# TPTE: Text-Guided Patch Token Exploitation for Unsupervised Fine-Grained Representation Learning

SHUNAN MAO and HAO CHEN, School of Computer Science, Peking University, Beijing, China
YAOWEI WANG, Peng Cheng Laboratory, Shenzhen, China
WEI ZENG and SHILIANG ZHANG, School of Computer Science, Peking University, Beijing, China

Recent advances in pre-trained vision-language models have successfully boosted the performance of unsupervised image representation in many vision tasks. Most of existing works focus on learning global visual features with Transformers and neglect detailed local cues, leading to suboptimal performance in fine-grained vision tasks. In this article, we propose a text-guided patch token exploitation framework to enhance the discriminative power of unsupervised representation by exploiting more detailed local features. Our text-guided decoder extracts local features with the guidance of texts or learned prompts describing discriminative object parts. We hence introduce a local-global relation distillation loss to promote the joint optimization of local and global features. The proposed method allows to flexibly extract either global or global-local features as the image representation. It significantly outperforms previous methods in fine-grained image retrieval and base-to-new fine-grained classification tasks. For instance, our Recall@1 metric surpasses the recent unsupervised retrieval method STML by 6.0% on the SOP dataset. The code is publicly available at https://github.com/maosnhehe/TPTE.

CCS Concepts: • **Computing methodologies** → **Image representations**;

Additional Key Words and Phrases: Image Retrieval, Fine-grained, Cross modal

**ACM Reference format:**
Shunan Mao, Hao Chen, Yaowei Wang, Wei Zeng, and Shiliang Zhang. 2024. TPTE: Text-Guided Patch Token Exploitation for Unsupervised Fine-Grained Representation Learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 11, Article 352 (November 2024), 18 pages.
https://doi.org/10.1145/3673657

This work is supported in part by ZTE-PKU joint research project of task scheduling of video computing power network (project no. IA20230629009), in part by the Natural Science Foundation of China under Grant No. U20B2052, in part by the China Postdoctoral Science Foundation under Grant No. 2023M730056, in part by the Natural Science Foundation of China under Grant No. 61936011.
Authors' Contact Information: Shunan Mao, School of Computer Science, Peking University, Beijing, China; e-mail: snmao@pku.edu.cn; Hao Chen, School of Computer Science, Peking University, Beijing, China; e-mail: hchen@pku.edu.cn; Yaowei Wang, Peng Cheng Laboratory, Shenzhen, China; e-mail: wangyw@pcl.ac.cn; Wei Zeng, School of Computer Science, Peking University, Beijing, China; e-mail: weizeng@pku.edu.cn; Shiliang Zhang (corresponding author), School of Computer Science, Peking University, Beijing, China; e-mail: slzhang.jdl@pku.edu.cn.

ACM 1551-6865/2024/11-ART352
https://doi.org/10.1145/3673657

## 1 Introduction

Contrastive pre-training of large-scale Vision-Language Models has demonstrated impressive zero-shot representation learning capabilities. Notably, models like CLIP [42] are pre-trained on extensive web-scale datasets containing over 400 million image-text pairs. Unlike vanilla supervised training, which is restricted to a closed set of concepts or classes, CLIP pre-training employs natural language. As a result, it creates a shared text-vision embedding space that is not restricted to a fixed set of classes. The flexibility of natural language allows CLIP to handle diverse and open-ended concepts.

Despite the impressive performance on generic vision-language tasks, CLIP still lacks the necessary discriminative power to properly distinguish fine-grained classes [18, 19, 35, 45, 52, 56, 60]. This can be attributed to two main issues. First, in the vision-language alignment process, only the global feature obtained from the [CLS] token is used to represent the image feature. Consequently, the other tokens extracted by the backbone model have not been thoroughly trained and utilized for vision-language alignment. This limitation may lead to a suboptimal alignment. Second, the constrastive loss in the CLIP pre-training process considers the anchor pair itself as the positive and all other pairs as negatives, which may hinder the model from accurately capturing fine-grained relationships between samples.

To amplify the discriminative ability of a model on fine-grained classes, previous works [23, 27, 58, 59] usually fine-tune a pre-trained model on a dataset with fine-grained categories. After lightweight prompt learning or adapter learning with the labeled data, the model yields better performance in fine-grained image recognition. Nevertheless, the process of annotating fine-grained data is labor-intensive, which requires not only accurate alignment of images with fine-grained categories but also comprehensive coverage of all potential categories.

In this article, we propose an unsupervised **Text-guided Patch Token Exploitation (TPTE)** framework to improve the discriminative ability of fine-grained categories by perceiving more detailed local features. As illustrated in Figure 1, besides the [CLS] token extracted by the visual encoder, we additionally leverage the text encoder to extract local features from image patch tokens. The inputs of text encoder can be learnable or knowledge-based text prompts indicating discriminative object parts. To facilitate the local feature extraction, we introduce a decoder inspired by the DETR framework [5]. The decoder extracts local features from image patch tokens with the guidance of text queries. The outputs correspond to discriminative local attributes, providing crucial clues in describing instances within the context of fine-grained tasks.

To train global and local features with unlabeled data, we introduce two kinds of loss functions. As neighbor instances are more likely to be in the same class, we first exploit clues from neighbor instances within the feature space. The relationship is refined with k-reciprocal nearest neighbors. Then a local-global relation distillation loss is applied to promote information exchange between local and global feature branches. Global features are robust to local noises, leading to a more stable k-reciprocal nearest neighbor. Local features usually focus on specific regions, which can be more sensitive to fine-grained details. Therefore, global features are applied to provide a reliable neighbor relationship, and the trained local features capture local detail cues to further boost the discriminative power. In addition, we propose a novel soft-weight contrastive loss to supervise the learning of both global and local features. Conventional contrastive loss only considers the anchor itself as the positive, which hinders the model from exploring similarities between different views of a class. Differently, we propose to mine meaningful nearest neighbor information and use neighbor instances as possible positives in contrastive learning. This strategy leverages neighboring instances to boost the feature discriminative capabilities.

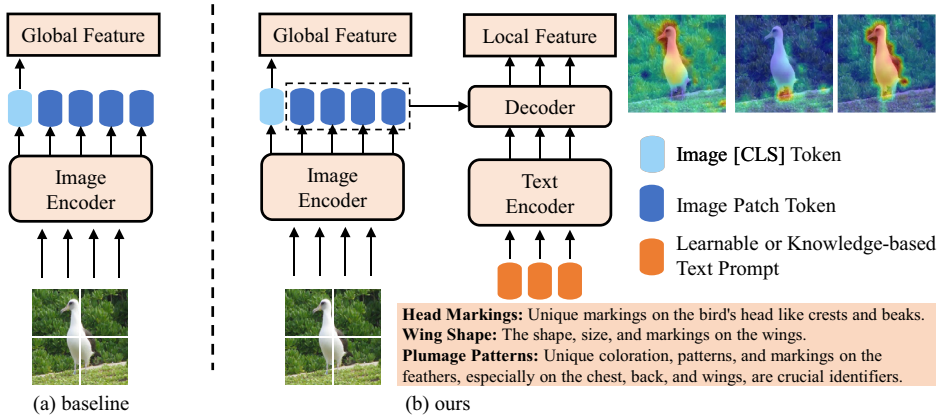


Fig. 1. Comparison of the baseline and our proposed method. (a) Baseline: an image representation is built with the global feature (the [CLS] token) of the visual encoder. (b) Ours: in addition to the global feature, we leverage text prompts to extract fine-grained local features from patch tokens.

We validate the effectiveness of our method on three fine-grained image retrieval datasets and four fine-grained image classification datasets, respectively. Our results exhibit substantial improvements over recent methods. Notably, the proposed method surpasses the state-of-the-art unsupervised retrieval approach by 4.6% on the CUB dataset [49] and by 6.0% on the SOP dataset [38]. Furthermore, experiments conducted on base-to-new image classification tasks reveal that our approach simultaneously enhances both the discrimination and generalization capacities of representations generated by CLIP [42].

To summarize, our method contributes in the following ways: (1) We introduce a decoder-like local feature extractor to effectively exploit the local details from the visual patch tokens and text encoder of the CLIP model. (2) We propose a novel local-global relation loss and a soft-weight contrastive loss to train both global and local features, which effectively enhance the discriminative ability of the model. (3) Through extensive experimentation on both unsupervised image retrieval and base-to-new image classification tasks, we demonstrate the effectiveness of our proposed approach in improving the performance of CLIP across diverse fine-grained vision tasks.

## 2 Related Works

### 2.1 Vision-Language Pre-Training and Fine-Tuning

Vision-language pre-training aims to learn a versatile and transferable model that can effectively handle both vision and language modalities [4, 32, 33, 47, 48, 54]. The pre-trained models like CLIP [42] and ALIGN [22] have demonstrated great potential in acquiring generic visual representations and enabling zero-shot transfer to downstream classification tasks. However, despite the inclusion of similar semantics in the large-scale pre-training sets, the pre-trained models often show inferior performance compared to specialized models trained on specific downstream tasks.

Some works use few-shot downstream data to train a small adapter on the pre-trained model to address this issue. CoOP [59] and CoCOOP [58] focus on tuning the classification layers for downstream tasks. VPT [23] and CLIP-A [13] choose to add a few parameters to the visual model. MUST [31] does not restrict the size of the downstream data, but the labels of each image are unknown. To address this, they utilize the class names in the training set and assign pseudo labels for each training image using the text encoder. However, the availability and reliability of fine-grained

class names can be problematic. For instance, learning representations of online products [38] in this way proves challenging because the detailed types are always unavailable or meaningless.

### 2.2 Attribute Learning

Attributes can provide rich information for learning the correlation between fine-grained categories. Therefore, many works on fine-grained tasks have embraced attribute learning strategies. One kind of attribute learning methods is to learn an attribute feature for each sample. Early works [34, 44] train the backbone model with an auxiliary attribute loss. However, these methods require the datasets to be annotated with attribute labels. Recent works [20, 30, 46, 53] learn the attribute features with class-level supervision. Another kind of attribute learning methods [7, 8, 12, 21, 36, 50] do not explicitly extract the attribute feature for each sample. Instead, all samples in the dataset share the same attribute prototypes. They utilize the attribute prototypes as guidance to discover the more discriminative region features. Some of these methods [21, 50] generate attribute prototypes with learnable parameters. They add an additional layer to the backbone to make the feature focus on discriminative parts. Others [7, 8] use handcrafted attributes like "blue head."

### 2.3 Unsupervised Representation Learning

Unsupervised representation learning involves learning discriminative representations on a fine-grained dataset from a pre-trained model without using any human-annotated labels. It can also be divided into pre-training methods and fine-tuning methods.

Pre-training methods commonly employ self-training strategies. They typically fall into two categories. First, some approaches [9, 16] leverage contrastive learning. These methods strive to enrich the feature space by strategically pushing apart representations of different samples. Secondly, others [15, 57] opt for a masked image modeling approach. These methodologies revolve around the idea of generating a feature space through the reconstruction of masked patches of images. However, these methods often fail to fully exploit the semantic information embedded in fine-grained data.

Some fine-tuning methods also adopt the self-learning framework. Wu et al. [51], Ye and Shen [55] assign unique labels to each training sample and employ contrastive learning to create a meaningful feature space. However, they may struggle to effectively model variations within each latent class in fine-grained downstream tasks. Pseudo labeling is another popular unsupervised fine-tuning method, which assign pseudo labels to unlabeled training data. Some approaches [6, 25, 26] generate pseudo labels for the entire training set using offline clustering algorithms like $k$-means. However, these auxiliary algorithms bring in high computational complexity and prevent the end-to-end training. An alternative approach, as seen in STML [28], generates pseudo labels within mini-batches and leverages contextualized semantic similarity computed from the teacher model to distill feature relations to the student model.

### 2.4 Difference from Prior Approaches

Since annotating fine-grained data is labor-intensive, this paper introduces an unsupervised framework to tune the CLIP model for fine-grained vision tasks. Compared with prior unsupervised representation learning methods, this paper proposes a text-guided patch exploitation method. The modality alignment in the CLIP model enables our model to generate attribute prototypes with learnable or knowledge-based text prompts. The learned local attribute features effectively enhance the discrimination ability of our model on fine-grained classes. Compared with previous attribute learning methods that require class-level annotation [8, 21] and even attribute-level annotation [8] in the training set, our method avoids the cumbersome annotation process and learns fine-grained

attribute features in an unsupervised manner. To the best of our knowledge, this is an original research on unsupervised fine-grained representation learning with local information exploitation.

## 3 Methodology

This section first presents an overview of our unsupervised fine-grained representation learning framework, named as TPTE. We then describe each module of TPTE in detail, including the proposed architecture and loss functions.

### 3.1 Overview

CLIP [42] is a widely used backbone trained on a vast dataset of image-text pairs. Its robust semantic capabilities make it an promising pre-training model for downstream tasks such as image retrieval and zero-shot learning. The architecture of CLIP consists of two distinct encoders: the visual encoder $V(\cdot)$ and the text encoder $T(\cdot)$. The output of the visual encoder $V(\cdot)$ with the input image $x_i$ has two parts: the global feature $g_i$ from the [CLS] token and the local patch features $L_i$ from the patch tokens. We can readily formulate image retrieval and zero-shot classification using the global feature as

$$\begin{aligned} &\text{image retrieval:} && \arg\min_{x_j \in R} \|g_i - g_j\|, \\ &\text{zero-shot classification:} && \arg\min_{y_c \in Y} \|g_i - T(y_c)\|, \end{aligned} \tag{1}$$

where $x_j$ is the $j$th image in the gallery dataset $R$, and $y_c$ is the class name of the $c$th category. Since the global feature of the image encoder and the text encoder have been aligned in the same space, the result of the argmin function can be the retrieved image, and the identified category name, respectively.

To fine-tune the clip visual encoder, previous works [13] add a few learnable parameters, called the adapter, to the visual model, meanwhile keeping the original visual model fixed. To train the adapter with unlabeled data, the most intuitive way is self-supervised learning with a conventional contrastive loss, which can be formulated as follows:

$$\mathcal{L}^{contrast} = -\sum_i \log \frac{e^{g_i^T g_i}}{e^{g_i^T g_i} + \sum_{j \neq i} e^{g_i^T g_j}}, \tag{2}$$

where $g_i$ is the global feature of $x_i$ extracted by the adapted visual encoder.

The above strategies provide a feasible baseline for image retrieval and classification. However, it suffers from the two problems in fine-grained tasks. (1) It neglects the local information in $L_i$, which is important to discriminate fine-grained categories. The outputs of $V(\cdot)$ correlated to the local patches of the input image are not used in Equation (1). (2) The conventional contrastive loss considers all other samples as negatives, making it unable to capture meaningful relationships between samples.

To address the first issue, we propose an unsupervised TPTE to learn the local information with unlabeled data. Given an image $x_i$, we aim to extract its global features and discriminative local features by referring to text prompts $\{t_m\}$ describing local object parts. Text prompts $\{t_m\}$ can be generated by **Large Language Model (LLM)**, or learned through end-to-end training. We use the text encoder to extract the text queries $\{T(t_m)\}$. The adapted visual encoder is used for extracting the visual tokens $V(x_i)$, including the global token $g_i$ and the patch tokens $L_i$. Both the text queries $\{T(t_m)\}$ and visual features $V(x_i)$ are used as the inputs of TPTE. The output of TPTE is a global feature $g_i$ and a local feature $l_i$. The conceptual formulation is followed and the detailed structure

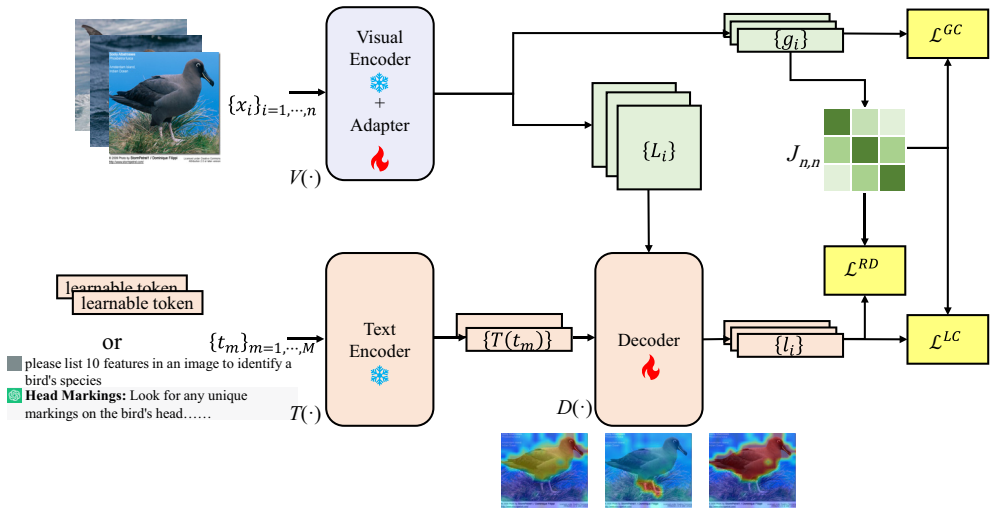


Fig. 2. Unsupervised learning framework of TPTE. For each input image, the image encoder with Adapter extracts one global feature $g_i$ and patch tokens $L_i$. Given several learnable or knowledge-based sentences as text prompts $t_m$, the text encoder outputs text queries $T(t_m)$. The text queries exploit the patch tokens through a decoder and output the discriminative local feature $l_i$. The global and local features are then jointed optimized *w.r.t.* the proposed losses along with other features in the batch.

is discussed in Section 3.2

$$g_i, l_i = \text{TPTE}(\{T(t_m)\}, V(x_i)). \tag{3}$$

To address the second issue, we introduce two kinds of loss functions tailored to the intrinsic characteristics of fine-grained data. Different from the conventional contrastive loss in Equation (2), we consider the fine-grained relationships between samples present within the mini-batch. We leverage neighbor clues embedded within the feature space, which inherently convey the genuine categories of the unlabeled images. Therefore, we propose soft-weight contrastive loss functions $\mathcal{L}^{GC}$ and $\mathcal{L}^{LC}$, which are applied to supervise the learning of global feature and local feature, respectively. In addition, we propose a local-global relation distillation loss $\mathcal{L}^{RD}$ to further align the local and global features together. The overall loss function is followed and the details are discussed in Section 3.3

$$\mathcal{L} = \mathcal{L}^{RD}(\{l_i\}; \{g_i\}) + \mathcal{L}^{GC}(\{g_i\}) + \mathcal{L}^{LC}(\{l_i\}). \tag{4}$$

Figure 2 presents the framework of our proposed method. Details of TPTE and the loss computation will be presented in the following parts.

### 3.2 Structure of TPTE

The outputs of TPTE include the global features $g_i$ and the local feature $l_i$. To get $g_i$, we directly use the [CLS] token generated by the adapted visual model. To make $l_i$ capture local details that are intrinsic to fine-grained tasks, we introduce an additional local branch to extract local features as illustrated in Figure 2.

The inputs of the local branch include the visual patch tokens $L_i$ and task-related text prompts $\{t_m\}_{m=1,\cdots,M}$. The main structure of the local branch is a DETR-like [5] decoder $D(\cdot)$. The decoder plays a pivotal role by exploiting diverse patch cues, thereby facilitating the extraction of attribute information from these patch tokens. The outputs of the decoder are max-pooled together to form

a local feature representation. This process can be mathematically formulated as follows:

$$l_i = \text{MaxPool}(\{D(T(t_m), L_i)\}_{m=1,\cdots,M}), \tag{5}$$

where $l_i$ is the local feature of the image $x_i$, $L_i$ is the local patch tokens from the visual encoder, $t_m$ is the $m$th input text prompts describing discriminative object parts. We use the resulting local feature $l_i$ to represent more detailed attribute features of the image $x_i$.

Text prompts $\{t_m\}_{m=1,\cdots,M}$ are expected to provide hints of discriminative local parts to facilitate the learning of local features. It can either be learned through end-to-end training or leveraging knowledge encoded in existing LLMs. A simple way to learn them is by treating them as a set of learnable embedding of DETR [5]. Following previous prompt engineering works [59], we represent a text prompt $t_m$ as,

$$t_m = [SOT, p_m, EOT], \tag{6}$$

where $SOT$ denotes the "start of the text" embedding token, $p_m$ denotes a learnable embedding, and $EOT$ denotes the "end of the text" embedding token, respectively. We experimentally set the number of learnable text queries to

Text prompts can also be generated by referring to off-the-shelf LLMs such as ChatGPT [3, 43]. For different fine-grained datasets, we design different questions. For example, we ask the LLM with the question "Please list local $M$ features in an image to identify a bird's species (a car's model)" for the CUB (Cars196) dataset. The LLM will generate $M$ sentences $\{s_m\}_{m=1,\cdots,M}$ describing discriminative local parts. The knowledge-based text prompts could be constructed as

$$t_m = [SOT, \tau(s_m), EOT], \tag{7}$$

where $\tau(\cdot)$ is a tokenizer turning the words into word tokens. Performance of different text prompts will be tested in Section 4.

### 3.3 Loss Computation

As illustrated in Figure 2, we increase the discriminative power of the final feature by training the adapter in the visual encoder and the decoder $D(\cdot)$. Our method involves two types of losses, i.e., the local-global relation distillation loss and the soft-weight contrastive loss, respectively. Note that, optimizing local features leads to better adapter parameters and image patch tokens. They hence also facilitate the learning of better global feature.

*Local-Global Relation Distillation Loss.* Since the unlabeled data have neither class-level annotation nor attribute-level annotation, we utilize the relationship computed with global features as pseudo annotations to supervise the learning of local features.

We initialize the final layers of adapters as zero, which makes the initial outputs of the visual encoder with and without adapters are the same. This ensures a better initialization to global features than those local features generated by the randomly initialized decoder $D(\cdot)$. It also makes the global feature a more reliable teacher to distill local features.

As different channels of local and global features may represent different cues, it is not appropriate to directly distill features [14, 37]. We compute similarity cues with global features as the guidance to train local features, i.e., we distill the similarity matrix among samples within a mini-batch. To compute reliable similarity cues, we leverage the Jaccard similarity according to the k-reciprocal nearest neighbors to refine the relationship between local features.

As discussed in [28, 41], k-reciprocal nearest neighbors are more robust to the outliers and stable during the training. We formulate the k-reciprocal nearest neighbors of global feature $g_i$ as $R_k(i) = \{g_j | g_j \in N_k(g_i) \wedge g_i \in N_k(g_j)\}$, where $N_k(g_i)$ is the k-nearest neighbor set including $g_i$ itself. If two images are similar, their k-reciprocal nearest neighbors are likely to share more

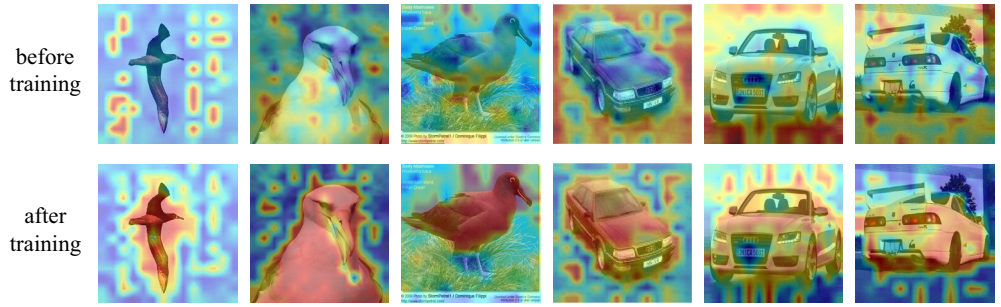


Fig. 3. Visualization of similarity between global feature $g_i$ and local patch tokens $L_i$ before and after training with $\mathcal{L}^{RD}$. Red color denotes higher similarity.

identical samples [41]. The overlap between $\{R_k(i)\}$ is hence adopted to computer the Jaccard similarity between $g_i$ and $g_j$, i.e.,

$$J_{i,j} = \frac{|R_k(i) \cap R_k(j)|}{|R_k(i) \cup R_k(j)|}. \tag{8}$$

The computed Jaccard similarity is hence used to compute the relation distillation loss between global and local features as:

$$\mathcal{L}^{RD}(\{l_i\};\{g_i\}) = -\sum_{i,j} \phi(J_{i,j}) \log\left[\frac{\phi(J_{i,j})}{\phi(l_i^T l_j)}\right], \tag{9}$$

where $\phi$ is the softmax operation, and $l_i^T l_j$ is the cosine similarity between local features $l_i$ and $l_j$.

Local features $\{l_i\}$ are optimized by learning the relationship of global features. Since the gradient of $l_i$ can be back-propagated to the local patch tokens $L_i$ and then to the adapter of the visual encoder, the extraction of the global feature $g_i$ is also influenced by $\mathcal{L}^{RD}$. As shown in the first row of Figure 3, the pre-trained visual encoder cannot produce meaningful local features. After training with $\mathcal{L}^{RD}$, the visual encoder could produce global features depicting discriminative object regions. It indicates that, $\mathcal{L}^{RD}$ enables a joint optimization between local and global features. In other words, it boosts the performance of local features, and also injects discriminative local cues into global features. We hence could only extract global features during inference to save computational cost. Experimental results will be presented in Section 4.

*Soft-Weight Contrastive Loss.* We further propose soft-weight contrastive loss functions to optimize global and local features. Finding positive pairs is crucial for unsupervised training. As ground-truth labels are unavailable, we leverage the nearest neighbors of each feature to get its positive samples as well as to re-weight negative samples.

We follow the approach presented in STML [28] to construct mini-batches containing potential positive samples. To construct a mini-batch with the size $b$, we randomly sample $\frac{b}{k}$ anchors, then search $k$ nearest neighbors for each anchor in the entire training set. These nearest neighbors are subsequently employed as positive samples for each anchor. In addition, conventional contrastive loss considers the negative samples with the same weight, ignoring the relationship between samples. The Jaccard similarity in Equation (8) can represent a more reliable relationship. Samples with larger Jaccard similarity are more likely to be within the same class. Therefore, we utilize the Jaccard similarity to assign soft weights in contrastive learning to avoid treating samples within the same class as negative samples.

With these considerations, the loss functions for the global feature $g_i$ and local feature $l_i$ of the image $x_i$ can be formulated as

$$\mathcal{L}^{GC}(\{g_i\}) = -\sum_i \sum_{p \in N_k(i)} \log \left[ \frac{e^{g_i^T g_p}}{e^{g_i^T g_p} + \sum_{j \notin N_k(i)} (1 - J_{i,j}) e^{g_i^T g_j}} \right],$$
$$\mathcal{L}^{LC}(\{l_i\}) = -\sum_i \sum_{p \in N_k(i)} \log \left[ \frac{e^{l_i^T l_p}}{e^{l_i^T l_p} + \sum_{j \notin N_k(i)} (1 - J_{i,j}) e^{l_i^T l_j}} \right], \tag{10}$$

where $N_k(i)$ is the k nearest neighbors of the sample $x_i$ and $J_{i,j}$ is the Jaccard distance in Equation (8).

The final loss in Equation (4) is hence implemented with above loss functions in Equations (9) and (10).

## 4 Experiments

### 4.1 Datasets and Evaluation Settings

We evaluate the effectiveness of the proposed method in two distinct scenarios: the unsupervised image retrieval and the base-to-new class generalization. In both cases, images in the original training set are utilized for training. Their training labels are unknown. It should be noted that, our approach differs from the MUST framework [31] in that, we do not depend on the class names in the training set. Our method thus shows better potentials to overcome limitations of relying on extra class information during training.

Experiments of unsupervised image retrieval are conducted on three benchmark datasets: CUB-200-2011 (CUB) [49], StanfordCars (Cars196) [29], and **Stanford Online Product (SOP)** [38]. To make fair comparisons, we follow the standard protocol outlined in Kim et al. [28] for the train-test splits of these datasets. We utilize Recall@k as the evaluation metric. This metric reflects the fraction of queries that have at least one relevant sample among their k nearest neighbors. By employing this metric, we can quantitatively assess the retrieval performance and draw meaningful comparisons between our method and existing approaches on these benchmark datasets. We evaluate the performances of base-to-new class generalization on four benchmark datasets: CUB [49], Cars196 [29], Food101 [2], and OxfordPets [39]. To make fair evaluations, we adopt the train-test splits as prescribed in CoCoOP [58]. To measure the efficacy of our method, we compute the zero-shot classification accuracy for both the base sets and the new sets.

### 4.2 Implementation Details

ResNet50 [17] and VIT-B/16 [11] pre-trained by CLIP [42] are adopted as backbones. The dimensions of global features and local queries are all set as the output size of the backbone network, which is 1024 for ResNet50 and 512 for VIT-B/16. The default depth of the proposed decoder is set as 3 and the default number $M$ of text queries is set as 40. Our method allows to flexible extract both global features and local features during inference. As the global feature is jointly optimized by the local feature with the distillation loss, it presents better discriminative power to local details. We only extract global features as the default option.

To allow mini-batches including diverse and relevant samples, we construct mini-batches following STML. The total batch size is $b = 120$ for all datasets. For each mini-batch, we randomly sample $\frac{b}{k}$ anchors and then search for $k$ nearest neighbors for each anchor. Since the SOP dataset has a small number of samples per class, we set $k = 2$. For other training sets, we set $k = 5$. The epoch is set as 20 for the retrieval task and 10 for base-to-new generalization task. All experiments are implemented by pytorch [40] with 2 NVIDIA RTX 3090 GPUs.

Table 1. Retrieval Results on Three Fine-Grained Datasets with Two Backbones of CNN and Transformer

| Methods | Arch. | CUB | | | Cars196 | | | SOP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@2 | R@4 | R@1 | R@2 | R@4 | R@1 | R@10 | R@100 |
| ImageNet† [17] | $R50^{2048}$ | 54.8 | 67.6 | 78.3 | 44.8 | 57.1 | 68.4 | 50.3 | 65.9 | 79.7 |
| CLIP [42] | $R50^{1024}$ | 52.8 | 66.0 | 77.6 | 71.9 | 83.0 | 90.3 | 56.9 | 73.1 | 85.6 |
| ROUL† [25] | $R50^{128}$ | 55.7 | – | – | 49.3 | – | – | 58.5 | – | – |
| CBML† [26] | $R50^{512}$ | 59.2 | – | – | 48.8 | – | – | 59.5 | – | – |
| BDAI† [10] | $GoogLeNet^{512}$ | 60.3 | 72.4 | 82.2 | 48.9 | 60.5 | 72.0 | - | - | - |
| STML† [28] | $GoogLeNet^{512}$ | 60.6 | 71.7 | 81.5 | 50.5 | 61.8 | 71.7 | 65.3 | 79.8 | 89.8 |
| STML [28] | $R50^{1024}$ | *67.3* | *78.2* | *86.5* | *74.2* | *86.4* | *91.1* | *70.8* | *82.9* | *91.6* |
| ours | $R50^{1024}$ | **70.9** | **80.4** | **88.0** | **78.2** | **89.1** | **93.3** | **71.9** | **84.8** | **93.0** |
| ImageNet† [11] | $VIT\text{-}B/16^{768}$ | 53.5 | 65.8 | 76.4 | 49.2 | 60.7 | 71.7 | 52.0 | 66.7 | 79.5 |
| CLIP [42] | $VIT\text{-}B/16^{512}$ | 69.4 | 80.8 | 88.4 | 85.2 | 91.9 | 96.5 | 59.8 | 76.3 | 88.2 |
| MUST* [31] | $VIT\text{-}B/16^{512}$ | *75.6* | *84.7* | **91.1** | *89.5* | *94.6* | **97.6** | – | – | – |
| CBML [26] | $VIT\text{-}B/16^{512}$ | 71.4 | 80.8 | 89.2 | 86.3 | 92.5 | 97.4 | 61.7 | 77.9 | 89.6 |
| STML [28] | $VIT\text{-}B/16^{512}$ | 72.3 | 81.5 | 90.2 | 86.2 | 92.6 | 97.0 | *69.4* | *81.3* | *91.8* |
| ours | $VIT\text{-}B/16^{512}$ | **76.9** | **85.4** | *91.0* | **90.2** | **94.9** | 97.5 | **75.4** | **86.9** | **93.7** |

"R50" and "VIT-B/16" denote that the backbones are Resnet50 and VIT-Base/16, respectively. "ImageNet" and "CLIP" denote two kinds of pre-training methods. Methods with/without † are implemented on the ImageNet/CLIP pre-trained backbone, respectively. Results with ImageNet pre-trained models are reported in their papers. "MUST*" is a weakly supervised method that requires the names of training classes. The highest and second-highest performances are marked with **bold** and italic, respectively.

### 4.3 Comparison with Recent Methods

*Unsupervised Fine-Grained Image Retrieval.* We compare our method with several unsupervised metric learning methods [10, 25, 26, 28] on two backbones, i.e., ResNet50 and VIT-B/16. We also apply our method to ResNet50 and VIT-B/16 pre-trained by CLIP [42]. Results are summarized in Table 1. Note that, our method needs to work with pre-trained CLIP. We thus also implement some methods on CLIP pre-trained backbone with code released by their authors to make a more fair comparison.

Results in Table 1 indicate that, our method achieves competitive performance among compared methods. It outperforms unsupervised methods [10, 25, 26, 28] by clear margins. For example, using the same backbone VIT-B, our method outperforms SMTL [28] by 4.6%, 4.0%, and 6.0% on three datasets, respectively. A weakly supervised CLIP tuning method MUST [31] is also compared. MUST does not need the image-level annotation but requires the class names in the training set. Since SOP [38] does not provide the class name, MUST cannot work on it. Our method substantially outperforms MUST in R@1 and R@2, and gets comparable performance with it in R@4, without leveraging extra class name cues. We hence conclude that, our method achieves promising performance in unsupervised fine-grained image retrieval.

*Base-to-New Class Generalization.* This experiment compares our method against CLIP [42], CoOP [59], CoCoOP [58], VPT [23], CLIP-A [13], KART [24], and MaPLe [27] under the base-to-new class generalization setting in Table 2. It is clear that, supervised methods like CoOP can bridge the relationship between images and the class names, exhibiting promising performance in the base classes. However, they cannot generalize well on unseen classes. As discussed in CoCoOP [58], the CoOP method confronts challenges of overfitting on the base classes, thus is not effective in generalizing to new classes.

Table 2. Base-to-New Class Generalization Results

(a) CUB

| | Base | New | H |
|---|---|---|---|
| *no fine-tuning data* | | | |
| CLIP [42] | 58.7 | 70.3 | 63.9 |
| *16 labeled images per class* | | | |
| CoOP [59] | **79.2** | 53.3 | 63.9 |
| CoCoOP [58] | 67.1 | 74.1 | 70.4 |
| VPT [23] | *68.5* | 70.4 | 69.4 |
| CLIP-A [13] | 68.3 | 70.8 | 69.5 |
| KART [24] | 67.5 | 74.2 | *70.7* |
| MaPLe [27] | 68.1 | *74.3* | **71.0** |
| *unlabeled images* | | | |
| Ours | 67.5 | **75.0** | **71.0** |

(b) Cars196

| | Base | New | H |
|---|---|---|---|
| *no fine-tuning data* | | | |
| CLIP [42] | 63.3 | 74.9 | 68.6 |
| *16 labeled images per class* | | | |
| CoOP [59] | *78.1* | 60.4 | 68.1 |
| CoCoOP [58] | 70.5 | 73.6 | 72.0 |
| VPT [23] | **79.0** | 60.7 | 68.7 |
| CLIP-A [13] | 70.5 | 73.3 | 71.9 |
| KART [24] | 69.5 | 66.2 | 67.8 |
| MaPLe [27] | 72.9 | *74.0* | *73.5* |
| *unlabeled images* | | | |
| Ours | 72.8 | **75.2** | **74.0** |

(c) Food101

| | Base | New | H |
|---|---|---|---|
| *no fine-tuning data* | | | |
| CLIP [42] | 90.1 | 91.2 | 90.7 |
| *16 labeled images per class* | | | |
| CoOP [59] | 88.3 | 82.3 | 85.2 |
| CoCoOP [58] | **90.7** | 91.3 | *91.0* |
| VPT [23] | 87.8 | 85.3 | 86.5 |
| CLIP-A [13] | *90.3* | 91.2 | 90.8 |
| KART [24] | 86.1 | 87.1 | 86.6 |
| MaPLe [27] | **90.7** | *92.1* | **91.4** |
| unlabeled images | | | |
| Ours | *90.3* | **92.5** | **91.4** |

(d) OxfordPets

| | Base | New | H |
|---|---|---|---|
| no fine-tuning data | | | |
| CLIP [42] | 91.1 | 97.2 | 94.1 |
| *16 labeled images per class* | | | |
| CoOP [59] | 93.7 | 95.3 | 94.5 |
| CoCoOP [58] | *95.2* | *97.7* | *96.4* |
| VPT [23] | 94.1 | 94.6 | 94.4 |
| CLIP-A [13] | 94.8 | 97.0 | 95.9 |
| KART [24] | 93.1 | 96.5 | 94.8 |
| MaPLe [27] | **95.4** | **97.8** | **96.6** |
| *unlabeled images* | | | |
| Ours | **95.4** | **97.8** | **96.6** |

"H" denotes the harmonic mean of performances on base and new classes. The highest and second-highest performances are marked with **bold** and italic, respectively.

Compared with recent methods [23, 27] in Table 2, our method consistently achieves the best performance in terms of new class performances and the overall harmonious balance. It also can be observed that, our method also exhibits reasonably good performance on base classes, even if it is not fine-tuned on base classes with labeled data. Since the appearance variance of Food101 is commonly large, fine-tuning with few-shot samples like CoOP may cause overfitting. Our method uses more training samples, thus gets better performance. This result indicates that our method boosts the fine-grained feature discriminative power while keeping the CLIP zero-shot capability.

### 4.4 Ablation Study

This section adopts fine-grained image retrieval task on CUB dataset to test effects of each component and important parameters in our method.

Table 3. Retrieval Results on CUB Under Different Decoder Depths, Numbers of Text Queries $M$ and Neighbor Sizes $k$

(a) Decoder Depth

| Decoder Depth | R@1 | R@2 | R@4 |
|---|---|---|---|
| no decoder | 71.2 | 81.3 | 88.7 |
| 1 | 73.3 | 83.7 | 90.6 |
| 2 | 75.9 | 84.7 | 90.9 |
| 3 | **76.9** | **85.4** | 91.0 |
| 5 | 76.7 | **85.4** | **91.1** |

(b) Numbers of Text Queries $M$

| $M$ | R@1 | R@2 | R@4 |
|---|---|---|---|
| 5 | 74.7 | 84.7 | 90.8 |
| 10 | 75.5 | 84.6 | 90.8 |
| 20 | 76.4 | 85.0 | 90.9 |
| 40 | **76.9** | 85.4 | **91.0** |
| 80 | 76.8 | **85.5** | **91.0** |

(c) Neighbor Size $k$

| $k$ | R@1 | R@2 | R@4 |
|---|---|---|---|
| 2 | 69.1 | 80.2 | 88.6 |
| 3 | 74.6 | 84.3 | 90.7 |
| 5 | **76.9** | 85.4 | 91.0 |
| 10 | 76.8 | **86.2** | **91.8** |
| 20 | 76.6 | 85.6 | 91.7 |

VIT-B/16 is used as backbone model. Bold is a conventional presentation of the best performance under a fair competition.

Table 4. Retrieval Results on CUB with Different Query Types

| Type of $t$ | $T(\cdot)$ | number | R@1 | R@2 | R@4 |
|---|---|---|---|---|---|
| learnable | – | 10 | 73.4 | 84.0 | 90.8 |
| ChatGPT [43] | ✓ | 10 | 74.3 | 84.0 | 90.7 |
| ERINE [1] | ✓ | 10 | 74.0 | 83.9 | 90.8 |
| learnable | ✓ | 10 | 75.5 | 84.6 | 90.8 |
| learnable | – | 20 | 74.0 | 84.3 | 90.8 |
| ChatGPT [43] | ✓ | 20 | 74.4 | 84.1 | 90.8 |
| ERINE [1] | ✓ | 20 | 74.2 | 84.0 | 90.8 |
| learnable | ✓ | 20 | **76.4** | **85.0** | **90.9** |

"Type of $t$" denotes the generalization method of the text prompts. "learnable" denotes that $t$ is a learnable prompt while "ChatGPT" and "ERINE" denote that $t$ is generated by large language models. The symbol "✓" under "$T(\cdot)$" denotes that the text encoder is applied to $t$. Bold is a conventional presentation of the best performance under a fair competition.

*Parameter Analysis.* Table 3(a) presents the impact of the depth of the proposed decoder. Table 3(b) showcases the influence of the number of text queries used. Table 3(c) presents the impact of the neighbor size $k$ during the batch sampling and loss computation. The results in Table 3(a) and (b) highlight that a decoder with a reasonably large depth and width leads to noticeable performance enhancements compared to the baseline. As the depth and width of the decoder increase, there is a substantial performance improvement. Setting a too large width degrades the performance. Based on these observations, we determine the default values for the depth and width parameters as 3 and 40, which could achieve a reasonable tradeoff between accuracy and efficiency. The performance in Table 3(c) is stable when the neighbor size $k \geq 5$. We determine the default values as 5.

*Performance of Different Types of Text Prompts.* Table 4 delves into the impact of different types of text prompts. We compare the performance achieved using learnable prompt sentences with that generated by LLM [1, 3, 43]. The results reveal that both LLM-generated sentences and learnable sentences boost the performance. As the number of prompts increases, the performance of learnable prompts continues to improve. Differently, ChatGPT [43] and ERINE [1] fail to offer extra useful textual prompts. Additionally, the utilization of the text encoder further enhances performance. This demonstrates the effectiveness of our method, i.e., adopting both learnable text prompts and the text encoder to learn more fine-grained visual features.

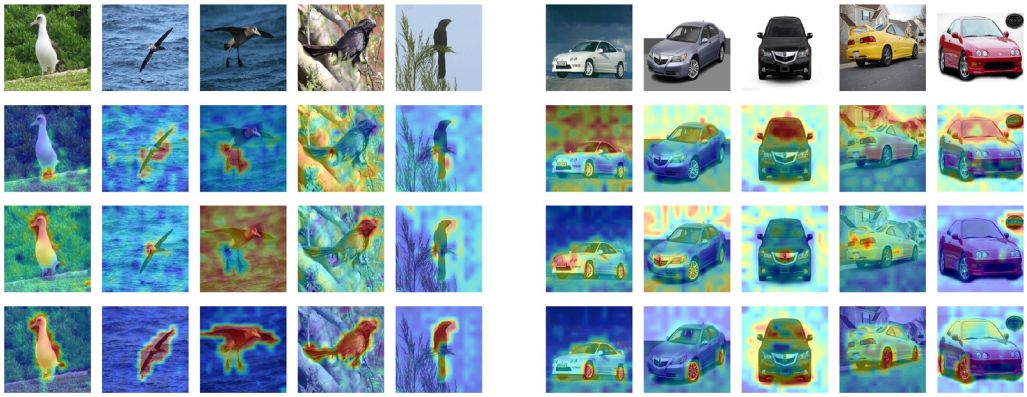

Fig. 4. Visualization of attention maps of learnable text prompts on CUB and Cars, respectively. For each dataset, the similarities between queries generated by three text prompts and five images are shown.

Table 5. Retrieval Results of Learned Global and Local Features, as Well as the Fused Feature on CUB

| Training Stage | | Testing Stage | | R@1 | R@2 | R@4 | MACs (G) |
|---|---|---|---|---|---|---|---|
| Global | Local | Global | Local | | | | |
| ✓ | | ✓ | | 71.2 | 81.3 | 88.7 | 9.63 |
| ✓ | ✓ | ✓ | ✓ | 76.0 | **85.6** | **91.8** | 10.02 |
| ✓ | ✓ | | ✓ | 74.1 | 84.2 | 90.8 | 10.02 |
| ✓ | ✓ | ✓ | | **76.9** | 85.4 | 91.0 | 9.63 |

"MACs (G)" denotes the computational cost for a single input image during inference. Bold is a conventional presentation of the best performance under a fair competition. MACs, multiply-accumulate operations.

Figure 4 illustrates the similarity between patch tokens and queries generated by learnable text prompts. It demonstrates that learnable prompts present the capacity to acquire local semantic attributes, even if they are randomly initialized. For instance, one of the text prompts on the CUB dataset exhibits attention toward the feet of the birds. The other text prompt has attention toward the birds' heads in the next row.

*Evaluation on Local and Global Features.* Our local-global relation distillation loss optimizes the visual encoder to produce better image patch tokens. As the visual encoder is also used to extract global features, our method jointly optimizes local and global features during training. This experiment tests the performance of learned local and global features, as well as the fused feature in Table 5. The fused feature in the table is the average of global and local features. It can be observed that, after training both branches, either the single global feature or the local feature can achieve better performance than the baseline. The global features perform better than local features, and achieve comparable performance with the fused feature. This could be because better image patch tokens lead to more discriminative global features after training. The last column shows the computational cost for a single image input. Since $\{T(t_m)\}$ are the same for all images, they can be offline extracted, which takes 71.18G **Multiply-Accumulate Operations (MACs)**. We adopt global features during inference in subsequent experiments for better efficiency.

*Effectiveness of Loss Functions.* This experiment tests the impact of the three loss functions. Experimental results in Table 6 indicate the importance of those loss functions. Specifically, removing the two contrastive loss functions leads to a degradation of the Recall@1 metric by 3.6%

Table 6. Effectiveness of Proposed Loss Functions on Unsupervised Image Retrieval

| $L^{GC}$ | $L^{LC}$ | $L^{RD}$ | R@1 | R@2 | R@4 |
|---|---|---|---|---|---|
| ✓ | ✓ | ✓ | **76.9** | **85.4** | **91.0** |
| – | ✓ | ✓ | 73.3 | 83.7 | 90.6 |
| ✓ | – | ✓ | 74.1 | 84.2 | 91.0 |
| ✓ | ✓ | – | 71.3 | 82.3 | 90.1 |
| Contrast | ✓ | ✓ | 72.1 | 82.5 | 90.2 |
| ✓ | Contrast | ✓ | 74.1 | 84.3 | 91.1 |
| ✓ | ✓ | KL | 69.9 | 80.9 | 89.0 |

"✓" indicates the incorporation of the proposed loss functions. "Contrast" and "KL" represent the self-supervised contrastive loss and KL divergence on features instead of the proposed loss functions, respectively. Bold is a conventional presentation of the best performance under a fair competition.

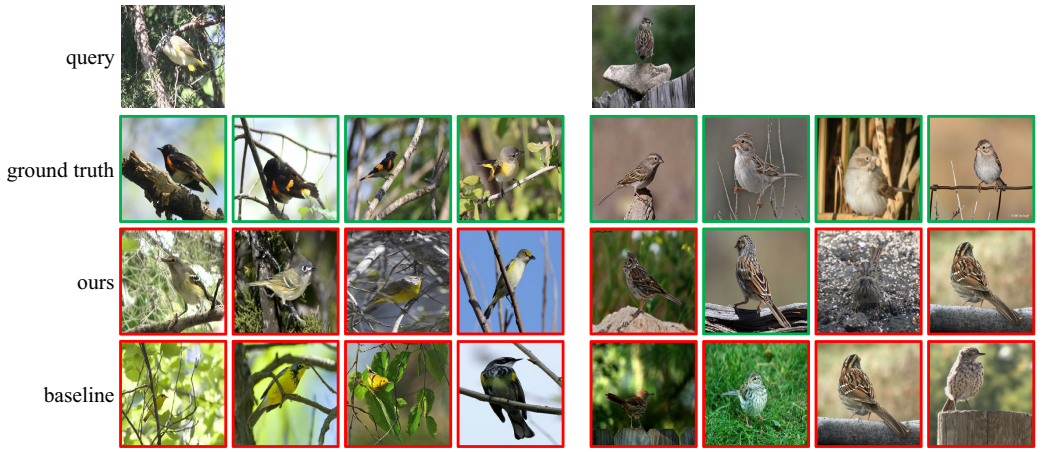


Fig. 5. Visualization of failure cases. The green and red bounding boxes denote the true positive results and false positive results, respectively.

and 2.8%, respectively. The absence of the distillation loss results in a more substantial degradation of 5.6% in R@1. To further show their effectiveness, we replace the proposed loss functions with some vanilla loss functions. "Contrast" refers to the general self-paced contrastive loss, where only the anchor sample is regarded as the positive sample. "**Kullback–Leibler (KL)**" designates the direct KL divergence loss between global and local features, which demonstrates notably worse performance. Those results emphasize the importance of integrating fine-grained relations into the contrastive loss framework. It also indicates the promising performance of considering the Jaccard similarity between k-reciprocal nearest neighbors for distillation. We hence could conclude that the proposed loss functions are important and effective in boosting the model performance.

*Discussions.* Figure 5 shows some failure cases of our method. When the intra-class variance is large, especially when the birds in the same class have different feathers or postures, both the baseline method and the proposed method may fail to find the true positives. For such cases, some

additional knowledge, like the location or sounds of birds, or more gallery images may be useful to improve the retrieval accuracy.

## 5 Conclusion

In conclusion, this paper presents a novel unsupervised fine-tuning approach for CLIP, aiming at enhancing its discriminative capabilities in fine-grained vision tasks. The proposed method incorporates a decoder to exploit previously untapped patch tokens with text guidance to get discriminative local features. The proposed local-global relation distillation loss utilizes the coherent k-reciprocal similarity on global features to supervise the learning of local features. The distillation loss optimizes the visual encoder, hence also boosts the performance of global features. The effectiveness of the proposed method is evident in image retrieval and base-to-new classification tasks. For instance, on the SOP dataset, the Recall@1 metric surpasses the state-of-the-art unsupervised retrieval method by a significant margin of 6.0%. This contribution showcases the potentials of our approach to improving the CLIP performance across various fine-grained vision tasks.

## References


[1] baidu. [n. d.]. Retrieved from https://cloud.baidu.com/product/wenxinworkshop

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision*. Springer, 446–461.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the Advances in Neural Information Processing Systems,* Vol. 33, 1877–1901.

[4] Shaofei Cai, Liang Li, Xinzhe Han, Shan Huang, Qi Tian, and Qingming Huang. 2023. Semantic and correlation disentangled graph convolutions for multilabel image recognition. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*. Springer, 213–229.

[6] Hao Chen, Benoit Lagadec, and François Bremond. 2021. ICE: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14960–14969.

[7] Shiming Chen, Ziming Hong, Wenjin Hou, Guo-Sen Xie, Yibing Song, Jian Zhao, Xinge You, Shuicheng Yan, and Ling Shao. 2023. TransZero++: Cross attribute-guided transformer for zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 11 (2023), 12844–12861.

[8] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. 2022. Transzero: Attribute-guided transformer for zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, 330–338.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1597–1607.

[10] Haoyang Cheng, Hongliang Li, Qingbo Wu, Heqian Qiu, Xiaoliang Zhang, Fanman Meng, and Taijin Zhao. 2023. Disturbed augmentation invariance for unsupervised visual representation learning. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 11 (2023), 6924–6938.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.

[12] Zhanzhou Feng, Jiaming Xu, Lei Ma, and Shiliang Zhang. 2024. Efficient video transformers via spatial-temporal token merging for action recognition. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 4 (2024), 1–21.

[13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* 132, 2 (2024), 581–595.

[14] Jianping Gou, Liyuan Sun, Baosheng Yu, Shaohua Wan, and Dacheng Tao. 2023. Hierarchical multi-attention transfer for knowledge distillation. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)* 20, 2 (2023), 1–20.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 16000–16009.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 9729–9738.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 770–778.

[18] Xiangteng He, Yuxin Peng, and Junjie Zhao. 2019. Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization. *International Journal of Computer Vision* 127 (2019), 1235–1255.

[19] Yunqing Hu, Xuan Jin, Yin Zhang, Haiwen Hong, Jingfeng Zhang, Yuan He, and Hui Xue. 2021a. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4239–4248.

[20] Yutao Hu, Xuhui Liu, Baochang Zhang, Jungong Han, and Xianbin Cao. 2021b. Alignment enhancement network for fine-grained visual categorization. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 1s (2021), 1–20.

[21] Dat Huynh and Ehsan Elhamifar. 2020. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4483–4493.

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4904–4916.

[23] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*. Springer, 709–727.

[24] Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. 2023. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15670–15680.

[25] Shichao Kan, Yigang Cen, Yang Li, Vladimir Mladenovic, and Zhihai He. 2021. Relative order analysis and optimization for unsupervised deep metric learning. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 13999–14008.

[26] Shichao Kan, Zhiquan He, Yigang Cen, Yang Li, Vladimir Mladenovic, and Zhihai He. 2023. Contrastive Bayesian analysis for deep metric learning. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 45, 06 (2023), 7220–7238.

[27] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

[28] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. 2022. Self-taught metric learning without labels. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 7431–7441.

[29] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the International Conference on Computer Vision workshops*, 554–561.

[30] Hao Li, Xiaopeng Zhang, Qi Tian, and Hongkai Xiong. 2020. Attribute mix: Semantic data augmentation for fine grained recognition. In *Proceedings of the IEEE International Conference on Visual Communications and Image Processing (VCIP '20)*. IEEE, 243–246.

[31] Junnan Li, Silvio Savarese, and Steven Hoi. 2022. Masked Unsupervised Self-training for Label-free Image Classification. In *Proceedings of the International Conference on Learning Representations*.

[32] Zhixin Li, Lan Lin, Canlong Zhang, Huifang Ma, Weizhong Zhao, and Zhiping Shi. 2021. A semi-supervised learning approach based on adaptive weighted fusion for automatic image annotation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 1 (2021), 1–23.

[33] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. 2022. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3003–3018.

[34] Xiao Liu, Jiang Wang, Shilei Wen, Errui Ding, and Yuanqing Lin. 2017. Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4190–4196.

[35] Kang Ma, Ying Fu, Dezhi Zheng, Yunjie Peng, Chunshui Cao, and Yongzhen Huang. 2023. Fine-grained unsupervised domain adaptation for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11313–11322.

[36] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 4 (2021), 1–23.

[37] Xiushan Nie, Yang Shi, Ziyu Meng, Jin Huang, Weili Guan, and Yilong Yin. 2023. Complex scenario image retrieval via deep similarity-aware hashing. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)* 20, 4 (2023), 1–24.

[38] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 4004–4012.

[39] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*. IEEE, 3498–3505.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in Neural Information Processing Systems,* Vol. 32, 8026–8037.

[41] Danfeng Qin, Stephan Gammeter, Lukas Bossard, Till Quack, and Luc Van Gool. 2011. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*. IEEE, 777–784.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 8748–8763.

[43] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[44] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. 2016. Deep attributes driven multi-camera person re-identification. In *Proceedings of the European Conference on Computer Vision*. Springer, 475–491.

[45] Hongbo Sun, Xiangteng He, and Yuxin Peng. 2022. Sim-trans: Structure information modeling transformer for fine-grained visual categorization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5853–5861.

[46] Min Tan, Fu Yuan, Jun Yu, Guijun Wang, and Xiaoling Gu. 2022. Fine-grained image classification via multi-scale selective hierarchical biquadratic pooling. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 1s (2022), 1–23.

[47] Wei Tang, Liang Li, Xuejing Liu, Lu Jin, Jinhui Tang, and Zechao Li. 2023. Context disentangling and prototype inheriting for robust visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46 (2023), 3213–3229.

[48] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. 2024. SMART: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46 (2024), 4926–4943.

[49] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. California Institute of Technology.

[50] Shijie Wang, Zhihui Wang, Haojie Li, and Wanli Ouyang. 2022. Category-specific nuance exploration network for fine-grained object retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, 2513–2521.

[51] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 3733–3742.

[52] Jian Xiao and Xiaojun Bi. 2023. Model-guided generative adversarial networks for unsupervised fine-grained image generation. *IEEE Transactions on Multimedia* 26 (2023), 1188–1199.

[53] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. 2020. Attribute prototype network for zero-shot learning. In *Proceedings of the Advances in Neural Information Processing Systems,* Vol. 33, 21969–21980.

[54] Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. 2024. Pink: Unveiling the power of referential comprehension for multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 13838–13848.

[55] Mang Ye and Jianbing Shen. 2020. Probabilistic structural latent representation for unsupervised embedding. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 5457–5466.

[56] Wei Zhao, Cai Xu, Ziyu Guan, Xunlian Wu, Wanqing Zhao, Qiguang Miao, Xiaofei He, and Quan Wang. 2021. TelecomNet: Tag-based weakly-supervised modally cooperative hashing network for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 7940–7954.

[57] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. 2022. Image BERT pre-training with online tokenizer. In *Proceedings of the International Conference on Learning Representations*.

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 16816–16825.
[59] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.
[60] Qiangxi Zhu, Wenlan Kuang, and Zhixin Li. 2023. A collaborative gated attention network for fine-grained visual classification. *Displays* 79 (2023), 102468.

Received 8 January 2024; revised 25 April 2024; accepted 4 June 2024