Borrowing Knowledge From Pre-trained Language Model: A New Data-efficient Visual Learning Paradigm

Wenxuan Ma[1] Shuang Li[1‡] JinMing Zhang[1] Chi Harold Liu[1] Jingxuan Kang[2]
Yulin Wang[3] Gao Huang[3]
[1]Beijing Institute of Technology [2]University of Liverpool [3]Tsinghua University
{wenxuan, shuangli, jm-zhang}@bit.edu.cn liuchi02@gmail.com sgjkang3@liverpool.ac.uk
wang-yl19@mails.tsinghua.edu.cn gaohuang@tsinghua.edu.cn

Abstract

The development of vision models for real-world applications is hindered by the challenge of annotated data scarcity, which has necessitated the adoption of data-efficient visual learning techniques such as semi-supervised learning. Unfortunately, the prevalent cross-entropy supervision is limited by its focus on category discrimination while disregarding the semantic connection between concepts, which ultimately results in the suboptimal exploitation of scarce labeled data. To address this issue, this paper ***presents a novel approach that seeks to leverage linguistic knowledge for data-efficient visual learning****. The proposed approach, BorLan, Borrows knowledge from off-the-shelf pretrained Language models that are already endowed with rich semantics extracted from large corpora, to compensate the semantic deficiency due to limited annotation in visual training. Specifically, we design a distribution alignment objective, which guides the vision model to learn both semantic-aware and domain-agnostic representations for the task through linguistic knowledge. One significant advantage of this paradigm is its flexibility in combining various visual and linguistic models. Extensive experiments on semi-supervised learning, single domain generalization and few-shot learning validate its effectiveness. Code is available at* https://github.com/BIT-DA/BorLan.

1. Introduction

The tremendous accomplishment of deep learning in computer vision is mostly supported by large-scale labeled datasets [13, 46]. Nevertheless, in real-world scenarios, the acquisition of extensive labeled data through manual annotation for each specific task is a time-consuming and labor-exhaustive endeavor [11, 70]. As such, the development of data-efficient learning methods has become an imperative research direction aimed at enhancing the feasibility and practicality of deep neural networks [53, 61].

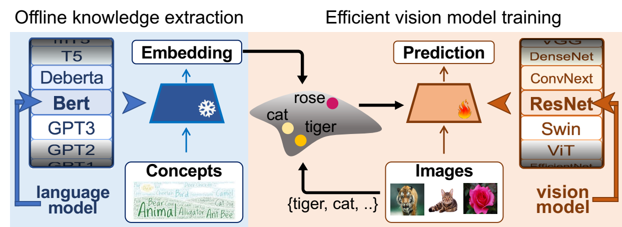

Figure 1: Illustration of BorLan. In both domains of language and vision, we can easily have access to various off-the-shelf models that are pretrained on large datasets in their respective modalities. This paper proposes a data-efficient visual learning paradigm (**black arrows**), aiming to improve various vision models on challenging data-scarce vision tasks by borrowing linguistic knowledge from frozen pretrained language models. In this way, we successfully leverage the rich semantics embedded in language modality to enhance data-efficiency in visual learning.

To mitigate the requirement for labeled data, techniques leveraging supplementary visual knowledge have been extensively investigated in vision community. For instance, transfer learning [63, 60, 27, 64] employs models pretrained on a large image dataset as the initialization, semi-supervised learning [26, 3, 44] exploits unlabeled data via self-training, and out-of-domain generalization [68, 58, 55] incorporates visual prior knowledge in the training using methods such as data augmentation [12]. However, their commonly adopted cross-entropy supervision mainly emphasizes category discrimination, while overlooking the semantic relevance between visual concepts. As a result, the learned image feature space may become distorted [25], and the inter-class relationships inferred by the model can become ambiguous, as shown in Fig 5. This observation motivates us to explore an additional form of supervision that can capture semantic information from image annotations

‡Corresponding author.

prior to their conversion into one-hot labels.

In this paper, we propose a novel approach to address the challenge of annotated data scarcity in vision tasks by leveraging off-the-shelf pretrained language models (PLMs), such as BERT [21] and GPT [41], to **provide explicit semantic guidance that is generalizable to various data-scarce scenarios**. PLMs are known to possess semantically rich embedding spaces, since they are pretrained on large corpora. Therefore, we borrow the general linguistic knowledge embedded within these models to enhance the data-efficiency of visual learning.

Particularly, vision models will benefit from two merits that the text embedding space of PLM possesses: (1) the semantic relationship between concepts could be reflected through text embedding similarities, *i.e.*, concept "cat" is more similar to "tiger" than "airplane"; (2) the concepts expressed in language are more domain-agnostic, which means they are less affected by styles of varying visual domains, *i.e.*, description "a photo of a cat" can be applied indiscriminately to cats in different kinds of environments. Therefore, by aligning image feature space towards the text embedding space, vision model can learn semantic relationships between concepts and domain-invariant knowledge for the given task.

More specifically, we combine a set of predetermined prompts with task-specific concepts to create the input sentences, and obtain text embeddings through the PLM. To capture all possible variants of each concept, we estimate the text embedding distribution of each concept using the generated embeddings. Finally, a distribution-aware knowledge transfer objective is optimized in its upper bound form to guide the vision model align its image representations with the text distribution. The framework is shown in Fig. 1.

Recently, motivated by the strong feature transferability and open-set recognition ability of the pretrained vision-language models (VLM) like CLIP [40] and ALIGN [20], a series of subsequent works adopt VLM to improve few-shot learning performance on data-scarce tasks [72, 71, 33, 15, 69, 19]. These VLM-based tuning methods such as CoOp [72] and Tip-Adapter [69] inherit and leverage the vision-language semantic connection established through joint pre-training on massive image-text pairs to efficiently adapt the model to specific tasks with few labeled samples. Different from them, our framework is designed to be more **flexible**, enabling knowledge transfer between various independently pretrained vision and language models and is also applicable on jointly pre-trained vision-language models.

We evaluate our method in three representative data-efficient learning scenarios: semi-supervised learning (SSL), single domain generalization (SDG), and few-shot learning (FSL). All scenarios pose serious challenges to vision models as they need to capture the high-level semantics within the training data instead of merely memorizing them. We empirically validate that our method consistently improves the performances of data-efficient training on a variety of benchmarks for these tasks, and we demonstrate that our method can promote vision models of different architectures and sizes, ranging from ResNet-50 [17] to Swin-Base [30], with the guidance knowledge obtained from a various choice of PLMs like BERT [21] and GPT [41].

We summarize our contributions in this work as follows:

- We present a novel data-efficient visual learning paradigm, named BorLan, that borrows lingnguistic knowledge from PLMs for explicit semantic guidance and as a complement to scarce visual data.
- We propose text embedding distribution-aware objective, enabling flexible combination of various independently or jointly pretrained vision and language models, and full parameter fine-tuning on specific visual tasks for better adaptation performance.
- Extensive experiments on three scenarios and various benchmarks are conducted to thoroughly validate our method and gain empirical insights.

2. Related Work

Data-efficient visual learning. It is demanding to learn a well-performed model on the given task when the annotated data is limited. To complement the inadequate labeled data, approaches in data-efficient visual learning seek additional knowledge from other sources. Transfer learning [63, 60, 27, 64] transfers the knowledge from models pretrained on large-scale database to the data-scarce tasks. However, the tuned model may bias towards the limited labeled data in the new task [53] and results in feature distortion of the original smooth model [25]. Semi-supervised learning [26, 48, 44, 67, 2] utilizes unlabeled data to explore the intrinsic data structure [56], in which pseudo-labeling technique is widely adopted [26, 42]. However, pseudo-labels are inevitably noisy and the inaccurate labels lead to confirmation bias [6] hence limiting the model performance. Out-of-domain generalization [58, 34] leverage visual knowledge priors to construct image augmentations, and help the model to learn domain-invariant [1] or causal [29, 34] features to generalize beyond the limited training data. Nevertheless, most popular augmentation techniques such as color jittering [8] and mixup [68] can hardly reflect inter-class semantic relationships.

In addition to the pros and cons of each of these technologies, the visual supervision unanimously adopted by them, such as cross-entropy loss, may overlook the semantic information of the concepts by turning class names into one-hot labels. By contrast, linguistic supervision naturally contains rich semantics and is thus potentially more beneficial to serve as visual training guidance in data insufficient tasks. Our method takes a step toward this direction by constructing additional supervision through pretrained language models. Besides its own effectiveness, our method can be regarded as an orthogonal complement to those visual knowledge-based data-efficient learning methods.

Enhance vision models by language. Recently, improving the visual model with the power of language is shown to be effective and promising. The vision-language model (VLM) pretraining based on contrastive learning [40, 20] demonstrates strong feature transferability and open-set recognition ability. These methods focus on learning general representations that can quickly adapt to different downstream tasks. However, they require massive amounts of image-text pairs to establish the connection between image and language semantics. To improve the data efficiency in this paradigm, DeCLIP [28] explores the data correlation both within and across modalities, LiT [66], Frozen [49] and either leverage pretrained image or language models as improved starting points.

Apart from improving the pretraining strategy, a stream of researches [72, 71, 33, 15, 69, 19] is conducted to enhance the few-shot learning performance using pretrained VLMs, which shares similar goals to this article and is referred as VLM-based efficient tuning methods. These approaches leverage the image-text connection learned by a pair of vision and language models, and adapt to downstream tasks efficiently through adjusting a small set of parameters such as text prompts [72, 33] or image keys [69]. However, they have two common limitations. First, these methods rely on coupled or jointly pretrained vision-language models to form a retrieval-based classification head, thus cannot be naturally extended to individually pretrained image models. Second, they keep the visual encoder frozen during model adaptation, thus restrict the model potential of improvement compared to end-to-end fine-tuning. Different from the VLM-based efficient tuning methods, our method decouples the pretrained vision and language models and enables the parameter within the vision backbone to be updated, therefore enjoys more flexibility in model selection for both modalities and possesses greater potential in model adaptation for downstream tasks.

There are other methods leveraging linguistic knowledge with different purposes: K-LITE [43] focuses on external knowledge utilization, LocTex [32] stresses localization and VisualGPT [7] targets at image captioning. Our proposed BorLan focuses on exploiting the semantic richness of the language feature space for visual learning guidance.

3. Method

In real-world applications, training a deep vision model to achieve satisfying performance could be challenging due to the scarcity of label supervision. Therefore, data-efficient training strategies are essential in practical scenarios. We consider semi-supervised learning (SSL), single domain generalization (SDG) and few-shot learning (FSL) as typical scenarios that demand for data-efficient training techniques. In SSL, the training data includes labeled data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_l} \in \mathcal{X} \times \mathcal{Y}$ and unlabeled data $\{(\boldsymbol{u}_i)\}_{i=1}^{n_u} \in \mathcal{X}$, where usually the labeled data set size n_l is much smaller than unlabeled set size n_u. For FSL, only a class-balanced labeled set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_f \times K}$ is provided where n_f indicates number of shots and K is the total category number. While the test data are sampled from the same distribution as the training data in both scenarios, they are not in the SDG setting. In SDG, all the training data are sampled from a single source domain: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_s} \in \mathcal{X}_s \times \mathcal{Y}$, but the test data are expected from arbitrary unseen target domain $\mathcal{X}_t \times \mathcal{Y}$ that shares the same category space.

A vision model could generally be regard as a classification head G on top of a feature extractor F, and has a basic optimization objective of minimizing empirical risk:

$$\mathcal{L}_{emp} = \frac{1}{n} \sum_{i=1}^{n} \ell(G(F(\boldsymbol{x}_i)), y_i), \tag{1}$$

where $\ell(\cdot, \cdot)$ typically takes the form of cross-entropy loss in classification tasks. On this basis, data-efficient training methods generally involve additional training objectives like $\mathcal{L}_u(\boldsymbol{u})$ or $\mathcal{L}_{aug}(\boldsymbol{x}, y; \alpha)$, leveraging unlabeled data $\boldsymbol{u}$ or certain data augmentation technique α, respectively.

This section will first introduce the construction of a new form of objective $\mathcal{L}_{text}(\boldsymbol{x}, y; T)$ by gathering knowledge from a frozen pretrained language model T such as BERT [21]. Then, it specifies the application of the proposed objective on the three data-scarce scenarios.

3.1. Beneficial Supervision from Language Models

We would like to borrow knowledge from language modality to complement the supervision insufficiency in data-scarce vision tasks. Vision-language pretraining set a nice example by learning from web-scaled image-text pairs, but it is a laborious and inflexible approach that has to train a pair of vision-and-language models with a great amount of paired data. Instead, we seek for a more friendly solution that acquires knowledge from pretrained language models.

PLMs like BERT [21] and GPT [41] have shown great success in natural language processing. They learn contextualized word embeddings from large corpus, and capture rich linguistic knowledge within their pretrained model weights [62]. Given that a variety of powerful and off-the-shelf PLM publicly available, it is our interest to investigate how to extract from them the linguistic knowledge beneficial to vision model training.

Normally, to calculate the loss ℓ in Eq. (1), category labels y are turned into one-hot vectors. This common practice encourages the model to discriminate each concept, yet inevitably loses the semantic relationship between them. As a consequence, it is difficult for the vision model to learn the

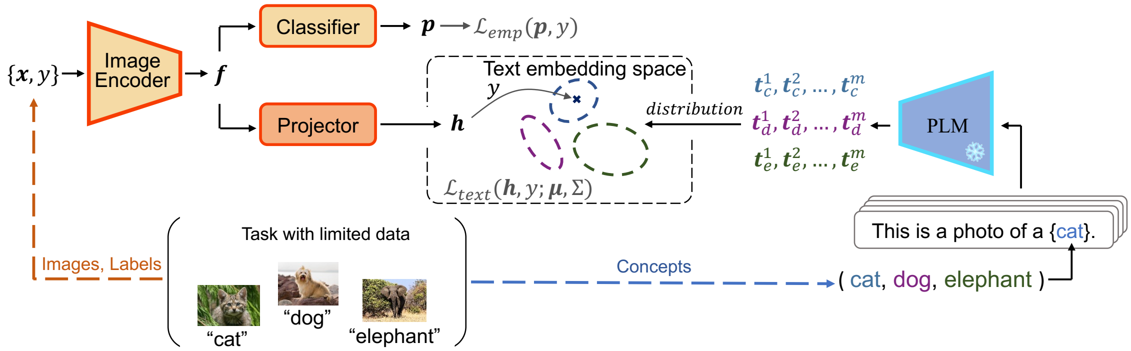

Figure 2: Illustration of the proposed Language embedding space supervision framework. Given a specific vision task with limited labeled data, before training the vision model, we insert category names of into the predetermined prompts to construct input sentences, which are then passed through a frozen pretrained language model (PLM). The generated text embeddings $\boldsymbol{t}$ are utilized to estimate category-wise embedding distributions (as shown in dashed ellipses) in the text embedding space. During training, the language-guided alignment loss $\mathcal{L}_{text}$ is computed besides the standard cross-entropy loss to transfer the linguistic knowledge from PLM to the vision model.

connection between concepts of the task, especially when labeled data is scarce. In contrast, the *text embedding space* generated by PLM contains rich semantics that have two favorable properties: (i) semantic relationship between concepts are reflected through text embedding similarities, (ii) concepts expressed in language are more domain-agnostic. Therefore, we propose to align the feature distribution of the vision model towards the text embedding distribution to help it capture semantics omitted in original visual training.

Given the category names of a specific task $\{\mathcal{W}_k\}_{k=1}^K$ where K is the total category number, we use a set of predetermined prompts (*e.g.*, 'This is a photo of a { }') to complete the input sentences. Specifically, assuming the prompt set has size m, then we can totally obtain mK sentence embeddings by feeding the inputs into a frozen pretrained language model T. These embeddings are then normalized and are denoted as $\{\boldsymbol{t}_1^{(k)}, \boldsymbol{t}_2^{(k)}, ..., \boldsymbol{t}_m^{(k)}\}_{k=1}^K \in \mathbb{R}^{d_{text}}$ where d_{text} is the dimension of the text embedding space.

To conduct feature alignment between image representations and these obtained text embeddings, we initialize a new projector network H on top of the image encoder to obtain the image representations. For the labeled training data $(\boldsymbol{x}, y)$, we compute its normalized representation $\boldsymbol{h} = \frac{H(F(\boldsymbol{x}))}{||H(F(\boldsymbol{x}))||_2} \in \mathbb{R}^{d_{text}}$ and utilize the contrastive loss that regards $\{\boldsymbol{t}_1^{(y)}, \boldsymbol{t}_2^{(y)}, ..., \boldsymbol{t}_m^{(y)}\}$ as positive samples and those in the rest categories as negative samples. The loss is as follows:

$$\mathcal{L}_{text}^{sample}(\boldsymbol{x}, y; T)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_p}\sum_{p=1}^{m_p}\left[-\log\frac{e^{\tau \boldsymbol{h}_i^\top \boldsymbol{t}_p^{(y_i)}}}{e^{\tau \boldsymbol{h}_i^\top \boldsymbol{t}_p^{(y_i)}} + \sum\limits_{k\neq y_i}^{K}\frac{1}{m_n}\sum\limits_{q=1}^{m_n} e^{\tau \boldsymbol{h}_i^\top \boldsymbol{t}_q^{(k)}}}\right], \tag{2}$$

where τ is the temperature hyperparameter, and m_p, m_n denotes the number of positive and negative samples, respectively. Note that here we have $m_p = m_n = m$, and we distinguish these notations only for the derivation in § 3.2.

Despite that the loss in Eq. (2) is an applicable objective, directly optimizing it creates a dilemma regarding the number of handcrafted prompts: small m could not provide enough supervision whereas large m requires heavy labor on prompt engineering. Moreover, as shown in Fig. 4, a few poorly designed prompts lead text embeddings to form "prompt cluster" instead of "concept cluster", making them toxic to feature alignment and thus requires extra effort for manually removal. To overcome these issues, we propose an improved version of $\mathcal{L}_{text}^{sample}$ from a distributional perspective. Specifically, by viewing the text embeddings with the same concept as samples from an underlying distribution of the concept, the image representations can directly align to the distribution, as shown in the following.

3.2. Alignment Between Image Features and Language Concept Distributions

Our modification begins by assuming that text embeddings with input sentences describing the same concept follow a Gaussian distribution in the embedding space. Its mean vector can be viewed as the prototypical embedding of the concept whereas its variance represents the concept in different contexts. Therefore, the parameters for the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ of concept k are estimated through the handcrafted embeddings as:

$$\boldsymbol{\mu}^{(k)} = \frac{\sum_{j=1}^m \boldsymbol{t}_j^{(k)}}{m}, \; \boldsymbol{\Sigma}^{(k)} = \frac{\sum_{j=1}^m (\boldsymbol{t}_j^{(k)} - \boldsymbol{\mu}^{(k)})(\boldsymbol{t}_j^{(k)} - \boldsymbol{\mu}^{(k)})^\top}{m-1}. \tag{3}$$

Once all the concept distributions are estimated, we can sample infinite positive and negative samples and take the limitation of Eq. (2) as m_p and m_n goes to infinity:

$$\begin{aligned}\mathcal{L}_{text}^{\infty} &= \lim_{m_p\to\infty, m_n\to\infty} \mathcal{L}_{text}^{sample} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}^{(y_i)}\sim\mathcal{N}^{(y_i)}}\left[-\log \frac{e^{\tau \boldsymbol{h}_i^\top \boldsymbol{t}^{(y_i)}}}{e^{\tau \boldsymbol{h}_i^\top \boldsymbol{t}^{(y_i)}} + \sum_{k\neq y_i}^{K}\mathbb{E}_{\boldsymbol{t}^{(k)}}\left[e^{\tau \boldsymbol{h}_i^\top \boldsymbol{t}^{(k)}}\right]}\right].\end{aligned} \tag{4}$$

Then, following the derivation of [54], we can further obtain its upper bound using Jensen's inequality and moment generation function (detailed derivation in supplementary):

$$\begin{aligned}\mathcal{L}_{text}^{\infty} &\leq \bar{\mathcal{L}}_{text}^{\infty} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[-\log\frac{e^{\mathcal{F}(\boldsymbol{h}_i, y_i)}}{\sum_{k=1}^{K} e^{\mathcal{F}(\boldsymbol{h}_i, k)}} + \frac{\tau^2}{2}\boldsymbol{h}_i^\top \boldsymbol{\Sigma}^{(y_i)}\boldsymbol{h}_i\right] \\ &\stackrel{\text{def}}{=} \mathcal{L}_{text}(\boldsymbol{x}, y; \boldsymbol{\mu}, \boldsymbol{\Sigma}),\end{aligned} \tag{5}$$

where $\mathcal{F}(\boldsymbol{h}, y) \stackrel{\text{def}}{=} \tau\boldsymbol{h}^\top\boldsymbol{\mu}^{(y)} + \tau^2\boldsymbol{h}^\top\boldsymbol{\Sigma}^{(y)}\boldsymbol{h}/2$, and $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ represents all the means and covariance matrices which only depends on the chosen language model T and the concepts $\mathcal{W}$ and will not be updated during training. To this end, we obtain the actual objective for our linguistic knowledge transfer, which can be seamlessly integrated with various data-scarce scenarios. Fig. 2 illustrates the whole process.

3.3. Application on three scenarios

Our method supports flexible training of vision models on a variety of real-world applications.

In SSL, we calculate $\mathcal{L}_{text}$ using both labeled and unlabeled data besides $\mathcal{L}_{emp}$. To incorporate unlabeled data $\boldsymbol{u}$, we assign pseudo label $\hat{y}$ based on network prediction. The objective is summarized as follows:

$$\mathcal{L}_{ssl} = \lambda_s\mathcal{L}_{emp}(\boldsymbol{x}, y) + \lambda_x\mathcal{L}_{text}(\boldsymbol{x}, y) + \lambda_u\mathcal{L}_{text}(\boldsymbol{u}, \hat{y}). \tag{6}$$

As in both SDG and FSL, we simply compute $\mathcal{L}_{text}$ on all labeled data available and combine it with $\mathcal{L}_{emp}$ as follows:

$$\mathcal{L}_{sdg} = \mathcal{L}_{fsl} = \lambda_s\mathcal{L}_{emp}(\boldsymbol{x}, y) + \lambda_x\mathcal{L}_{text}(\boldsymbol{x}, y). \tag{7}$$

Our method can also be applied to other data-efficient scenarios by simply adding $\mathcal{L}_{text}$ to labeled data or combining it with other techniques. We present here a general applicable algorithm (see Alg. 1), please refer to the supplementary for a more detailed algorithm.

Algorithm 1: Language Guided Vision Training

Input: Data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$; Concepts $\{\mathcal{W}_k\}_{k=1}^{K}$; Prompt set $\{\mathcal{P}_q\}_{q=1}^{m}$; Vision Backbone F; Pre-trained Language Model T.
Output: Language augmented Model for the Task: $G \circ F$ (G is the task-specific head).

```
// Obtain text embeddings.
```
Combine $\mathcal{P}$ with $\mathcal{W}$ to obtain complete input texts $\{\mathcal{P}_1\mathcal{W}_k, \mathcal{P}_2\mathcal{W}_k, ..., \mathcal{P}_m\mathcal{W}_k\}_{k=1}^{K}$, then obtain from T the output text embeddings $\{\boldsymbol{t}_1^{(k)}, ..., \boldsymbol{t}_m^{(k)}\}_{k=1}^{K}$;
for $k = 1, 2, \cdots, K$ **do**
    Estimate $\boldsymbol{\mu}_k$, Σ_k for concept $\mathcal{W}_k$ using Eq. (3);
end
```
// Train the vision model.
```
Initialize classifier G and projector H;
for $iter = 1, 2, \cdots, I$ **do**
    $\boldsymbol{f}_i \leftarrow F(\boldsymbol{x}_i)$;
    $\boldsymbol{p}_i \leftarrow G(\boldsymbol{f}_i)$, $\boldsymbol{h}_i \leftarrow normalize(H(\boldsymbol{f}_i))$;
    Compute $\mathcal{L}_{emp}(\boldsymbol{p}_i, y_i)$ by Eq. (1);
    Compute $\mathcal{L}_{text}(\boldsymbol{h}_i, y_i)$ by Eq. (5);
    $\mathcal{L} \leftarrow \mathcal{L}_{emp} + \lambda\mathcal{L}_{text}$;
    Update model F, G, H by $\mathcal{L}$;
end

4. Experiments

In this section, we evaluate our method on several benchmarks and make comprehensive analysis under three representative data-efficient learning settings: semi-supervised learning (SSL), single domain generalization (SDG) and few-shot learning (FSL). For SSL, following [53], we adopt *CIFAR-100* [24], *FGVC Aircraft* [35], *Stanford Cars* [23] and *CUB-200-2011* [52] to cover from general to fine-grained classification tasks. For SDG, following [22], we evaluate our method on small-sized *Office-Home* [51] and large-scaled *DomainNet* [39]. As for FSL, we follow [72] and conduct experiments on *Caltech101* [14], *FGVC Aircraft*, *DTD* [37], *EuroSAT* [18], *Oxford Flowers* [36], *Oxford Pets* [38], *Stanford Cars*, *Food-101* [4], *SUN397* [57] and *UCF101* [45].

Implementation Details. We use 80 handcrafted prompts proposed in CLIP [40] to obtain text embeddings. We set all λ_s, λ_x and λ_u as 1.0 in SSL and FSL while setting λ_s as 0.3 in SDG. Temperature τ is fixed as $\frac{1}{0.07}$. SGD with a momentum of 0.9 is adopted as the optimizer. The learning rate is set as 1e-3 for the visual backbone in most experiments and a $10\times$ larger value is applied for the classifier and projector in SSL and SDG. The projector is an MLP consists of "FC-ReLU-BN-FC", where the output dimension depends on the text embedding dimension d_{text}. More details can be found in the supplementary.

4.1. Semi-supervised Learning

Our baselines include two type of methods. Vanilla fine-tuning, co-tuning [64] and LP-FT [25] use only the labeled data provided. Five semi-supervised methods [26, 44, 9, 53,

Table 1: Classification accuracy (%) of our method and various baselines on three fine-grained classification benchmarks (backbone: ResNet-50 pretrained on ImageNet-1k). Our method is denoted as BorLan-[language model]-[vision model].

Method	*FGVC Aircraft*					*Stanford Cars*					*CUB-200*				
	15%	30%	50%	65%	80%	15%	30%	50%	65%	80%	15%	30%	50%	65%	80%
Fine-tuning (supervised baseline)	39.57	57.46	67.93	71.31	76.89	36.77	60.63	75.10	79.01	81.07	45.25	59.68	70.12	71.18	71.84
Co-Tuning [64] *(NeurIPS'20)*	44.09	61.65	72.73	–	–	46.02	69.09	80.66	–	–	52.58	66.47	74.64	–	–
LP-FT [25] *(ICLR'22)*	43.51	59.13	68.35	71.88	77.61	40.79	62.54	76.38	80.22	83.64	46.18	59.13	71.86	71.99	72.20
Pseudo-Labeling [26] *(ICML'13)*	46.83	62.77	73.21	–	–	40.93	67.02	78.71	–	–	45.33	62.02	72.30	–	–
FixMatch [44] *(NeurIPS'20)*	55.53	71.35	78.34	–	–	49.86	77.54	84.78	–	–	44.06	63.54	75.96	–	–
SimCLRv2 [9] *(NeurIPS'20)*	40.78	59.03	68.54	–	–	45.74	61.70	77.49	–	–	45.74	62.70	71.07	–	–
Self-Tuning [53] *(ICML'21)*	64.11	76.03	81.22	86.98	88.33	72.50	83.58	88.11	90.77	90.82	64.17	75.13	80.22	80.69	81.34
DebiasMatch [59] *(CVPR'22)*	59.54	71.23	77.10	79.31	81.19	75.39	86.10	89.98	90.55	91.27	64.67	75.05	77.73	78.11	79.28
BorLan-Bert-L-ResNet-50	**71.05**	**83.41**	**87.22**	**88.81**	**90.19**	**79.34**	**88.78**	**91.46**	**92.30**	**92.38**	**65.96**	**75.70**	**80.91**	**81.86**	**82.79**

Table 2: Classification accuracy (%) on *CIFAR-100* provided with only 400 labels, 2, 500 labels and 10, 000 labels.

Method	400	2.5k	10k
FixMatch [44] *(NeurIPS'20)*	42.03	70.01	78.31
ReMixMatch [2] *(ICLR'20)*	46.85	68.56	79.09
Co-Tuning [64]*(NeurIPS'20)*	42.42	69.06	77.78
FlexMatch [67] *(NeurIPS'21)*	43.11	69.87	77.30
Self-Tuning [53] *(ICML'21)*	52.83	75.84	82.43
BorLan-Bert-L-EfficientNet-B2	**55.18**	**76.93**	**83.44**

59] leverage both labeled and unlabeled data. All methods including our BorLan use vision models (such as ResNet-50 [17]) pretrained on ImageNet-1k [13] as backbone. As for the language model, we adopt the representative pre-trained Bert-Large [21] to produce text embeddings. In the rest of this section, we denote our configuration in a unified format of "BorLan-[language model]-[vision model]" (*i.e.*, BorLan-Bert-L-ResNet-50).

Results on Three Fine-grained Datasets. We evaluate BorLan's performances using labeled dataset of proportion ranging from 15% to 80%. The results are shown in table 1. Our method achieves the best performances on all tasks on the three benchmarks. More significant improvements can be observed when the proportion of labeled data is smaller: we surpass Self-Tuning [53] by 6.94%, 6.84% and 1.79% on three benchmarks under 15% labeled data setting. Meanwhile, FixMatch [44] and DebiasMatch [59] are representative semi-supervised baselines that utilizes both strong and weak augmentations to achieve better exploitation of the unlabeled data. Our method, through transferring the linguistic knowledge to the vision model, outperforms the two opponents without using any strong augmentation techniques. Moreover, our method surpasses SimCLRv2 [9], which distills visual knowledge from a teacher vision model. Different from SimCLRv2, we distill knowledge from a language model that learns rich semantics through large corpus and achieves better results.

Results on CIFAR100. Following [53], we adopt the pretrained EfficientNet-B2 [47] model as backbone. We report the results in table 2, including several representative

Table 3: Classification accuracy (%) on two fine-grained classification benchmarks using different pre-training methods. The method of pre-training is written in brackets.

Dataset	Method	Labeling Ratio		
		15%	30%	50%
FGVC Aircraft	Pseudo-Labeling [26] *(ICML'13)*	46.83	62.77	73.21
	FixMatch [44] *(NeurIPS'20)*	55.53	71.35	78.34
	BorLan-Bert-L-RN-50 (Supervised)	71.05	83.41	87.22
	BorLan-Bert-L-RN-50 (MoCov2 [10])	**74.26**	**86.11**	**88.25**
Stanford Cars	Pseudo-Labeling [26] *(ICML'13)*	40.93	67.02	78.71
	FixMatch [44] *(NeurIPS'20)*	49.86	77.54	84.78
	BorLan-Bert-L-RN-50 (Supervised)	**79.34**	**88.78**	91.46
	BorLan-Bert-L-ViT-B (MAE [16])	76.79	87.31	**91.58**

semi-supervised learning methods [44, 67, 2, 53] as baselines. Similar conclusion can be drawn from the table: our method surpasses all the baselines on all three tasks, and gets the most performance boost on the task with the least data available (400 labels only). In addition, it shows that our method can be applied to various pure image pretrained backbones (ResNet and EfficientNet), which is a major advantage compared to VLM-based methods like CoOp [72].

Experiments with Self-supervised Pre-trained Vision Models. Our method is effective not only on image models pre-trained in supervised manner, but also on those models pre-trained in prevalent self-supervised manner. Table 3 shows the results of our method with backbones using vanilla supervised pre-training, MoCov2 [10] and MAE [16] respectively. We can observe that both MoCov2 pre-trained model and MAE pre-trained model achieves competitive results to supervised pre-trained model. We think the reason why MAE pre-trained backbone performs a little worse than the other variant is that ViT-B, with its more powerful learning capabilities, is more likely to be influenced by noisy pseudo labels in the early stage, which suggests that our method could be integrated with more advanced pseudo-labeling strategies to achieve higher results. This is left for future exploration.

4.2. Single Source Domain Generalization

In single domain generalization, the vision model is trained on only one domain and is tested on multiple target domains. Hence, this setting is more difficult than the

Table 4: Target domain accuracy (%) for single domain generalization on *Office-Home*. Backbone ResNet-50 and ConvNext-S are pretrained on ImageNet-1k, and Swin-B is pretrained on ImageNet-22k. † denotes our implementation.

Image Model	Method	Source:Ar			Source:Cl			Source:Pr			Source:Rw			Avg.
		Cl	Pr	Rw	Ar	Pr	Rw	Ar	Cl	Rw	Ar	Cl	Pr	
ResNet-50 #Param: 23M	ERM†	43.71	67.60	73.78	51.03	60.90	63.32	52.73	38.81	72.21	64.80	44.17	76.89	59.16
	ERM [22](ECCV'22)	46.80	64.40	71.20	52.50	62.50	63.60	49.50	42.50	72.30	66.10	49.00	77.20	58.40
	FACT† [58](CVPR'21)	49.12	64.63	73.30	54.80	62.53	64.60	52.08	45.22	72.34	67.12	48.41	78.08	61.02
	CIRL† [34](CVPR'22)	**50.61**	64.79	72.80	55.79	63.03	65.02	52.41	**46.76**	71.88	65.22	**54.71**	77.09	61.68
	BorLan-Bert-L-ResNet-50	47.97	**70.49**	**76.54**	**57.85**	**66.91**	**69.22**	**57.15**	44.01	**76.11**	**68.64**	48.82	**79.68**	**63.62**
ConvNext-S #Param: 49M	ERM†	54.08	75.86	79.96	66.66	74.26	75.14	64.44	51.30	78.56	71.61	53.14	81.63	68.89
	ERM [22](ECCV'22)	53.40	72.70	78.60	67.50	72.90	75.40	61.80	49.00	80.00	72.20	52.70	80.90	67.90
	CIRL† [34](CVPR'22)	**60.85**	76.57	80.95	69.03	74.45	75.17	66.58	**58.58**	82.40	73.40	**59.17**	83.10	71.69
	BorLan-Bert-L-ConvNext-S	59.54	**80.11**	**84.48**	**71.57**	**79.09**	**81.04**	**70.13**	56.54	**83.82**	**75.90**	57.07	**85.51**	**73.73**
Swin-B #Param: 86M	ERM†	69.32	83.97	87.99	81.52	84.91	86.68	79.21	66.59	87.56	82.80	67.05	89.15	80.56
	ERM [22](ECCV'22)	70.70	86.10	88.50	80.60	84.30	86.70	77.90	66.10	88.30	82.60	69.10	90.40	81.00
	CIRL† [34](CVPR'22)	71.93	84.17	87.02	79.32	84.40	86.75	78.54	67.60	88.68	82.74	**72.13**	89.60	81.07
	BorLan-Bert-L-Swin-B	**73.26**	**86.75**	**90.34**	**85.21**	**88.24**	**89.92**	**82.82**	**70.49**	**90.96**	**84.38**	70.31	**90.27**	**83.58**

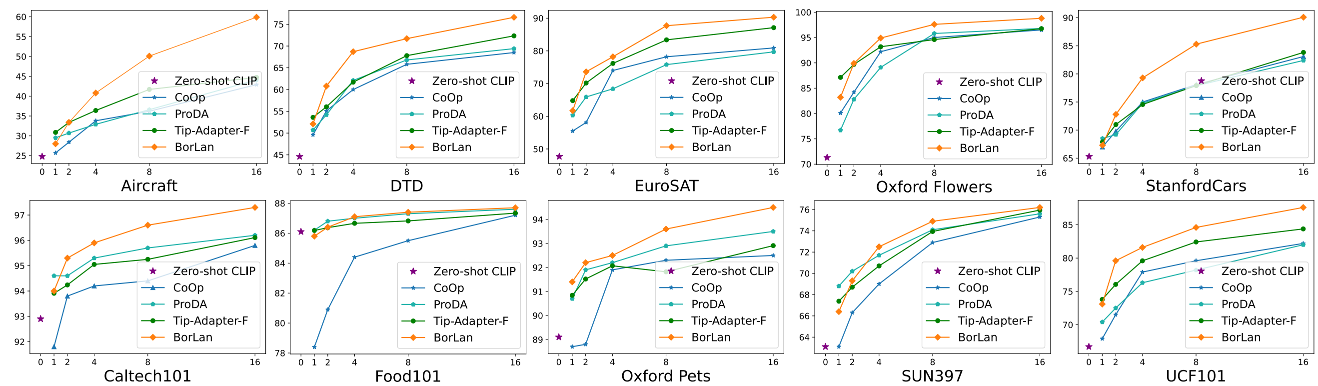

Figure 3: Few-shot learning accuracy (%) on ten datasets compared with VLM-based efficient tuning methods (vision backbone: CLIP ViT-B/16). ProDA is reported using our implementation results based on the official code.

classical domain generalization where multiple source domains are available, but it is also more common in realistic data-scarce scenarios. To validate the effectiveness of our method, we set three baselines: ERM refers to training the model with vanilla cross-entropy loss on all labeled data, FACT [58] and CIRL [34] are strong algorithms in DG. In addition, since recent discovery shows that network architecture and pretraining dataset have large impacts on domain transfer tasks [22], we also examine our method on ConvNext-Small (S) [31] and Swin-Transformer-Base (B) [30] pretrained on ImageNet-1k and -22k, respectively.

Results on Office-Home. The result is shown in table 4. When using ResNet-50 as backbone, BorLan outperforms ERM and CIRL by an average accuracy of 5.22% and 1.94%, respectively. After changing the network architecture and pretraining dataset, our method continues to improve on these vision models, achieving an average performance boost of 5.83% on ConvNext-S and 2.58% on Swin-B. These improvements prove that linguistic knowledge from pretrained language models could serve as ideal complement in enhancing visual feature transferability, and our method is applicable to various vision models.

Results on DomainNet. We show results on more challenging DomainNet benchmark in table 5. Each column reports the average accuracy of five results, with their common target/test domain as the column's title. For example, the number under *clp* is the average performance of five models trained on *inf*, *pnt*, *qdr*, *rel* and *skt* respectively. Compare to vanilla ERM, BorLan improves the generalization performance on all three vision models: ResNet, ConvNext and Swin-Transformer, which indicates the flexibility and the scalability of leveraging linguistic knowledge from PLM. The results also demonstrate that our method is equally effective on large-scaled dataset. Full results can be found in supplementary.

Table 5: Target domain average accuracy (%) ↑ for SDG on large-scaled benchmark *DomainNet*. Backbone ResNet-50 and ConvNext-S are pretrained on ImageNet-1k, and Swin-B on ImageNet-22k. †denotes our implementation results.

Image Model	Method	clp	inf	pnt	qdr	rel	skt	Avg.
ResNet-101 #Param: 42M	ERM [22](ECCV'22)	38.98	12.92	30.98	9.08	41.44	**32.90**	27.72
	ERM†	38.88	14.79	32.16	8.42	43.98	31.05	28.31
	CIRL† [34](CVPR'22)	39.96	13.05	31.20	**9.54**	41.56	31.28	27.77
	BorLan-Bert-L-ResNet-101	**40.21**	**15.62**	**33.09**	9.28	**44.27**	32.17	**29.11**
ConvNext-S #Param: 49M	ERM [22](ECCV'22)	48.34	16.20	38.78	9.50	52.18	39.36	34.06
	ERM†	46.00	17.55	40.02	8.82	54.44	37.07	33.98
	CIRL† [34](CVPR'22)	**48.58**	16.66	41.00	**10.56**	52.34	37.26	34.40
	BorLan-Bert-L-ConvNext-S	47.69	**18.54**	**41.89**	9.39	**56.10**	**39.50**	**35.52**
Swin-B #Param: 86M	ERM [22](ECCV'22)	56.74	21.48	45.80	12.42	60.22	45.50	40.36
	ERM†	55.24	21.62	47.89	10.48	61.68	44.07	40.16
	CIRL† [34](CVPR'22)	58.43	22.70	46.51	13.38	64.01	46.16	41.87
	BorLan-Bert-L-Swin-B	**59.82**	**25.61**	**52.31**	**13.43**	**67.91**	**49.04**	**44.69**

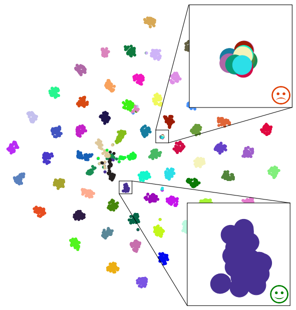

Figure 4: T-SNE visualization of the Bert-L text embeddings spaces on *Office-Home*. Color represents different categories. Best viewed in color.

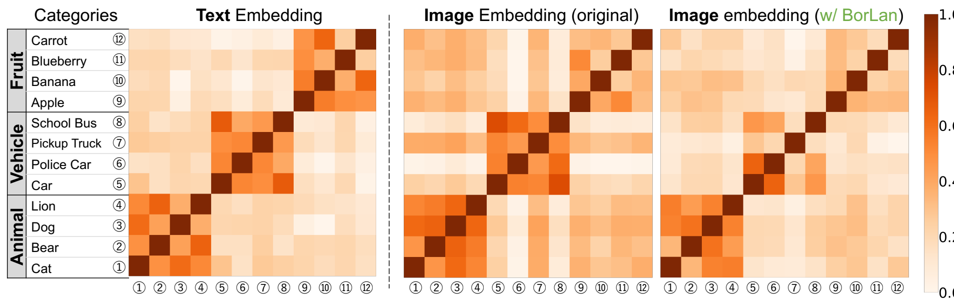

Figure 5: Normalized cosine similarity between the mean text embeddings or image embeddings of 12 selected categories in *DomainNet*. Category indexes are rearranged according to their semantics to form three groups which are shown in the gray boxes in the left. Text embedding space (*left*) can reflect the concept or category similarity, and our method helps the image model (*mid*) learn these semantics (*right*).

4.3. Few-shot Learning

Vision language models (VLM) possess strong zero-shot ability utilizing the image-text semantic connection learned from massive image-text pairs, yet they struggle on further improvements in few-shot learning with vanilla linear probing [40]. Through reusing such cross-modality connection, VLM-based efficient tuning methods successfully improve the few-shot learning performance of CLIP pretrained model [72, 71]. Now we show that BorLan can also enhance CLIP's few-shot ability via the proposed semantic guidance and therefore is beneficial to both pure image pretrained model and pretrained VLM.

We compare BorLan's few-shot learning performance on ten standard benchmarks against CoOp [72], ProDA [33] and Tip-Adapter [69], using CLIP ViT-B/16 as common vision backbone. The result is shown in Fig. 3, where similar trends can be discovered on all datasets. Using 1 shot or 2 shots, BorLan achieves either comparable or a little worse performance compared to the top-performed method. However, as the shot number increases, BorLan continues to achieve large improvements and significantly outperforms all baselines. We speculate on the following reasons.

On one hand, different from VLM-based efficient tuning methods that can inherit the powerful image-text connection obtained in the pretraining stage, our method, with a tunable and decoupled vision encoder plus a new classification head, needs to establish the image-text connection from the beginning using the limited data. As a consequence, When the labeled data is extremely scarce, it is not enough for BorLan to build strong cross-modality connection, and the improvement may not be significant. It can be regarded as a price for our method's increased flexibility in model decoupling. On the other hand, BorLan's capability of full-parameter fine-tuning shows its advantage as the labeled data increases. This is because CoOp and Tip-Adapter can be viewed as language-guided linear probing methods given that both their vision and language encoder is kept frozen to maintain the aforementioned image-text connection.

To summarize, the observed trends in Fig. 3 reflect the difference between the two language-guided paradigm, while our approach BorLan has the advantage of being more adaptable in the few-shot learning scenario.

4.4. Analytical Experiments

Ablation Study. We conduct ablation study on the two losses as in table 6. Firstly, we replace our alignment loss $\mathcal{L}_{text}$ with $\mathcal{L}_{text}^{sample}$ in Eq. (2) and vary the value of the prompt set size m. The results show that the performance increases as m increases, yet it still underperforms our method using $\mathcal{L}_{text}$ even when m is set to 80. Secondly, we remove the cross-entropy loss $\mathcal{L}_{emp}$ (together with the classifier), and instead use fixed text mean vectors as class prototypes for prediction. The results prove that using cross-entropy loss and a trainable classification head makes to model significantly more adaptable and thus cannot be replaced.

Table 6: Ablations of two losses on *Aircraft* (*Air*) and *StanfordCars* (*Car*) datasets.

		$m=5$	$m=10$	$m=20$	$m=40$	$m=80$	∞ ($\mathcal{L}_{text}$)
$\mathcal{L}_{text}^{sample}$	Acc. (*Air*-15%)	68.42	69.11	70.51	70.56	70.66	**71.05**
	Acc. (*Car*-15%)	75.71	75.96	76.54	77.20	78.02	**79.34**
		Air-15%	*Air*-30%	*Air*-50%	*Car*-15%	*Car*-30%	*Car*-50%
$\mathcal{L}_{emp}$	w/o $\mathcal{L}_{emp}$	63.10	71.86	75.07	63.84	77.61	81.95
	w/ $\mathcal{L}_{emp}$	**71.05**	**83.41**	**87.22**	**79.34**	**88.78**	**91.46**

To demonstrate the flexibility of our method, we ablate different PLM on two image backbones: ResNet-50 pretrained on ImageNet-1k and Swin-B pretrained on ImageNet-22k. We also examine effectiveness using the text encoder in CLIP and language model (denoted as $\text{CLIP}_{\text{text}}$). The average accuracy on *Office-Home* of each combination are shown in table 7. The results validate that our method allows free combination between a variety of pretrained image and pretrained language models.

Table 7: Vision and language model ablations on *Office-Home*. Average accuracies (%) are reported.

Image ╲ Language Model	w/o	$\text{CLIP}_{\text{text}}$	Bert-L	mT5-L	GPT2-L
ResNet-50 (IN-1k)	59.16	63.37	63.62	63.64	**63.77**
Swin-B (IN-22k)	80.56	83.41	83.58	83.59	**83.83**

Visualization of the Text Embedding Space. To obtain a more intuitive understanding of the text embedding space, we conduct two experiments to study its properties. Fig. 4 demonstrate the t-SNE visualization results on the text embeddings spaces generated from Bert-L. Embeddings from the same category are painted with the same color. We notice that while most text embeddings in the same category from compact clusters (*denoted as "concept cluster"*), a few of them with the same prompt form an individual cluster (*denoted as "prompt cluster"*), as shown in the zoomed-in view. This is because the prompt template occasionally have large impact on the output embedding and overshadows the concepts that we fill into it. In conclusion, the visualization provides a more intuitive reason why it is necessary to adopt $\mathcal{L}_{text}$ rather than $\mathcal{L}_{text}^{sample}$ to mitigate their negative influence.

Fig. 5 demonstrates the normalized cosine similarity between the means of text embedding distributions and image embeddings of 12 selected categories in DomainNet (*left*). We can observe strong correlation between cosine similarity of text embeddings and the semantic relevance of concepts. For instance, category "cat" is closer to "dog" than "banana". In contrast, original image embeddings (*mid*) learned by one-hot labeled doesn's possess this property. Our method, by aligning image feature to text embedding, helps the vision model learn these semantics (*right*).

Table 8: Clustering and transferability analysis of our method on *Office-Home* (image model: ResNet-50).

Metric	Method	Source:Ar	Source:Cl	Source:Pr	Source:Rw	Avg.
C-H Score ↑	ERM	37.80	42.12	36.29	16.00	33.05
	BorLan	**58.08**	**52.73**	**47.86**	**24.95**	**45.91**
LogME ↑	ERM	1.040	1.032	0.994	0.850	0.979
	BorLan	**1.090**	**1.052**	**1.016**	**0.874**	**1.008**

Transferability Analysis. To quantitatively measure the transferability improvement by our method, we compare between the vision model trained by vanilla ERM and our method using two standard metrics: clustering metric Calinski-Harabasz Index [5] and transferability metric LogME [65]. Specifically, we use the fixed model trained on source domain to generate features in each target domain. Then we leverage the true labels of these target features to calculate both metrics. Table 8 shows the results on Office-Home, where each number represents the average score on three target domains. It is obvious that our method achieve better scores on both metrics, proving that linguistic knowledge is beneficial to representation learning.

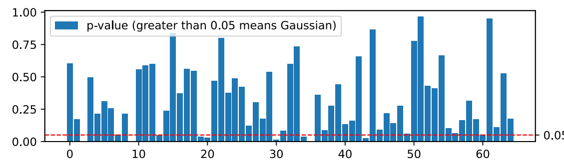

Figure 6: Normality test for 65 concepts in *Office-Home*.

Normality Test of Text Embeddings. To validate the Gaussian assumption in our text embedding space (i.e., the text embeddings generated from PLM are sampled from Gaussian distribution for each concept), we conduct a normality test on each group of concept text features and the results are shown in Fig. 6. The results demonstrate that in the majority of concepts, features have p-values greater than the significance level of 0.05, showing that they are very likely to be Gaussian distributed.

Table 9: Comparison between knowledge distillation from large vision teacher models and large language models (Ours) on *Office-Home* in SDG(student model: ResNet-50). Language teacher improves generalization more.

Teacher	None	ConvNext-S	ConvNext-B	Swin-B	Bert-S [50] (Ours)
#Param	–	49M	87M	86M	29M
Acc. (%)	59.16	61.12	61.28	61.31	**63.57**

Comparison to Knowledge Distillation. Our method can generally be regarded as distilling knowledge from language teacher to vision students, thus we compare it with classical vision knowledge distillation. As shown in table 9, transferring knowledge from language model achieves better generalization improvements on student model, showing that language teacher is able to transfer more semantics.

5. Conclusion

This paper proposes a generalizable data-efficient visual learning paradigm BorLan that leverages linguistic knowledge from pre-trained language model to provide explicit semantic guidance as complementary supervision. The proposed paradigm is designed to allow a flexible combination of various visual and linguistic models, and the proposed objective can transfer the semantic information from text embeddings to visual feature space. Extensive experiments on SSL, SDG and FSL are conducted to validate the effectiveness of this new paradigm in data-efficient learning.

Acknowledgement

This paper was supported by National Key R&D Program of China (No. 2021YFB3301503), and also supported by the National Natural Science Foundation of China under Grant No. 62272046, 62132019, 61872337.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR, abs/1907.02893*, 2019. 2

[2] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *ICLR*, 2020. 2, 6

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 1

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014. 5

[5] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974. 9

[6] Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debiased self-training for semi-supervised learning. In *NeurIPS*, 2022. 2

[7] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *CVPR*, pages 18030–18040, 2022. 3

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 2

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, pages 22243–22255, 2020. 6

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR, abs/2003.04297*, 2020. 6

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 1

[12] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, pages 702–703, 2020. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1, 6

[14] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*, pages 178–178, 2004. 5

[15] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *CoRR, abs/2110.04544*, 2021. 2, 3

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 6

[18] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J-STARS*, 12(7):2217–2226, 2019. 5

[19] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *CoRR, abs/2204.03649*, 2022. 2, 3

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 2, 3

[21] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. 2, 3, 6

[22] Donghyun Kim, Kaihong Wang, Stan Sclaroff, and Kate Saenko. A broad study of pre-training for domain generalization and adaptation. In *ECCV*, 2022. 5, 7

[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *CVPR Workshops*, pages 554–561, 2013. 5

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[25] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022. 1, 2, 5, 6

[26] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013. 1, 2, 6

[27] Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, Zeyu Chen, and Jun Huan. Delta: Deep learning transfer using feature map with attention for convolutional networks. In *ICLR*, 2019. 1, 2

[28] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022. 3

[29] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. In *NeurIPS*, pages 6155–6170, 2021. 2

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 11012–11022, 2021. 2, 7

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 7

[32] Zhijian Liu, Simon Stent, Jie Li, John Gideon, and Song Han. Loctex: Learning data-efficient visual representations from localized textual supervision. In *ICCV*, pages 2167–2176, 2021. 3

[33] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, pages 5206–5215, 2022. 2, 3, 8

[34] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *CVPR*, pages 8046–8056, 2022. 2, 7

[35] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR, abs/1306.5151*, 2013. 5

[36] Nilsback Maria-Elena and Zisserman. Andrew. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 5

[37] Cimpoi Mircea, Maji Subhransu, Kokkinos Iasonas, Mohamed Sammy, and Vedaldi. Andrea. Describing textures in the wild. In *CVPR*, 2014. 5

[38] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012. 5

[39] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019. 5

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 3, 5, 8

[41] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2, 3

[42] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, 2021. 2

[43] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. K-lite: Learning transferable visual models with external knowledge. *CoRR, abs/2204.09222*, 2022. 3

[44] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *CoRR, abs/2001.07685*, 2020. 1, 2, 6

[45] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR, abs/1212.0402*, 2012. 5

[46] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, pages 843–852, 2017. 1

[47] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 6

[48] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2

[49] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *NeurIPS*, pages 200–212, 2021. 3

[50] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *CoRR, abs/1908.08962*, 2019. 9

[51] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 5

[52] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[53] Ximei Wang, Jinghan Gao, Mingsheng Long, and Jianmin Wang. Self-tuning for data-efficient deep learning. In *ICML*, pages 10738–10748, 2021. 1, 2, 5, 6

[54] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *NeurIPS*, 32, 2019. 5

[55] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *ICCV*, pages 834–843, 2021. 1

[56] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *ICLR*, 2022. 2

[57] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010. 5

[58] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pages 14383–14392, 2021. 1, 2, 7

[59] Wang Xudong, Wu Zhirong, Lian Long, and X Yu Stella. Debiased learning from naturally imbalanced pseudo-labels for zero-shot and semi-supervised learning. In *CVPR*, 2022. 6

[60] LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, pages 2825–2834, 2018. 1, 2

[61] Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, and Gao Huang. Resolution adaptive networks for efficient inference. In *CVPR*, pages 2369–2378, 2020. 1

[62] Liang Yao, Chengsheng Mao, and Yuan Luo. Kg-bert: Bert for knowledge graph completion. *CoRR, abs/1909.03193*, 2019. 3

[63] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, pages 3320–3328, 2014. 1, 2

[64] Kaichao You, Zhi Kou, Mingsheng Long, and Jianmin Wang. Co-tuning for transfer learning. In *NeurIPS*, pages 17236–17246, 2020. 1, 2, 5, 6

[65] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *ICML*, pages 12133–12143. PMLR, 2021. 9

[66] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18123–18133, 2022. 3

[67] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, pages 18408–18419, 2021. 2, 6

[68] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2017. 1, 2

[69] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, pages 493–510, 2022. 2, 3, 8

[70] Ziwei Zheng, Le Yang, Yulin Wang, Miao Zhang, Lijun He, Gao Huang, and Fan Li. Dynamic spatial focus for efficient compressed video action recognition. *TCSVT*, 2023. 1

[71] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 2, 3, 8

[72] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 2, 3, 5, 6, 8