} = \{\mathcal{V}(\mathbf{x})\}_{i=1}^{|\mathcal{C}|}$ and the $\mathcal{C}$ can denote base classes $\mathcal{C}^{base}$ or $[\mathcal{C}^{base},\mathcal{C}^{inc}]$ during the incremental session. With the relation adjacent matrix, the reprojected visual prototypes using graph convolutional networks [50] in Fig. 3 is formally presented as:

$$\mathbf{U} = \mathtt{ReLU}\left(\widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{V}\mathbf{W}^v\right) \in \mathbb{R}^{|\mathcal{C}|\times D_V}, \tag{5}$$

where $\mathbf{W}^v \in \mathbb{R}^{D_V\times D_V}$ is the learnable graph weights. Here we set the output dimensions of the text and visual features are aligned $D_V = C_T$ for subsequent fusion operations. $\widetilde{\mathbf{D}} = \sum_j \widetilde{\mathbf{A}}_{\mathbf{i},\mathbf{j}}$ is the normalized diagonal matrix. $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I} \in \mathbb{R}^{|\mathcal{C}|\times|\mathcal{C}|}$ denotes the text relationship with self-loop and $\mathbf{I}$ denotes the identity matrix.

Text-Vision Prototypical Fusion: With the graph relation transferring from text features, the updated visual prototypes

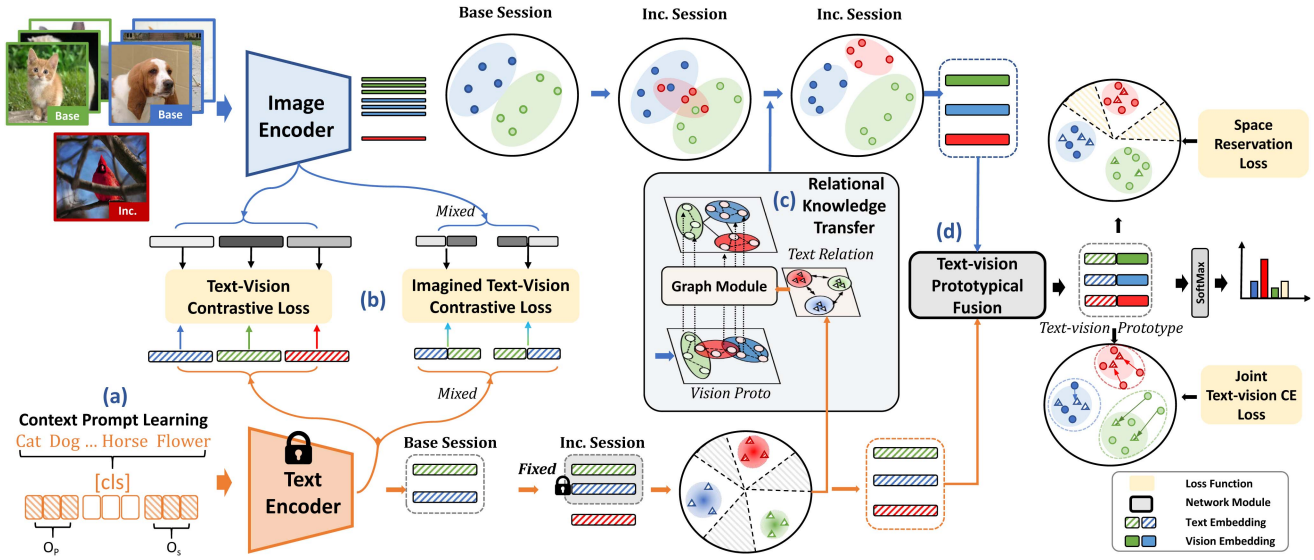

Fig. 3. The proposed Language-inspired Relation Transfer (LRT) approach consists of two essential modules. 1) Relational knowledge transfer module first transfers the text-wise relationship to the visual prototypes and a text-vision prototypical fusion module for knowledge fusion. 2) Image-Text alignment module introduces context prompt learning for fast adaptation and proposes the imagined contrastive learning for multi-modal alignment in few-shot class incremental learning.

$\mathbf{U}$ are reprojected in a topologically distinguishable space for recognition. In this paper, we argue that text prototypes in Fig. 2(b) and visual prototypes in a) are both beneficial for FSCIL tasks, i.e., the text prototypes provide well-generalized representations when there are insufficient training samples, meanwhile, the visual prototypes provide clear visual clues when supervised with sufficient training data. Unlike the predominant image-text prediction methods [2], [25], our model in Fig. 2(c) relies on the joint text-vision prototypes instead of the conventional `fc` layers. Considering the alignment during contrastive learning, we directly fuse the visual prototypes $\mathbf{U} \in \mathbb{R}^{|\mathcal{C}| \times D_V}$ and $\mathbf{T} \in \mathbb{R}^{|\mathcal{C}| \times D_T}$ with a learnable weight τ. Hence for any query image $\mathbf{x}$, the joint similarity scores from (3) are updated as:

$$\hat{\mathbf{p}}_i = \frac{e^{\tau \cdot \xi(\mathcal{V}(\mathbf{x}), \mathbf{T}_i^\top) + \xi(\mathcal{V}(\mathbf{x}), \mathbf{U}_i^\top)}}{\sum_{j=i}^{N} e^{\tau \cdot \xi(\mathcal{V}(\mathbf{x}), \mathbf{T}_j^\top) + \xi(\mathcal{V}(\mathbf{x}), \mathbf{U}_j^\top)}}. \tag{6}$$

This operation can be theoretically replaced by other concatenation or attention-based fusion strategies. Despite its simplicity, we found it works well under different scenarios, which are discussed later. We use these fused text-vision prototypes for both training and inference during base and incremental sessions. Benefiting from this prototypical design, the text knowledge can be taken as a part of visual encoders, and during inference time, we only use the visual backbones without any additional computation costs.

D. Aligning Text With Image in FSCIL

Vanilla contrastive learning adopts the handcrafted text prompt, e.g., '`a photo of a [cls].`' to get language embedding, which is aligned in the same space during the contrastive pretraining. However, during the downstream supervised learning process on visual encoders, it accompanies a clear domain gap between the text and vision embeddings. To solve this, we propose to find the generalized alignment strategy when only texts of labels (e.g., `[cat]`) are available for training.

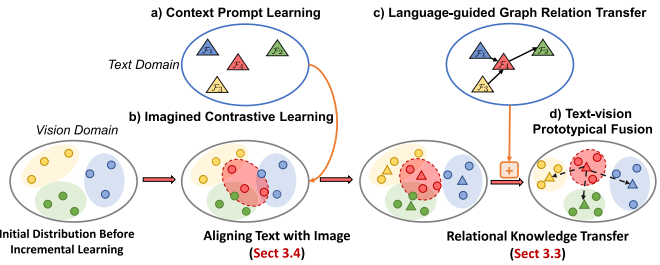

Fig. 4. The motivation and modules of the proposed LRT. Our proposed LRT is composed of an aligning stage to conduct a multimodal alignment with few-shot downstream data and a transferring stage to transfer the text knowledge to the vision domain.

We make two major improvements for this multimodal alignment in FSCIL: 1) during the incremental learning session, we propose a context prompt learning method (Fig. 4(a)) for fast adaptation of pretrained language knowledge on few-shot novel classes; 2) during the base training session, we propose the imagined contrastive learning (Fig. 4(b)) to alleviate the imbalance of multi-modality data (i.e., sufficient visual training data while insufficient text descriptions). Besides, we also propose a space reservation in the base session to construct compact base prototypes to "reserve" space for subsequent incremental prototypes.

Imagined Contrastive Learning: The other aforementioned challenge is caused by the insufficiency of text inputs compared to image data. The contrastive learning in Fig. 5(a) is easy to overfit on training data when only label text is available, e.g., N text phrases for N classes. To alleviate this phenomenon, we introduce a contrastive loss conducted during imagined texts and images in Fig. 5(b), which is composed of two steps. i) the text

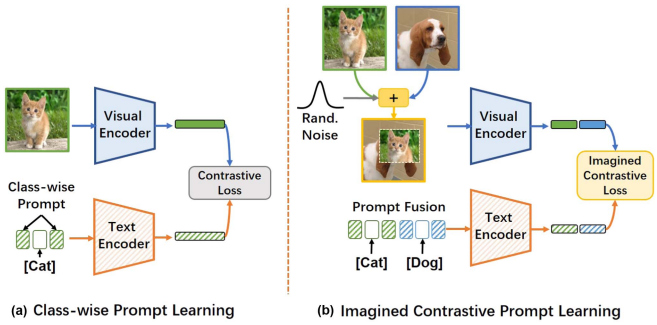

Fig. 5. Illustrations of different text-vision learning loss. (a) Class-wise context prompt Learning. (b) Multi-modality imagined contrastive learning: two images using mixing strategy [51] are aligned with their corresponding prompt fusion texts.

prompts of two random classes are concatenated including the learnable ones:

$$f(\mathbf{t}_i, \mathbf{t}_j) = [\mathbf{o}_i^p, \texttt{cls}_i, \mathbf{o}_i^s, \mathbf{o}_j^p, \texttt{cls}_j, \mathbf{o}_j^s]. \tag{7}$$

ii) two visual images are fused averaged using CutMix [51] or other alternative intra-mixing methods:

$$m(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{M} \cdot \mathbf{x}_i + (\mathbb{1} - \mathbf{M}) \cdot \mathbf{x}_j, \tag{8}$$

where $\mathbf{M} \in \mathbb{R}^{H' \times W'}$ denotes the sampled masks. We control the mask proportions $\frac{H' \times W'}{H \times W}$ of $\mathbf{M}$ are sampled from $(0.4, 0.6)$ to match the text concatenation while introducing randomness. For two mixed samples i, j in batch $\mathcal{B}$, the imagined contrastive learning has the form:

$$\mathcal{L}_{\text{im}}(\mathcal{B}; \mathbf{o}, \theta) = -\sum_{i,j} \log \frac{e^{\xi(\mathcal{V}(m(\mathbf{x}_i, \mathbf{x}_j); \theta), \mathcal{T}(f(\mathbf{t}_i, \mathbf{t}_j; \mathbf{o})))}}{\sum_{p,q \in \mathcal{B}} e^{\xi(\mathcal{V}(m(\mathbf{x}_i, \mathbf{x}_j); \theta), \mathcal{T}(f(\mathbf{t}_p, \mathbf{t}_q; \mathbf{o})))}}. \tag{9}$$

While for each positive pair $m(\mathbf{x}_i, \mathbf{x}_j)$ and $f(\mathbf{t}_i, \mathbf{t}_j)$, the negative samples are other mixtures with different $p \neq i, q \neq j$ in the same batch.

Learning With Space Reservation: Besides the aforementioned contrastive learning, here we introduce the auxiliary margin-based cross-entropy $\mathcal{L}_{\text{M-CE}}$ for space reservation for incremental classes, which helps to project the classes into a normalized hypersphere. Here we introduce the margin-based softmax loss, e.g., ArcFace [52] to help the "space reservation", which alleviates the overfitting on base classes. This indicates that the base classes would be distributed more compactly and would not fulfill the overall manifold space, thus the reserved space can be retained for representing the incremental classes. To be specific, this margin loss has the following form:

$$\mathcal{L}_{\text{M-CE}} = -\log \frac{e^{s \cos(\alpha_{\mathbf{y}_i} + m)}}{e^{s \cos(\alpha_{\mathbf{y}_i} + m)} + \sum_{j=1, j \neq \mathbf{y}_i}^{N} e^{s \cos(\alpha_j)}}, \tag{10}$$

where we empirically set the scale s as 1 with $m = 0.4$ to maintain the magnitude of loss constraints. The $\cos \alpha_j$ denotes the cosine similarity between prototype $\hat{\mathbf{p}}_j$ and visual features $\mathbf{U}_j$.

Context Prompt for Fast Adaptation: During incremental sessions in FSCIL, only $M \leq 5$ samples can be used for training, which makes the visual fine-tuning process a major obstacle.

Algorithm 1: Language-Inspired Relation Transfer (LRT).

Input: Base session dataset $\mathcal{D}^1 = \{(\mathbf{x}_i^1, \mathbf{y}_i^k, \mathbf{t}_i^s)\}_{i=1}^{|\mathcal{D}^1|}$.
Incremental session dataset $\mathcal{D}^1 \ldots \mathcal{D}^S$.
Output: Visual encoder:$\mathcal{V}$, text prompts:$\{\mathbf{o}_i\}_{i=1}^{|\mathcal{C}|}$

Initialize Visual $\mathcal{V}(\cdot)$ and Text encoder $\mathcal{T}(\cdot)$ with CLIP alignment.
▷ *Base Session Pretraining*
Random Init. text prompt $f(\mathbf{t}_i) = [\mathbf{o}_i^p, \texttt{cls}_i, \mathbf{o}_i^s]$
Random Init. visual prototypes $\mathbf{V}$
for $\forall((\mathbf{x}_i^1, \mathbf{y}_i^k, \mathbf{t}_i^1)) \in \mathcal{D}^1$ **do**
  Extract text features: $\mathbf{T}_i = \mathcal{T}(f(\mathbf{t}_i)), i = 1 \ldots |\mathcal{C}^1|$
  Calculate text relationship $\mathbf{A}_{i,j}$ by (4)
  Update visual prototypes $\mathbf{V}$ as $\mathbf{U}$ in (5)
  Fuse text-vision knowledge by (6) to obtain $\hat{\mathbf{p}}$
  Optimize $\mathcal{L}_{\text{CE}}(\hat{\mathbf{p}}, \mathbf{y})$ and $\mathcal{L}_{\text{M-CE}}(\hat{\mathbf{p}}, \mathbf{y})$
  Construct mixed image $\hat{\mathbf{x}}$ by CutMix $m(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{M} \cdot \mathbf{x}_i + (\mathbb{1} - \mathbf{M}) \cdot \mathbf{x}_j$
  Construct mixed text prompt $f(\mathbf{t}_i, \mathbf{t}_j) = [\mathbf{o}_i^p, \texttt{cls}_i, \mathbf{o}_i^s, \mathbf{o}_j^p, \texttt{cls}_j, \mathbf{o}_j^s]$
  Calculate imagined contrastive loss $\mathcal{L}_{\text{im}}^{(i,j)}(\cdot; \mathbf{o}, \theta)$ in (10)
  Conduct base session optimization $\mathcal{L}_{\text{base}}$ in (12) with fixed text encoder $\mathcal{T}$
▷ *Incremental Session Fast Adaptation*
Transfer other learned weights and prompts to incremental session.
Random Init. text prompt $f(\mathbf{t}_i) = [\mathbf{o}_i^p, \texttt{cls}_i, \mathbf{o}_i^s]$ for new classes $\mathcal{C}^s$
for $\forall((\mathbf{x}_i^s, \mathbf{y}_i^k, \mathbf{t}_i^s)) \in \mathcal{D}^s$ **do**
  Init. visual prototypes $\mathbf{V}$ by (2) using $\mathbf{x}$
  Extract text features: $\mathbf{T}_i = \mathcal{T}(f(\mathbf{t}_i)), i = 1 \ldots |\mathcal{C}^s|$
  Conduct incremental session text-vision fast alignment $\mathcal{L}_{\text{inc}} = \mathcal{L}_{\text{conp}}(\cdot; \mathbf{o}_{\{p,s\}})$ in (11)
return Optimized visual encoder $\mathcal{V}$, text prompts $\{\mathbf{o}_i\}_{i=1}^{|\mathcal{C}|}$

To start from another view, as the joint prediction in (6) is also determined by the text embeddings $\mathbf{T} = \mathcal{T}(\mathbf{t})$, we propose to construct class-wise learnable prompt instead of the hand-crafted ones in CLIP. We empirically use prefix $\mathbf{o}^p$ and suffix $\mathbf{o}^s$ learnable prompt for each class, which is accomplished as a whole learnable sentence $f(\mathbf{t}_i) = [\mathbf{o}_i^p, \texttt{cls}_i, \mathbf{o}_i^s]$. As in Fig. 2(c), during the incremental session, we now fix the visual and text encoders and only learnable context prompts for each class are fine-tuned to minimize the language-vision domain gap:

$$\mathcal{L}_{\text{conp}}(\cdot; \mathbf{o}_{\{p,s\}}) = \Sigma_{(\mathbf{x},\mathbf{y},\mathbf{t}) \in \mathcal{D}^s} \mathbf{y} \log \mathcal{S}(\xi(\mathcal{V}(\mathbf{x}), \mathcal{T}(f(\mathbf{t}; \mathbf{o})))), \tag{11}$$

where $\mathcal{S}$ denotes the $\texttt{SoftMax}$ function with pre-learnt weight τ. With the class-wise learnable prompt, few-shot samples can be depicted with learnable sentences, and the domain gap of unseen categories is fast mitigated.

Overall Training Scheme: The overall training follows the few-shot class-incremental learning paradigm. 1) During the base session, the visual prototypes are randomly initialized as

TABLE I
CLASSIFICATION ACCURACY ON *mini*IMAGENET DATASET FOR 5-WAY 5-SHOT INCREMENTAL LEARNING

Method	Pub. Year	Accuracy in each session ↑									Avg.	Δ_{imp}
		1	2	3	4	5	6	7	8	9		
Finetune [15]	-	61.31	27.22	16.37	6.08	2.54	1.56	1.93	2.60	1.40	13.45	(+0.00)
iCaRL [9]*	CVPR 17	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.21	33.28	(+19.84)
Rebalance [53] *	CVPR 19	61.31	47.80	39.31	31.91	25.68	21.35	18.67	17.24	14.17	30.83	(+17.38)
TOPIC [15]	CVPR 20	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42	39.64	(+26.19)
FSLL+SS [19]	CVPR 20	68.85	63.14	59.24	55.23	52.24	49.65	47.74	45.23	43.92	53.92	(+40.47)
IDLVQ-C [20]	ICLR 20	64.77	59.87	55.93	52.62	49.88	47.55	44.83	43.14	41.84	51.16	(+37.71)
CEC [18]	CVPR 21	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	57.75	(+44.30)
F2M [22]	NeurIPS 21	67.28	63.80	60.38	57.06	54.08	51.39	48.82	46.58	44.65	54.89	(+41.44)
MetaFSCIL [54]	CVPR 22	72.04	67.94	63.77	60.29	57.58	55.16	52.90	50.79	49.19	58.85	(+45.40)
LIMIT [45]	TPAMI 22	72.32	68.47	64.30	60.78	57.95	55.07	52.70	50.72	49.19	59.05	(+45.61)
FACT [23]	CVPR 22	72.56	69.63	66.38	62.77	60.60	57.33	54.34	52.16	50.49	60.70	(+47.25)
C-FSCIL [21]	CVPR 22	76.40	71.14	66.46	63.29	60.42	57.46	54.78	53.11	51.41	61.61	(+48.16)
Base-V (FT)	-	72.88	67.65	63.09	59.09	55.54	52.79	49.97	47.87	45.59	57.16	(+43.71)
CLIP (0-shot)	-	65.18	65.05	63.20	62.58	62.49	62.54	61.33	60.98	60.62	62.67	(+49.21)
Ours (LRT)	-	**90.17**	**85.82**	**81.70**	**78.12**	**75.04**	**71.71**	**68.88**	**66.74**	**65.34**	**75.94**	**(+62.49)**

∗: performances reported by [15]. Δ_{imp}: averaged relative improvements across all sessions compared to the finetune baseline.

classifiers and the learning objective is joint optimization of three terms i.e., the cross-entropy $\mathcal{L}_{\text{CE}}$ between the fused image-text prediction and ground truth label, the space reservation constraints $\mathcal{L}_{\text{M-CE}}$, and the imagined contrastive learning $\mathcal{L}_{\text{im}}$:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{CE}}(\hat{\mathbf{p}}, \mathbf{y}) + \lambda_{\text{m}}\mathcal{L}_{\text{M-CE}}(\hat{\mathbf{p}}, \mathbf{y}) + \lambda_{\text{im}}\mathcal{L}_{\text{im}}(\mathbf{x}, \mathbf{t}). \quad (12)$$

The relational knowledge transfer module in Section III-C is consistent during the base and incremental sessions. These three loss functions are jointly optimized during the base session to construct a generalized text-to-image feature transferring space. 2) While in the incremental learning session, we first construct vision prototypes following (2) and then finetune the $\mathcal{L}_{\text{inc}} = \mathcal{L}_{\text{conp}}(\cdot; \mathbf{o}_{\{p,s\}})$ to fast mitigate the domain gap among text and vision while representing the visual samples with learnable text prompts. Note that we only conduct the prompt learning $\mathcal{L}_{\text{conp}}$ during incremental sessions, alleviating the overfitting of multi-modal feature space caused by extreme few-shot samples. The detailed training algorithm is shown in Algorithm 1, which adopts different training strategies during the base training session and incremental training session. As there are only a few visual samples for training, we only conduct the fast adaptation with a few learnable text prompts of the current training classes, while fixing the prompts of previously seen sessions.

For inference time, we first use learned text prompts $\{\mathbf{o}_i\}_{i=1}^{|\mathcal{C}|}$ to update text prototypes and thus drop the heavy text encoder for inference, keeping other modules including knowledge transfer module consistent with the training phase. The final prediction is measured by fused text-vision prototypes in (6).

IV. EXPERIMENTS

A. Experimental Settings

Dataset and Evaluations: In this experiment, following the splits in prevailing works [15], [18], [21], we mainly conduct ablations on two widely-used benchmark datasets, i.e., *mini*ImageNet [27] and CIFAR-100 dataset [28]. *mini*ImageNet [27] contains 100 different semantic classes, which are divided into 60 base classes and 40 few-shot classes for 8 incremental sessions. In the base sessions, each class has 600 images with a resolution of 84×84, while in the few-shot session, only 5 images of each class are used for training. Besides, we conduct experiments on the large-scale ImageNet100 [27] dataset of over 128k images following [23] with the image resolution of 224×224. Similarly, ImageNet100 contains 100 different semantic classes, which are divided into 60 base classes and 40 few-shot classes for 8 incremental sessions. Besides, the CIFAR-100 dataset [28] is also divided into 60 base classes and 40 few-shot incremental classes, with the resolution of 32×32. The final evaluations are conducted on classes across all training sessions.

Implementation Details: To conduct fair comparisons with state-of-the-art works, we follow [18] to conduct the supervised training with identical data augmentation strategies. We adopt the lightweight ResNet-50 [1] model pretrained by CLIP [2] to alleviate the additional parameters. The text prompt is set as '`a photo of a [cls].`' for fair comparisons with prevailing works. Following [18], The model is trained with the batch size of 128 with SGD for 100 epochs. The learning rate starts at 0.01 for both *mini*ImageNet [27] and CIFAR-100 dataset [28] and decays at 40 and 70 epochs. The text encoders are fixed across all the sessions, and the learnable prompt length is set as 4. Balanced weights λ_m, λ_{im} are set as 0.1 and 0.05 respectively. We resize the low-resolution image (84×84) to fit the positional encoding layer of CLIP models.

B. Comparison With State-of-the-Art

Results on miniImageNet: In Table II, we first conduct experiments on the widely-used challenging *mini*ImageNet dataset with state-of-the-art works, including several CIL methods [9], [53] and FSCIL methods [15], [18], [19], [21], [54]. Pioneer works [15] indicate learning with a naive finetuning strategy in the first line would lead to catastrophic forgetting on base sessions, while the CIL methods alleviate this difficulty by clear improvements. To validate the effectiveness of our method, we conduct a baseline finetuning only using the visual encoders (ResNet-CLIP) using the identical protocol (*Base-V*) with

TABLE II
CLASSIFICATION ACCURACY ON CIFAR100 DATASET FOR 5-WAY 5-SHOT INCREMENTAL LEARNING

Method	Pub. Year	Accuracy in each session ↑									Avg.	Δ_{imp}
		1	2	3	4	5	6	7	8	9		
Finetune [15]	-	64.10	39.61	15.37	9.80	6.67	3.80	3.70	3.14	2.65	16.53	(+0.00)
iCaRL [9]*	CVPR 17	64.10	53.28	41.69	34.13	27.93	25.06	20.41	15.48	13.73	32.87	(+16.33)
Rebalance [53] *	CVPR 19	64.10	53.05	43.96	36.97	31.61	26.73	21.23	16.78	13.54	34.21	(+17.68)
TOPIC [15]	CVPR 20	64.10	55.88	47.07	45.16	40.11	36.38	33.96	31.55	29.37	42.62	(+26.08)
FSLL+SS [19]	CVPR 20	66.76	55.52	52.20	49.17	46.23	44.64	43.07	41.20	39.57	48.71	(+32.17)
CEC [18]	CVPR 21	73.07	68.88	65.26	61.19	58.09	55.57	53.22	51.34	49.14	59.53	(+42.99)
F2M [22]	NeurIPS 21	61.71	62.05	59.01	55.58	52.55	49.96	48.08	46.28	44.67	53.32	(+36.78)
DSN [55]	TPAMI 22	73.00	68.83	64.82	62.24	59.16	56.96	54.04	51.57	49.35	60.00	(+43.47)
MetaFSCIL [54]	CVPR 22	74.50	70.10	66.84	62.77	59.48	56.52	54.36	52.56	49.97	60.79	(+44.25)
C-FSCIL [21]	CVPR 22	77.47	72.40	67.47	63.25	59.84	56.95	54.42	52.47	50.47	61.64	(+45.10)
LIMIT [45]	TPAMI 22	73.81	72.09	67.87	63.89	60.70	57.78	55.68	53.56	51.23	61.84	(+45.31)
FACT [23]	CVPR 22	74.60	72.09	67.56	63.52	61.38	58.36	56.28	54.24	52.10	62.23	(+45.70)
Base-V (FT)	-	67.37	62.37	58.00	54.27	51.11	48.32	45.71	43.57	41.50	52.47	(+35.93)
CLIP (0-shot)	-	39.78	38.25	36.97	34.32	33.26	32.07	31.99	31.24	30.24	34.24	(+17.71)
Ours (LRT)	-	**87.02**	**82.40**	**77.84**	**73.31**	**70.18**	**66.74**	**64.50**	**61.99**	**59.49**	**71.50**	**(+54.96)**

*: performances reported by [15]. Δ_{imp}: averaged relative improvements across all sessions compared to the finetune baseline.

TABLE III
COMPARISONS ON IMAGENET100 DATASET FOR 5-WAY 5-SHOT INCREMENTAL LEARNING

Method	Accuracy in ImageNet100 ↑									Avg.
	1	2	3	4	5	6	7	8	9	
CEC [18]*	84.77	80.03	76.66	73.10	69.30	65.88	64.27	62.91	60.04	70.77
FACT [23]*	86.00	80.94	77.66	75.34	70.40	66.72	64.82	63.15	60.98	71.67
Ours (LRT)	**91.43**	**87.03**	**83.83**	**79.34**	**76.15**	**72.05**	**70.18**	**68.52**	**65.90**	**77.16**

*: Methods are implemented using official codes.

prototypical learning of Fig. 2(a) in the incremental session, which shows slightly higher performance than earlier models. The zero-shot CLIP in Fig. 2(b) have a strong generalization ability and do not need any training data. The last session's accuracy of zero-shot CLIP remains at high accuracy. With our proposed Language-inspired Relation Transfer (LRT) model, the performance of the last session is improved by 19.7%, and also shows a clear margin i.e., 13.9% and 14.8% compared to the state-of-the-art C-FSCIL [21] and FACT [23] methods.

Results on CIFAR100: Note that CIFAR-100 is a low-resolution image recognition dataset and CLIP models face difficulties in conducting zero-shot testing. The first session shows about a 25% gap in accuracy and Base-V with fine-tuning surpasses the zero-shot CLIP models by 18% in the average accuracy. Similar to the performances on *mini*ImageNet, our model also shows a notable improvement on the CIFAR100 dataset. Existing models including CEC [18], FACT [23] with fixed or slightly tuned encoders (Fig. 2(a)) shows a clear performance improvement compared to the topological re-adjusting ones [15]. As this prevailing trend shows a performance bottleneck, our LRT steadily improves the performances of baseline finetuning (*Base-V*) by nearly 18.0% and surpasses the second best method [23] by 7.3% in the last session.

Results on ImageNet100: To verify the performance on large-scale datasets, we compare our proposed method with the source code of state-of-the-art methods CEC [18], FACT [23] and adopt the same augmentation from [23]. As in Table III FACT [23] improves the base and incremental learning sessions by 0.9%, while our proposed method (LRT) shows steady improvements and surpasses FACT [23] by a clear margin of 5.49%, implying the good potential of our model for large-scale datasets.

C. Performance Analysis

Ablations of Learning Paradigms: In Table IV, we conduct detailed ablations of our proposed learning paradigms. The first two lines show methods only using visual FineTuning (FT) and text Prompt Tuning (PT) respectively, which show inferior results on both base and incremental sessions. With the addition of the knowledge fusion module in the third line, the model shows a clear performance improvement, e.g., over 9.0% on the last session of CIFAR compared to the Base-V. The fourth and fifth line introduces the joint PT-visual tuning and the imagined contrastive loss, which presents steady improvements on both last session accuracy and average accuracy. The final line is our full model which can still boost the performance by using graph transformations (over 18% than visual baselines).

Effects of Text-Vision Relationship: We first conduct a visualized ablations with the Imagined Contrastive Training Loss of (12). In Fig. 6(a), we exhibit results that adopt the standard cross entropy $\mathcal{L}_{\mathrm{CE}}(\mathrm{V})$ for only using vision prototypes, and (6) and (12) denote the only image-text fusion and fusion with imagined contrastive loss. The upper left diagonal shows the last 10 base classes in the miniImageNet dataset, and the lower shows the confusion matrices on 10 new classes. Comparing Fig. 6(a) with (b), it can be found our proposed alignment strategies greatly improve the learning of new classes without forgetting the base knowledge.

TABLE IV
ABLATION STUDIES ON *mini*IMAGENET AND CIFAR100 BENCHMARKS

Text	Vision	T-V Trans	T-V Loss	*mini*ImageNet Acc ($\mathcal{D}^S$)	*mini*ImageNet Avg. Acc	*mini*ImageNet Δ_{avg}	CIFAR100 Acc ($\mathcal{D}^S$)	CIFAR100 Avg. Acc	CIFAR100 Δ_{avg}
-	FT	Only V	$\mathcal{L}_{\text{CE}}(\text{V})$	45.59	57.16	(+0.00)	41.50	52.47	(+0.00)
PT	Fixed	Only T	$\mathcal{L}_{\text{CE}}$(Eqn.(3))	42.54	52.87	(−4.30)	38.83	49.86	(−2.61)
Fixed	FT	$\mathcal{M}_{fuse}$	$\mathcal{L}_{\text{CE}}$(Eqn.(6))	50.60	64.98	(+7.81)	50.74	64.77	(+12.31)
PT	FT	$\mathcal{M}_{fuse}$	$\mathcal{L}_{\text{CE}}$(Eqn.(6))	54.33	67.01	(+9.85)	54.69	67.32	(+14.84)
PT	FT	$\mathcal{M}_{fuse}$	Eqn.(12)	56.00	67.88	(+10.72)	55.18	67.67	(+15.20)
PT	FT	$\mathcal{M}_{fuse}$+$\mathcal{M}_{graph}$	Eqn.(12)	**65.34**	**75.94**	**(+18.78)**	**59.49**	**71.50**	**(+19.03)**

Pt: learnable context prompt finetuning. Ft: standard visual finetuning. $\mathcal{M}_{fuse}$, $\mathcal{M}_{graph}$: the text-vision prototypical fusion and graph relationship transformation. Acc ($\mathcal{D}^S$): accuracy of the last session. Avg. Acc: averaged accuracy of all sessions. Δ_{avg} : relative improvements.

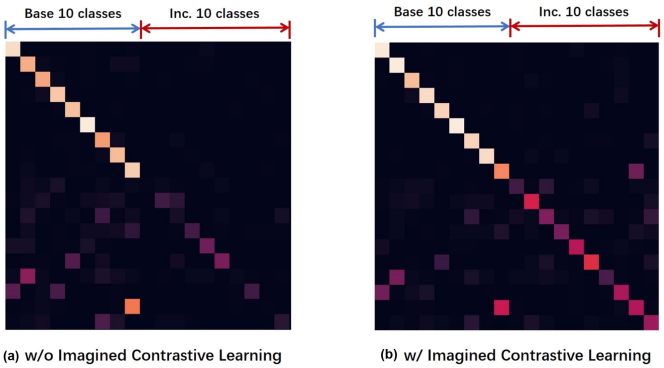

Fig. 6. Confusion matrices of different contrastive losses. The last 10 base classes and the first 10 incremental classes on *mini*ImageNet are zoomed in for ablation comparisons.

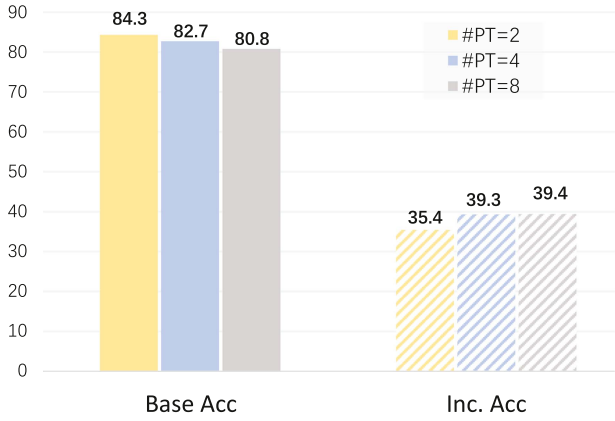

Fig. 7. Accuracies of base and incremental sessions with different learnable prompt lengths. Our method chooses prompt length #PT = 4 for the base and incremental trade-off.

TABLE V
PERFORMANCE ANALYSIS OF DIFFERENT TEXT-VISION (T → I) METHODS AND PROTOTYPICAL FUSION STRATEGIES ON *mini*IMAGENET

T → I Methods	Strategy	Acc ($\mathcal{D}^S$)	Avg. Acc	Δ_{avg}
Baseline (w/o T)	-	45.59	57.16	(+0.00)
$\mathcal{M}_{graph} + \mathcal{M}_{fuse}$	Proto Add.	42.74	51.25	(−5.91)
$\mathcal{M}_{graph} + \mathcal{M}_{fuse}$	Static	45.46	51.10	(−6.06)
$\mathcal{M}_{fuse}$	Learnable	56.00	67.88	(+10.72)
$\mathcal{M}_{graph}$	Learnable	60.30	72.61	(+15.45)
$\mathcal{M}_{graph} + \mathcal{M}_{fuse}$	Learnable	65.34	75.94	(+18.78)

Proto add. : Averaged summation of vision and text prototypes. Static: Learnable τ in (6) is set as 1.

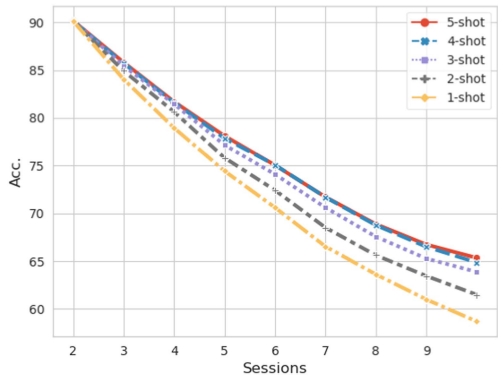

Fig. 8. Accuracy with different number of shots during incremental sessions on *mini*ImageNet dataset.

Besides visualization of loss functions, the fusion strategies are also important in our method, here we present several naive implementations in Table V. The first line shows the baseline visual fine-tuning model without the help of text information. The *Learnable* denotes our fusion methods using (6) and the *Static* denotes a fixed $\tau = 1$. The *Proto Add:* denotes we directly add two prototypes before measurement, i.e., $\xi(\mathbf{x}, \mathbf{V} + \mathbf{T})$. With the proposed graph module, the overall accuracy of base and incremental sessions shows a clear improvement. The experimental evidence shows the performance drops dramatically without the proper fusion of text and vision embedding since they are strongly aligned during the pretraining stage.

Effects of Context Prompt: Selecting the proper size of the prompt length is one of the factors that affect the final performance. We conduct hyper-parameter ablations by setting the prompt length (both prefix and suffix, #PT) as 2,4 and 8 respectively. Fig. 7 shows that using more learnable prompts can boost the incremental learning session with only a few given samples, e.g., 39.3% versus 35.4%. Nevertheless, finetuning more prompts will lead to the loss of base knowledge, e.g., 1.6% when extending the prompt number from 2 to 4. To achieve a good trade-off between the base and incremental sessions, we chose the prompt length of 4 in all our experiments.

Incremental Learning With Fewer Shots: Besides the exploration of common N-way 5-shot settings during incremental learning. Here we exhibit the results on fewer shots in Fig. 8, i.e.,

TABLE VI
PERFORMANCE ANALYSIS OF DIFFERENT INCREMENTAL SHOTS ON *MINI*IMAGENET

Incremental Shots	Base. Acc	Inc. Acc	Avg. Acc	Δ_{avg}
5-shot	82.68	**39.32**	**61.00**	(+0.00)
4-shot	83.15	37.42	60.28	(−0.72)
3-shot	83.77	34.02	58.90	(−2.10)
2-shot	84.22	27.40	55.81	(−5.19)
1-shot	**86.40**	17.12	51.76	(−9.24)

Δ_{avg} : relative improvements of averaged accuracy.

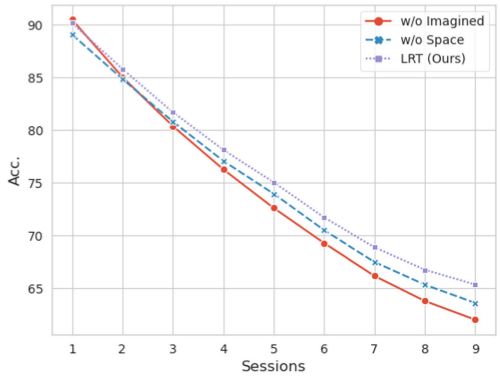

Fig. 9. Ablations of different loss functions. Imagined: imagined contrastive learning $\mathcal{L}_{\text{im}}$. Space: space reservation loss $\mathcal{L}_{\text{M-CE}}$.

from 1-shot to 4-shots. Starting from the same base accuracy of nearly 90%, the model with fewer shots performs a more notable performance drop than ours with 5 training shots. Whilst it is still acceptable (over 51% in Avg. Acc) compared to other earlier methods. However, when considering the base accuracy and incremental accuracy individually in Table VI, the performance of incremental sessions drops significantly. With the decrease of training shots, the accuracy on base sessions seems to retain the pretrained representations in base sessions, which shows higher performance (86.40% of 1-shot compared to 82.68% of 5-shots). But the accuracy of incremental sessions drops dramatically. This indicates that it is hard to align the visual concepts with proper text descriptions with only one sample. In other words, depicting objects with pretrained textual knowledge also needs more visual cases to alleviate overfitting.

Effect of Loss Constraints: We conduct experiments on different gratitude of hyperparameters i.e., λ_{im} for $\mathcal{L}_{\text{im}}$ and λ_{m} for $\mathcal{L}_{\text{M-CE}}$. The experimental results can be found in Table VIII. Enlarging or reducing the balanced weight would lead to a clear performance drop of 1.8% to 4.05%. Besides, only scaling up the $\mathcal{L}_{\text{M-CE}}$ constraints would lead to a clear base session drop, while other imagined contrastive learning and space reservation constraints mainly affect the ability to learn new concepts. The detailed ablations of $\mathcal{L}_{\text{im}}$ and $\mathcal{L}_{\text{M-CE}}$ in (12) is presented in Fig. 9. With the collaborative learning of these loss functions, the capability to learn novel concepts has been greatly enhanced.

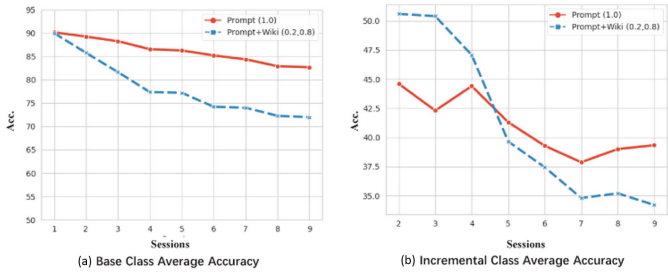

Fig. 10. Accuracies of base and the average of our methods and extensions on Wiki data [56]. Using Wiki data helps the fast understanding of incremental sessions in b), while leading to performance drops on the base sessions in a).

Understanding Objects From Wiki: One ideal scenario for understanding novel concepts is learning from descriptions from sufficient web data, which contains rich descriptions like "*The northern cardinal has a distinctive crest on the head...*". To achieve this, we collect the first 5 sentences from the Wikipedia articles corresponding to ImageNet categories provided by [56]. With this prior knowledge, we fuse the wiki data and our prompt with a ratio of $(8:2)$ with other settings identical. Fig. 10 exhibits the results using wiki data (blue dotted line) and our final model (red line). It can be found that the wiki data do provide rich knowledge for incremental classes (e.g., over 8% in session 2) with few learnable samples as in Fig. 10(b). However, as prompt learning takes a less important place during this learning, the base classes are less discriminative with the incremental sessions, which leads to clear drops during the base sessions in Fig. 10(a). We would like to leave this extension in our future work by incorporating advanced prompt fusion strategies when more text data are available.

D. Discussions

How Does LRT Help FSCIL? As aforementioned, the major challenge in few-shot class-incremental learning is to alleviate the forgetting of base classes while recognizing novel incremental classes. Some recent research [44] indicates that maintaining the base session performance but inferior incremental session performance would also lead to higher results, which are caused by the imbalanced number of classes in base and incremental sessions. Here we conduct detailed comparisons with two state-of-the-art methods i.e., CEC [18] and FACT [23] in Table IX. The *Base Acc.* denotes the averaged base class accuracy after the final incremental session, and similarly, we define the *Inc. Acc.* for all incremental classes. We then calculate the harmonic average accuracy (*Har. Acc.*) of incremental and base accuracy. Our proposed method achieves over 22% improvements in the incremental accuracy on *mini*ImageNet dataset. Moreover, increasing the length of text prompts (from *2PT* to *4PT*) leads to a slight performance drop (1.6%) on base classes, while boosting the incremental accuracy for over 3.9%.

Besides the comparison with recent methods, the other natural concern is: *why our proposed LRT improves the representation of incremental knowledge?* Keeping this in mind, we conduct

TABLE VII
KNOWLEDGE TRANSFER VERIFICATION ON *mini*IMAGENET BENCHMARKS

Text	Vision	T-V Trans	T-V Loss	*mini*ImageNet Base Acc.	*mini*ImageNet Inc. Acc.	*mini*ImageNet Har. Acc.	Δ_{har}
PT	FT	$\mathcal{M}_{fuse}$	Eqn.(6)	**83.78**	10.15	18.11	(+0.00)
PT	FT	$\mathcal{M}_{fuse}$	Eqn.(12)	76.97	24.55	37.07	(+18.96)
PT	FT	$\mathcal{M}_{fuse}$+$\mathcal{M}_{graph}$	Eqn.(12)	82.68	**39.32**	**53.29**	(+35.18)

Pt: learnable context prompt finetuning. Ft: standard visual finetuning. $\mathcal{M}_{fuse}$, $\mathcal{M}_{graph}$: the text-vision prototypical fusion and graph relationship transformation. Δ_{har} : relative improvements of harmonic Accuracy.

TABLE VIII
PERFORMANCE ANALYSIS OF BALANCED WEIGHTS λ_{M} AND λ_{IM} IN (12)

λ_{im}	λ_{m}	Acc ($\mathcal{D}^1$)	Acc ($\mathcal{D}^S$)	Avg. Acc	Δ_{avg}
0.1×	1×	89.93	61.28	73.49	(−2.45)
1×	0.1×	90.05	62.74	74.13	(−1.81)
1×	10×	87.63	61.13	71.89	(−4.05)
10×	1×	90.02	62.27	73.78	(−2.16)
1×	1×	90.17	65.34	75.94	(+0.0)

TABLE IX
DIFFERENT MEASUREMENT COMPARISONS ON PUBLIC BENCHMARKS

Dataset	**Method**	($\mathcal{D}^1$)	($\mathcal{D}^S$)	Base Acc.	Inc Acc.	Har. Acc.
mini-IN	CEC [18]	72.25	47.67	67.97	17.23	27.49
	FACT [23]	75.23	48.61	72.62	12.60	21.47
	Ours-2PT	89.95	64.74	**84.28**	35.42	49.88
	Ours-4PT	**90.17**	**65.34**	82.68	**39.32**	**53.29**
CIFAR	CEC [18]	73.07	49.10	67.90	20.90	31.96
	FACT [23]	78.65	51.19	71.02	21.45	32.94
	Ours-4PT	**87.02**	**59.49**	**78.68**	**30.70**	**44.17**

Models are evaluated using public codes.

ablations to verify how the knowledge transfer improves the learning of incremental sessions, as in Table VII. The results indicate that our proposed imagined image-text contrastive loss greatly improves the alignment of text and image domains and thus improves the incremental accuracy, i.e., from 10.15% to 24.55%. With our proposed graph module $\mathcal{M}_{graph}$ in the last line, the incremental accuracy can be improved to 39.32%, which results in a performance margin compared to the prevailing works.

Do the Performance Improvements Mainly Benefited From the CLIP Pretraining? There is no doubt that our proposed model benefits from the language-vision pretraining [2], which could lead to a performance boost even with the naive training scheme. We thus conduct detailed comparisons with the standard baseline and our base models (*Base-V*). The standard training baseline is modified from the decoupled prototype learning in [18]. The major difference between these two training procedures is the multi-modality pretraining by CLIP [2]. As in Table X, the pretrained CLIP model provides a performance gain in both base session and incremental session with an average of 4% on the *mini*ImageNet dataset. The CEC models surpass the baseline visual tuning models with an average of over 1.85% and FACT achieves an average accuracy of 70.97%. Under different circumstances, our proposed LRT is able to achieve steady improvements on both base and incremental sessions by a clear margin, which indicates our performance improvements do not mainly come from the strong pretraining models but develop the potential of multimodal knowledge transfer.

Can Models Learn From Long-Term Incremental Sessions? We conduct detailed experimental comparisons with CEC [18] and FACT [23] with the replaced CLIP backbones in Table XI. We split the *mini*ImageNet dataset into two parts (base session for 60 classes and incremental session for the rest 40 classes). The incremental stage consists of 20 sessions, where each session has 2 classes × 5 samples. The parameters of the network are fixed in CEC and FACT during the incremental session, which makes these methods show less forgetting in the long-term incremental learning but limits their crucial ability to learn novel concepts. Even under this setting, our proposed LRT still shows preferable performance improvements, i.e., 12.54% compared to CEC [18]. This is mainly because the incremental relationship is pre-learned in the textual encoders of CLIP models, and during the incremental sessions, its relationship is basically stable and does not suffer from catastrophic forgetting.

How Does Text Knowledge Help the Joint Embedding? The crucial idea of our multi-modal learning paradigm is to transfer the text domain knowledge to the image domains. Here we visualize the t-SNE results of the final prediction scores on *mini*ImageNet dataset. To present clear results, we only select the last 5 base classes (54∼59) and the first 5 incremental classes (60∼64) in the same space. We select the first 10 images of each class in Fig. 11. With the only learnable visual prototypes in Fig. 11(a), although the base classes can be distinguished before the incremental stages after continuous learning the incremental classes are entangled with the base classes in the feature embedding. While in Fig. 11(b), measurements using the text features show a clear distribution with little confusion. By using the joint measurement of text and image prototypes, final results in Fig. 11(c) show clear decision boundaries of each class clusters.

E. Limitations and Future Works

As the text domain also retains rich knowledge for understanding the object, simply using text prompts with few-shot visual samples still leads to insufficient representations, i.e., about 40% accuracy with 5-shot learning on *mini*ImageNet. It is caused by that only the class token [CLS] is used for multi-modal alignment. Other methods [2] indicate that using

TABLE X
COMPARISONS OF STANDARD TRAINING AND FINETUNING WITH CLIP VISUAL BACKBONES ON *mini*IMAGENET DATASET FOR 5-WAY 5-SHOT INCREMENTAL LEARNING

Method	Backbone	Accuracy in mini-ImageNet ↑									Avg.
		1	2	3	4	5	6	7	8	9	
Standard Base	ResNet-18	70.87	65.71	61.66	58.51	55.49	52.68	50.07	48.08	46.64	56.63 (+0.00)
CEC [18]	CLIP [2]	77.40	71.94	67.91	64.69	61.54	58.40	55.46	53.45	52.18	62.55 (+5.92)
FACT [23]	CLIP [2]	85.70	80.49	76.11	72.40	68.83	65.55	62.39	60.52	58.66	70.07 (+13.44)
Base-V (Ours)	CLIP [2]	72.56	69.63	66.38	62.77	60.60	57.33	54.34	52.16	50.49	60.70 (+4.03)
Ours (LRT)	CLIP [2]	**90.17**	**85.82**	**81.70**	**78.12**	**75.04**	**71.71**	**68.88**	**66.74**	**65.34**	**75.94** (+19.31)

Standard base: visual learning baseline proposed in section III-A.

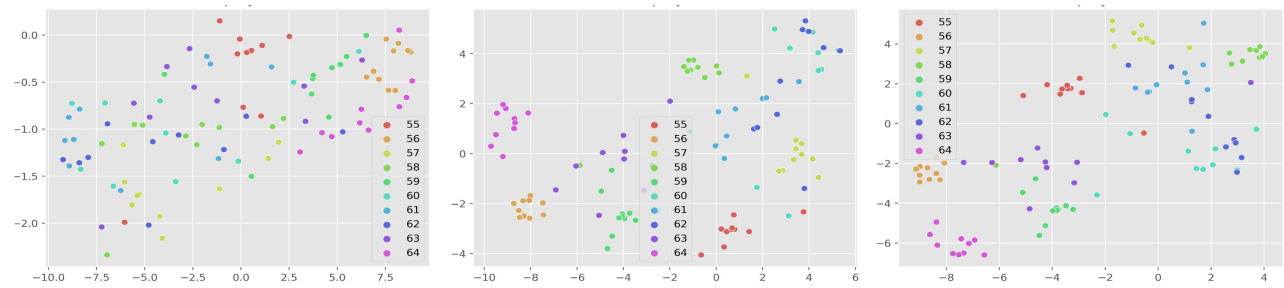

Fig. 11. T-SNE visualizations of prediction scores on *mini*ImageNet dataset. (a) Measurement only using visual prototypes. (b) Measurement only using text prototypes. (c) Our proposed joint fusion strategy with text-image measurements.

TABLE XI
LONG-TERM INCREMENTAL LEARNING (21 SESSIONS) ON ON *mini*IMAGENET FOR 5-WAY 5-SHOT CLASSIFICATION

Methods	Acc ($\mathcal{D}^1$)	Acc ($\mathcal{D}^{11}$)	Acc ($\mathcal{D}^{21}$)	Avg. Acc	Δ_{avg}
CEC-CLIP [18]	77.40	61.53	52.18	62.40	(+0.00)
FACT-CLIP [23]	85.70	68.83	58.66	69.94	(+7.54)
Ours (LRT)	90.13	73.41	63.64	74.64	(+12.24)

Proto add. : Averaged summation of vision and text prototypes. Static: Learnable τ in (6) is set as 1.

rich hand-crafted prompts may lead to higher performances, including "a good photo of [CLS]". Besides, compared to the visual representations with sufficient training samples, using text embedding as prototypes also lead to performance bottleneck, which might be caused by the insufficient description of local visual patterns.

One possible solution to solve this limitation is to design a dynamic bi-directional learning strategy for visual and text representations. When sufficient training samples are available (e.g., ImageNet), there should be also a re-adjustment of text embedding. In other words, we have only explored the data flow from $\mathtt{T} \rightarrow \mathtt{I}$ in this work, while the $\mathtt{I} \rightarrow \mathtt{T}$ relations are not fully discovered, which is also a promising direction for many downstream tasks.

V. CONCLUSION

In this paper, we make attempts to explore the few-shot class-incremental learning problem from a novel perspective by introducing generalized pertaining language knowledge as learning guidance. To achieve this, our approach proposes a new language-guided relation transfer module and a text-vision prototypical fusion module for joint text-vision representations. Beyond that, to align text with image data in FSCIL, we introduce context prompt learning for fast adaptation during training and an imagined contrastive loss to alleviate the data insufficiency during multi-modal alignment. Experimental results demonstrate that our proposed method surpasses the conventional single-modal methods by a large margin on benchmark datasets.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[2] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[3] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-SGD: Learning to learn quickly for few-shot learning," 2017, *arXiv: 1707.09835*.
[4] D. Rezende et al., "One-shot generalization in deep generative models," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1521–1529.
[5] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1842–1850.
[6] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4080–4090.
[7] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.
[8] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3637–3645.
[9] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2001–2010.
[10] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," 2018, *arXiv: 1812.00420*.

[11] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, "Experience replay for continual learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 32.
[12] J. Zhang et al., "Class-incremental learning via deep model consolidation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1131–1140.
[13] A. Mallya and S. Lazebnik, "PackNet: Adding multiple tasks to a single network by iterative pruning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7765–7773.
[14] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.
[15] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong, "Few-shot class-incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12 183–12 192.
[16] A. Cheraghian, S. Rahman, P. Fang, S. K. Roy, L. Petersson, and M. Harandi, "Semantic-aware knowledge distillation for few-shot class-incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2534–2543.
[17] S. Dong, X. Hong, X. Tao, X. Chang, X. Wei, and Y. Gong, "Few-shot class-incremental learning via relation knowledge distillation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1255–1263.
[18] C. Zhang, N. Song, G. Lin, Y. Zheng, P. Pan, and Y. Xu, "Few-shot incremental learning with continually evolved classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12 455–12 464.
[19] P. Mazumder, P. Singh, and P. Rai, "Few-shot lifelong learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2337–2345.
[20] K. Chen and C.-G. Lee, "Incremental few-shot learning via vector quantization in deep embedded space," in *Proc. Int. Conf. Learn. Representations*, 2020.
[21] M. Hersche, G. Karunaratne, G. Cherubini, L. Benini, A. Sebastian, and A. Rahimi, "Constrained few-shot class-incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9057–9067.
[22] G. Shi, J. Chen, W. Zhang, L.-M. Zhan, and X.-M. Wu, "Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 6747–6761.
[23] D.-W. Zhou, F.-Y. Wang, H.-J. Ye, L. Ma, S. Pu, and D.-C. Zhan, "Forward compatible few-shot class-incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9046–9056.
[24] L. H. Li et al., "Grounded language-image pre-training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10 965–10 975.
[25] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.
[26] M. Wortsman et al., "Robust fine-tuning of zero-shot models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7959–7971.
[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[28] A. Krizhevsky et al., "Learning multiple layers of features from tiny images," Master's thesis, Univ. Tront, 2009.
[29] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. Int. Conf. Learn. Representations*, 2017.
[30] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.
[31] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," 2018, *arXiv: 1803.02999*.
[32] E. Triantafillou, R. S. Zemel, and R. Urtasun, "Few-shot learning through an information retrieval lens," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 2255–2265.
[33] M. Ren et al., "Meta-learning for semi-supervised few-shot classification," in *Proc. Int. Conf. Learn. Representations*, 2018.
[34] B. N. Oreshkin, P. Rodriguez, and A. Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 719–729.
[35] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8012–8021.
[36] C. Zhang, Y. Cai, G. Lin, and C. Shen, "DeepEMD: Few-shot image classification with differentiable Earth mover's distance and structured classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12 203–12 213.
[37] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Netw.*, vol. 113, pp. 54–71, 2019.
[38] E. Belouadah and A. Popescu, "IL2M: Class incremental learning with dual memory," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 583–592.
[39] X. Hu, K. Tang, C. Miao, X.-S. Hua, and H. Zhang, "Distilling causal effect of data in class-incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3957–3966.
[40] X. Liu, M. Masana, L. Herranz, J. Van de Weijer, A. M. Lopez, and A. D. Bagdanov, "Rotate your networks: Better weight consolidation and less catastrophic forgetting," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 2262–2268.
[41] X. Tao, X. Chang, X. Hong, X. Wei, and Y. Gong, "Topology-preserving class-incremental learning," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 254–270.
[42] H. Zhao, Y. Fu, M. Kang, Q. Tian, F. Wu, and X. Li, "MgSvF: Multi-grained slow vs. fast framework for few-shot class-incremental learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1576–1588, Mar. 2024.
[43] Y. Zou, S. Zhang, Y. Li, and R. Li, "Margin-based few-shot class-incremental learning with class-level overfitting mitigation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, Art. no. 1977.
[44] C. Peng, K. Zhao, T. Wang, M. Li, and B. C. Lovell, "Few-shot class-incremental learning from an open-set perspective," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 382–397.
[45] D.-W. Zhou, H.-J. Ye, L. Ma, D. Xie, S. Pu, and D.-C. Zhan, "Few-shot class-incremental learning by sampling multi-phase tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12816–12831, Nov. 2023.
[46] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.
[47] S. Goel, H. Bansal, S. Bhatia, R. Rossi, V. Vinay, and A. Grover, "CYCLIP: Cyclic contrastive language-image pretraining," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 6704–6719.
[48] Y. Li et al., "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," in *Proc. Int. Conf. Learn. Representations*, 2021.
[49] Z. Wang et al., "Learning to prompt for continual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 139–149.
[50] M. Welling and T. N. Kipf, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2016.
[51] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6023–6032.
[52] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.
[53] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 831–839.
[54] Z. Chi, L. Gu, H. Liu, Y. Wang, Y. Yu, and J. Tang, "MetaFSCIL: A meta-learning approach for few-shot class incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14 166–14 175.
[55] B. Yang et al., "Dynamic support network for few-shot class incremental learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2945–2951, Mar. 2023.
[56] S. Bujwid and J. Sullivan, "Large-scale zero-shot image classification from rich and diverse textual descriptions," in *Proc. 3rd Workshop Beyond Vis. Lang. Integrating Real-World Knowl.*, 2021, pp. 38–52.

Yifan Zhao (Member, IEEE) received the BE degree from the Harbin Institute of Technology, in 2016, and the PhD degree from the School of Computer Science and Engineering, Beihang University, in 2021. He is currently an associated professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. He worked as a Boya Postdoc researcher with the School of Computer Science, Peking University. His research interests include computer vision and image/video understanding.

Jia Li (Senior Member, IEEE) received the BE degree from Tsinghua University, in 2005, and the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2011. He is currently a full professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. He is the author or coauthor of more than 100 technical articles in refereed journals and conferences, such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Computer Vision*, *IEEE Transactions on Image Processing*, CVPR, and ICCV. His research interests include computer vision and multimedia Big Data, especially the understanding and generation of visual contents. He is supported by the Research Funds for Excellent Young Researchers from National Nature Science Foundation of China since 2019. He was also selected into the Beijing Nova Program (2017) and ever received the Second-grade Science Award of Chinese Institute of Electronics (2018), two Excellent Doctoral Thesis Award from Chinese Academy of Sciences (2012) and the Beijing Municipal Education Commission (2012), and the First-Grade Science-Technology Progress Award from Ministry of Education, China (2010). He is an IET fellow, and a senior member of the ACM, CIE, and CCF.

Zeyin Song received the BE degree from Tianjin University, China, in 2021. She is currently working toward the MS degree with the School of Electronic and Computer Engineering, Peking University, China. Her research interests include continual learning and representation learning.

Yonghong Tian (Fellow, IEEE) is currently the dean with the School of Electronics and Computer Engineering, a Boya distinguished professor with the School of Computer Science, Peking University, China, and is also the deputy director of Artificial Intelligence Research, PengCheng Laboratory, Shenzhen, China. His research interests include neuromorphic vision, distributed machine learning, and multimedia Big Data. He is the author or coauthor of more than 300 technical articles in refereed journals and conferences. He was the recipient of the Chinese National Science Foundation for Distinguished Young Scholars in 2018, two National Science and Technology Awards, and three ministerial-level awards in China, and obtained the 2015 EURASIP Best Paper Award for *Journal on Image and Video Processing*, and the best paper award of IEEE BigMM 2018, and the 2022 IEEE SA Standards Medallion and SA Emerging Technology Award. He is a senior member of the CIE and CCF, and a member of the ACM.