

# Keyword-Based Diverse Image Retrieval by Semantics-aware Contrastive Learning and Transformer

Minyi Zhao\*  
zhaomy20@fudan.edu.cn  
School of Computer Science  
Fudan University  
Shanghai, China

Jinpeng Wang  
wjp20@mails.tsinghua.edu.cn  
Tsinghua Shenzhen Int'l Graduate  
School, Tsinghua University  
Shenzhen, China

Dongliang Liao†  
brightliao@tencent.com  
Wechat Group  
Tencent Inc.  
Guangzhou, China

Yiru Wang  
dorisywang@tencent.com  
Wechat Group  
Tencent Inc.  
Beijing, China

Huanzhong Duan  
boosterduan@tencent.com  
Wechat Group  
Tencent Inc.  
Beijing, China

Shuigeng Zhou†  
sgzhou@fudan.edu.cn  
School of Computer Science  
Fudan University  
Shanghai, China

## ABSTRACT

In addition to relevance, diversity is an important yet less studied performance metric of cross-modal image retrieval systems, which is critical to user experience. Existing solutions for diversity-aware image retrieval either explicitly post-process the raw retrieval results from standard retrieval systems or try to learn multi-vector representations of images to represent their diverse semantics. However, neither of them is good enough to balance relevance and diversity. On the one hand, standard retrieval systems are usually biased to common semantics and seldom exploit diversity-aware regularization in training, which makes it difficult to promote diversity by post-processing. On the other hand, multi-vector representation methods are not guaranteed to learn robust multiple projections. As a result, irrelevant images and images of rare or unique semantics may be projected inappropriately, which degrades the relevance and diversity of the results generated by some typical algorithms like top- $k$ . To cope with these problems, this paper presents a new method called CoLT that tries to generate much more representative and robust representations for accurately classifying images. Specifically, CoLT first extracts semantics-aware image features by enhancing the preliminary representations of an existing one-to-one cross-modal system with semantics-aware contrastive learning. Then, a transformer-based token classifier is developed to subsume all the features into their corresponding categories. Finally, a post-processing algorithm is designed to retrieve images from each category to form the final retrieval result. Extensive experiments on two real-world datasets Div400 and Div150Cred show that

CoLT can effectively boost diversity, and outperforms the existing methods as a whole (with a higher  $F1$  score).

## CCS CONCEPTS

• **Information systems** → **Information retrieval diversity**.

## KEYWORDS

Cross-modal retrieval, Keyword-based image retrieval, Diversification retrieval, Transformer

## ACM Reference Format:

Minyi Zhao, Jinpeng Wang, Dongliang Liao, Yiru Wang, Huanzhong Duan, and Shuigeng Zhou. 2023. Keyword-Based Diverse Image Retrieval by Semantics-aware Contrastive Learning and Transformer. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, July 23–27, 2023, Taipei, Taiwan. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3539618.3591705>

## 1 INTRODUCTION

With the popularity of the Web and its applications, increasing data are created and posted to the Web, which triggers the rapid development of web search and information retrieval techniques [41, 57]. Among them, cross-modal data retrieval [2, 3, 46, 46, 48, 62, 63, 71] enables users to conveniently acquire desirable information in different forms. A typical example is cross-modal image retrieval (CMIR in short) [5, 37, 49, 50, 59, 66]. CMIR takes a textual query as input to retrieve images with matched semantics, has been deployed in many web applications like Instagram and Flickr, and gains increasing research attention [12, 21, 27, 29, 54].

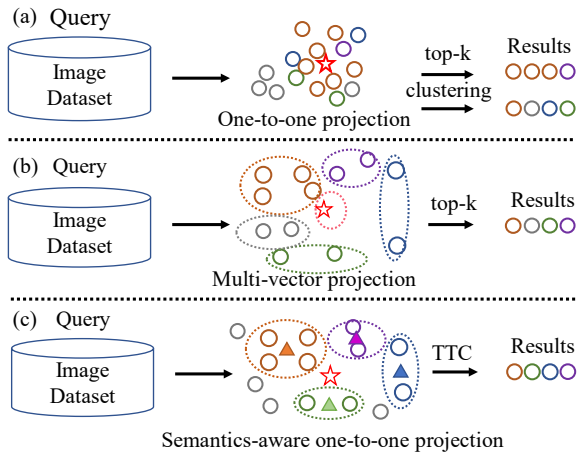
Nowadays, the relevance of cross-modal image retrieval systems has been significantly advanced by recent works [28, 30, 50, 60] and large-scale one-to-one pre-trained encoders [36, 53] with the help of massive image-text pairs crawled from the web. However, existing models are prone to return a list of retrieved images with similar semantics. Often, the queries submitted by ordinary users, especially those in the form of short keyword-based texts without concrete context [35, 47], very likely have broad semantics, and thus are semantically ambiguous and uncertain. For example, given the coarse-grained query “dog”, the user may expect dog images with diverse semantics (*e.g.* different breeds, colors, and body shapes).

\*Major part of this work was done while the author was an intern at Tencent.

† Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
*SIGIR '23, July 23–27, 2023, Taipei, Taiwan.*

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9408-6/23/07...\$15.00  
<https://doi.org/10.1145/3539618.3591705>



**Figure 1: Illustrations of (a) typical cross-modal image retrieval systems; (b) learning-based multi-vector retrieval systems; (c) our CoLT method. Red star represents the query. Points of different colors denote images of different semantics. Gray points represent irrelevant images, and triangles represent the prototypes of the corresponding semantics. Dotted circles denote the projection regions.**

Obviously, keyword-based queries are prone to match various retrieval results, but a list of images with similar semantics cannot meet the diverse requirements of different users, thus deteriorating their retrieval experience [43, 64].

To address the aforementioned drawback, the task of *keyword-based diverse image retrieval* [15, 17, 18, 20], is proposed, which takes a short keyword-based text as input to search a list of images with high relevance and rich semantic diversity. Recent approaches can be roughly divided into two groups. The first group is post-processing based approaches [10, 25, 33, 38–40, 40, 61]. These methods usually apply existing cross-modal encoders to extracting features. Then, various algorithms (e.g. re-ranking [61] and clustering [33]) are adopted to promote the diversity. However, these methods often cannot obtain a good retrieval list with balanced relevance and diversity, due to the limitations of *one-to-one projection*. For instance, as shown in Fig. 1(a), on the one hand, in typical one-to-one projection, (W1) the query feature (the red star) is likely to be surrounded by images of common semantics (the brown points) due to the long-tail distribution of the training data, which will make the top- $k$  result set full of images of similar semantics. On the other hand, (W2) image features with different semantics are less distinguishable because of the ignorance of modeling diversity [64], which will hurt the performance of some algorithms like clustering.

The second group is a set of learning-based approaches [1, 42, 43, 52, 64, 70] that try to use various techniques (e.g. graph [43], metric learning [5, 46] and multiple instance learning [55, 70]) to model the diversity. Compared with the one-to-one projection that projects each image to a vector in the latent space, these methods [42, 52, 64] embed each image (or text query) into multiple vectors around the relevant features to obtain their diverse representations for top- $k$  search, namely *multi-vector projection*. Unfortunately, such a projection is not robust enough and unable to handle images of rare

or unique semantics. As shown in Fig. 1(b), (W3) some irrelevant outliers (the grey points) will be mistakenly projected to represent diversity. Besides, (W4) some images of rare or unique semantics (the blue points), will very possibly be projected into some remote regions where the top- $k$  algorithm cannot reach.

To overcome the weaknesses (i.e., W1~W4) of the existing methods, in this paper we propose a novel approach called CoLT (the abbreviation of Semantics-aware Contrastive Learning and Transformer) for keyword-based image retrieval. In particular, to overcome W1, W2 and W3, CoLT extracts stable, representative and distinguishable image features with the help of a new *semantics-aware contrastive learning* (SCL) loss. As shown in Fig. 1(c), the core idea of SCL is to project images of similar semantics (e.g. dogs of the same breed) to vectors around their matched semantic prototype that keeps a proper distance from the other prototypes (e.g. dogs of different breeds and irrelevant images) and the query feature to better model the diversity. As for coping with images of rare semantics (W4), instead of utilizing top- $k$  algorithm as in existing works, CoLT employs a powerful *transformer-based token classifier* (TTC) to generate the final retrieval results. Specifically, in TTC the image and query features are concatenated as an input token sequence. Subsequently, TTC classifies each token into a relevant semantic category to distinguish the images of various semantics. Finally, a flexible post-processing algorithm is designed to select images from various semantic categories (both common and rare semantics), to form the final retrieval results. Such a design offers our method four-fold advantages: (i) *High semantic relevance*. CoLT improves the robust one-to-one projection of pre-trained cross-modal encoders, which is much more stable than recent multi-vector projection-based methods. (ii) *High semantic diversity*. CoLT not only makes the image features much more distinguishable via semantics-aware contrastive learning but also uses a transformer-based token classifier to mine rare semantics. (iii) *General and easy-to-use*. CoLT can be directly stacked at the end of various existing cross-modal encoders, without modifying their structures and parameters, and boost the performance in a plug-and-play manner. (iv) *Easy-to-control*. We can modify the post-processing algorithm in CoLT to flexibly balance semantic relevance and semantic diversity without re-implementing the model.

Contributions of this paper are summarized as follows: (1) We pinpoint the limitations of existing methods and present a novel approach called CoLT for keyword-based diverse image retrieval. CoLT first extracts high-quality and distinguishable semantics-aware features and then classifies the features to generate the final retrieval list. (2) We develop a semantics-aware contrastive loss in CoLT to extract more robust and representative features. (3) To better mine semantic diversity, we design a transformer-based token classifier to generate the retrieval results. (4) We conduct extensive experiments on two real-world datasets Div400 and Div150Cred, which show that our method can effectively boost the diversity, and outperforms the existing methods as a whole with a higher  $F1$  score.

**Table 1: A qualitative comparison between CoLT and major existing methods from three dimensions: feature projection, retrieval result generation and performance.**

Method	Projection	Generation	Performance
CLIP [36]	One-to-one	top- $k$	Low diversity
MMR [38]	One-to-one	Re-ranking	Low diversity
UMONS [40]	One-to-one	Clustering	Low relevance
VMIG [64]	Multi-vector	top- $k$	Medium relevance & diversity
CoLT (ours)	SCL	TTC	High relevance & diversity

## 2 RELATED WORK

### 2.1 Cross-Modal Image Retrieval

Typical cross-modal image retrieval methods [34] can be roughly divided into two categories: cross-modal similarity measurement based methods [44, 58, 65, 73] that directly calculate the cross-modal distance and common space learning-based methods [4, 31, 46, 56, 71] that map the query and images into a shared space via various techniques like attention mechanism and generative adversarial network etc. Nowadays, thanks to the transformer structure and pre-training techniques, large-scale pre-trained encoders (e.g. CLIP [36], ALIGN [22], GroupViT [53], and U-BERT [60]) have shown their superiority in relevance-based retrieval tasks. Although these methods have significantly improved the retrieval relevance, their ignorance of modeling semantic diversity hurts the semantic diversity of their retrieval lists.

### 2.2 Diverse Retrieval

Existing diverse retrieval approaches roughly fall into two groups. The first group is post-processing based methods [6, 10, 25, 33, 38–40, 40, 61], which usually use existing feature encoders [7, 8, 14, 26, 36] to generate features, then mine the diversity with a post-processing algorithm. Among them, [40] first filters irrelevant images, then clusters the rest via DBSCAN [9] to promote diversity. MMR [38] is proposed to re-rank the retrieval list to balance diversity and relevance. Bo and Gao [1] extract keywords to control the diversity of the results. The second group includes recently proposed learning-based methods, which aim to represent the semantic diversity in the latent space [1, 42, 43, 52, 64, 70]. In particular, Su et al. [43] propose a dynamic intent graph (GRAPH4DIV) to balance content and intent in a document. Song and Soleymani [42] utilize multiple transformers to extract visual features. Wu and Ngo [52] design an inactive word loss to expand the semantic concepts to represent various video contents. VMIG [64] embeds each image and text query into multiple vectors via multiple instance learning. Although these methods succeed in boosting the semantic diversity of the retrieval results, they perform unsatisfactorily in guaranteeing semantic relevance and mining images of rare semantics.

### 2.3 Differences between Our Method and Existing Works

To expound the differences between CoLT and typical existing methods, in Tab. 1 we present a qualitative comparison from three dimensions: how are images and queries projected? how are the

final retrieval results generated? and how is the performance in terms of both relevance and diversity? As presented in Tab. 1, recent pre-trained cross-modal encoders (e.g. CLIP [36]) cannot model semantic diversity well due to the limitations of the one-to-one projection. Two typical post-processing based methods MMR and UMONS are poor at either modeling diversity [38] due to the lack of an accurate diversity measurement mechanism or guaranteeing relevance due to clustering irrelevant features together. The recently proposed VMIG suffers from the robustness issue due to the uncertainty of multi-vector projection and the rare semantics handling issue caused by the top- $k$  search algorithm, which leads to undesirable performance. Our method CoLT is the only retrieval method that achieves both high semantic relevance and rich semantic diversity thanks to the proposed *semantics-aware contrastive learning* (SCL) and powerful *transformer-based token classifier* (TTC). Experiments and visualization studies demonstrate the advantages of our method.

## 3 METHODOLOGY

### 3.1 Overview

Given a text query  $Q$  and an image dataset  $\mathcal{D}$ , our aim is to generate a retrieval list  $\mathcal{R}$  that consists of  $K$  images of high semantic relevance and diversity. Fig. 2 shows the architecture of our method CoLT, which is composed of six components: a *fixed feature encoder*  $f$  that takes  $Q$  and  $\mathcal{D}$  as input to generate initial query feature  $h_q$  and visual features  $\{h_v^i\}$ , i.e.,  $\{h_q, \{h_v^i\}\} = f(Q, \mathcal{D})$ , a *visual feature re-encoder*  $g$  that re-encodes the visual features with the help of the *semantics-aware contrastive learning* (SCL) module, and the *transformer-based token classifier* (TTC)  $\phi$  that takes the query feature  $h_q$  and the re-encoded image features  $\hat{h}_v^i$  as an input token sequence to subsume each token into a suitable semantic category according to their representations. The TTC module consists of two sub-modules: the token classification transformer that is composed of  $L$  transformer encoder layers, and a fully-connected layer as the classifier. Finally, a *post-processing* module is adopted to select typical images from these categories as the final results. During training, a *token-wise data augmentation* module is used to make full use of the training data, which is employed between the SCL module and the TTC module.

### 3.2 Semantics-aware Contrastive Learning

In CoLT, we first use a fixed pre-trained feature encoder  $f$  to extract preliminary high-quality and robust visual features and query feature. Nevertheless, as mentioned above, these one-to-one projected features are not distinguishable enough to support effective diverse retrieval. Ergo, a visual feature re-encoder  $g$ , which is implemented by a multi-layer perception and powered by a novel semantics-aware contrastive learning is used to refine the image features to promote semantic diversity. In particular, for each visual feature  $h_v^i$ , we re-encode its representation as follows:

$$\hat{h}_v^i = h_v^i + \beta g(h_v^i), \quad (1)$$

where  $\beta$  is a hyper-parameter used to control the learned re-encoded feature  $g(h_v^i)$ . In what follows, we introduce the proposed semantics-aware contrastive loss in detail.

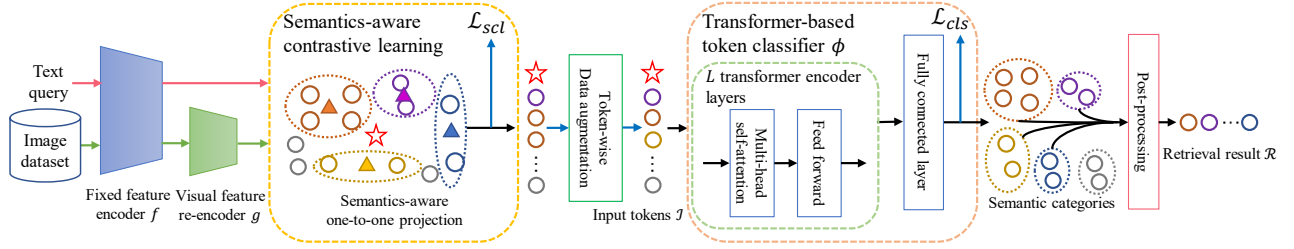


Figure 2: The architecture of CoLT. Blue lines are valid only during training.

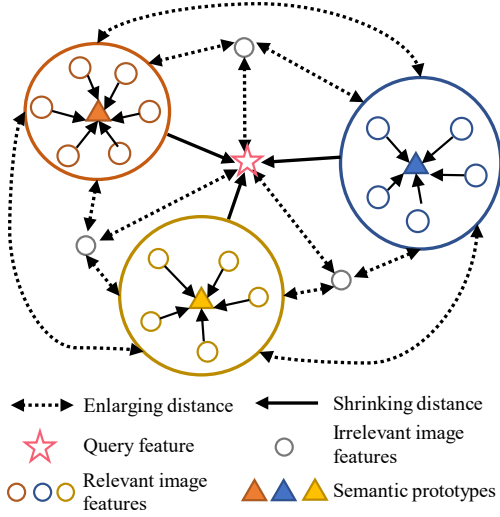


Figure 3: Illustration of semantics-aware contrastive learning. Different colors indicate various categories.

As shown in Fig. 3, the goal of semantics-aware contrastive learning (SCL) is: (1) Enlarging the distance between query feature and irrelevant features; (2) Enlarging the distance between relevant image features and irrelevant image features; (3) Enlarging the distance among image features of different semantics, which makes these features more distinguishable while benefiting diversity; (4) Shrinking the distance between the query feature and relevant image features, which can improve accuracy as (1) and (2); (5) Shrinking the distance among image features of similar semantics. In SCL, we use semantic category prototypes stored in a bank  $\mathcal{B}$  to efficiently compute (3) and (5), which can avoid inputting a large batch size. As a result, each image query will be projected to a position with suitable distance between the query feature, its matched semantic prototypes, unmatched semantic prototypes, and irrelevant image features.

Here we discuss the implementation of the proposed semantics-aware contrastive learning. In SCL, the positive pairs include (1) relevant image-query feature pairs and (2) relevant image-category prototype pairs, while the negative pairs are (3) irrelevant image-query feature pairs and (4) irrelevant image-category prototype pairs. For a query  $h_q$  with a set of relevant image features  $\{\hat{h}_v^{r,i}\}$  and a set of irrelevant image features  $\{\hat{h}_v^{ir,i}\}$ . Let  $\mathcal{B}(i)$  denotes the

$i$ -th semantic category prototype stored in the bank and  $G(\cdot)$  is a function that maps the image features to the corresponding indices of the matched semantic category prototypes, the loss of SCL can be formulated as follows:

$$\mathcal{L}_{scl} = -\log \frac{\underbrace{\sum_i \exp(h_q \cdot \hat{h}_v^{r,i} / \tau)}_{(1)} \cdot \underbrace{\sum_i \exp(\mathcal{B}(G(\hat{h}_v^{r,i})) \cdot \hat{h}_v^{r,i} / \tau)}_{(2)}}{\underbrace{\sum_i \exp(h_q \cdot \hat{h}_v^{ir,i} / \tau)}_{(3)} + \underbrace{\sum_{i,j} \exp(\mathcal{B}(j) \cdot \hat{h}_v^{ir,i} / \tau)}_{(4)} + (1) + (2)} \quad (2)$$

where  $\tau$  is a hyper-parameter used to control the temperature.

The category prototypes stored in the bank  $\mathcal{B}$  play an important role in the proposed SCL. Therefore, they need to be initialized and updated during training to obtain accurate and latest representations. Ergo, we use the fine-grained textual description features extracted by the fixed feature encoder to initialize the bank. As for update, exponential moving average (EMA) [11, 24, 68, 72] is utilized to update the category prototypes:

$$\mathcal{B}(G(\hat{h}_v^{r,i})) = \alpha \mathcal{B}(G(\hat{h}_v^{r,i})) + (1 - \alpha) \hat{h}_v^{r,i} \quad (3)$$

where  $\alpha$  is the momentum coefficient used to update the bank.

### 3.3 Transformer-based Token Classification

After obtaining a set of representation features, the next problem is how to generate the final result of high relevance and diversity. To this end, a powerful *transformer-based token classifier* (TTC) is developed to do feature fusion and token classification. Specifically, we treat each feature as a token, and concatenate the query feature  $h_q$  and  $N$  image features  $\{\hat{h}_v^i\}_{i=1}^N$  to form the input token sequence, i.e.,  $\mathcal{I} = [h_q, \{\hat{h}_v^i\}_{i=1}^N]$ . Here, to avoid irrelevant tokens, only  $N$  image features semantically most similar to the query feature are used. We sort the image features with respect to their cosine similarity with the query feature, and generate the corresponding ground truth  $\mathcal{Y} = \{y_i\}_{i=1}^{N+1}$  in terms of their fine-grained semantic annotations. It is worth mentioning that in TTC, all the irrelevant image features and the query feature are labeled with special indexes to further distinguish them. As shown in Fig. 2,  $L$  transformer encoder layers [45] powered by multi-head self-attention and feed forward layers are used to fuse these tokens. Subsequently, a fully connected layer is stacked as a classifier to do prediction. Formally, the predictions of TTC can be written as follows:

$$\{p_i\}_{i=1}^{N+1} = \phi(\mathcal{I}), \quad (4)$$

where  $p_i$  is the predicted distribution probability of the  $i$ -th token. Cross entropy loss is served as the classification loss to train TTC:

$$\mathcal{L}_{cls} = -\sum_i y_i \log(p_i). \quad (5)$$

After classifying each image token into an appropriate category, a post-processing algorithm  $t$  is applied to generate the final retrieval list. That is, selecting  $X$  images with the highest similarity to the query feature from each semantic category:

$$\mathcal{R} = t(\{p_i\}_{i=1}^{N+1}, X). \quad (6)$$

Finally, after selecting images from  $\lfloor k/X \rfloor$  semantic categories, a retrieval list  $\mathcal{R}$  of length  $k$  is obtained.

### 3.4 Token-wise Data Augmentation

Due to the lack of specific fine-grained annotations, directly training our model on  $\mathcal{D}$  is prone to over-fitting. Therefore, to better exploit the potential of the transformer-based token classification module, we employ token-wise data augmentation to enrich the input tokens  $\mathcal{I}$ . In particular, four different kernel operations are introduced:

**Query perturbation:** We perturb the query feature  $h_q$  as follows: MIXUP [67] the query feature with a relevant image feature  $\hat{h}_v^{r,i}$  with a probability of  $p_q$ . Formally, let  $\lambda \sim \text{Beta}(1.0, 1.0)$ , we generate the perturbed query feature as follows:

$$h_q = \max(\lambda, 1.0 - \lambda)h_q + \min(\lambda, 1.0 - \lambda)\hat{h}_v^{r,i}. \quad (7)$$

**Image perturbation:** We perturb the image feature  $\hat{h}_v^{r,i}$  as follows: MIXUP the image feature with a relevant query feature  $h_q$  with a probability of  $p_v$ . By sampling  $\lambda$  from  $\text{Beta}(1.0, 1.0)$ , we have

$$\hat{h}_v^{r,i} = \max(\lambda, 1.0 - \lambda)\hat{h}_v^{r,i} + \min(\lambda, 1.0 - \lambda)h_q. \quad (8)$$

**Deletion:** We delete an image feature with a probability of  $p_d$ .

**Copy:** We copy an image feature with a probability of  $p_c$ .

Among the 4 operations, query perturbation and image perturbation directly augment the features without modifying the semantics-aware representations, which is beneficial to the robustness of the model while the operations of deletion and copy can enhance the model's ability of distinguishing rare and similar tokens, respectively. Following the experience in [51], we perform data augmentation to the input tokens  $\mathcal{I}$  in such a manner: sampling each data augmentation operation in the following order: (1) query perturbation; (2) deletion; (3) copy; (4) image perturbation, then individually performing the selected operation on each token.

### 3.5 Training and Evaluation Algorithms

The training procedure of CoLT is presented in Alg. 1, which can be divided into three steps. First, the initial query feature and image features are extracted by  $f$  (L2). Then, we train the visual feature re-encoder by the proposed semantics-aware contrastive learning (L3-L9) to re-encode the preliminary features to semantics-aware ones. Finally, we take the query feature and the re-encoded image features as input to train the transformer-based token classifier with the help of token-wise data augmentation (L10-L16).

The evaluation procedure is given in Alg. 2, which is like this: we first generate the initial features (L2), then re-encode the image features (L3). Subsequently, take these features as tokens to generate the predicted distribution probabilities (L4-L5). Finally, using the post-processing algorithm  $t$  to generate the final retrieval list  $\mathcal{R}$ .

---

#### Algorithm 1 The training of CoLT.

---

- 1: **Input:** Fixed feature encoder  $f$ , visual feature re-encoder  $g$ , transformer-based token classifier  $\phi$ , query  $Q$ , and image dataset  $\mathcal{D}$
  - 2:  $h_q, \{h_v^i\} = f(Q, \mathcal{D})$
  - 3: initialize  $\mathcal{B}$  with fine-grained description
  - 4: **while**  $g$  is not converged **do**
  - 5:      $\hat{h}_v^i = h_v^i + \beta g(h_v^i)$
  - 6:      $\hat{h}_v^{r,i}, \hat{h}_v^{ir,i} \sim g(h_v^i)$
  - 7:     Compute  $\mathcal{L}_{scl}$  via Eq. (2)
  - 8:     Optimize  $g$  according to  $\mathcal{L}_{scl}$
  - 9:     Update  $\mathcal{B}$  via Eq. (3)
  - 10: **while**  $\phi$  is not converged **do**
  - 11:      $\mathcal{I} = [h_q, \{\hat{h}_v^i\}_{i=1}^N]$
  - 12:     Perform data augmentation to  $\mathcal{I}$  according to Sec. 3.4
  - 13:     Obtain the final  $\mathcal{I}$  according to Sec. 3.3
  - 14:      $\{p_i\}_{i=1}^{N+1} = \phi(\mathcal{I})$
  - 15:     Compute  $\mathcal{L}_{cls}$  via Eq. (5)
  - 16:     Optimize  $\phi$  according to  $\mathcal{L}_{cls}$
  - 17: **return**  $g$  and  $\phi$
- 

---

#### Algorithm 2 The evaluation of CoLT.

---

- 1: **Input:** Fixed feature encoder  $f$ , visual feature re-encoder  $g$ , transformer-based token classifier  $\phi$ , query  $Q$ , image dataset  $\mathcal{D}$ , and post-processing algorithm  $t$  with its hyper-parameter  $X$
  - 2:  $h_q, \{h_v^i\} = f(Q, \mathcal{D})$
  - 3:  $\hat{h}_v^i = h_v^i + \beta g(h_v^i)$
  - 4:  $\mathcal{I} = [h_q, \{\hat{h}_v^i\}_{i=1}^N]$
  - 5:  $\{p_i\}_{i=1}^{N+1} = \phi(\mathcal{I})$
  - 6:  $\mathcal{R} = t(\{p_i\}_{i=1}^{N+1}, X)$
  - 7: **return**  $\mathcal{R}$
- 

## 4 PERFORMANCE EVALUATION

### 4.1 Research Questions

In this section, we evaluate the proposed method by conducting extensive experiments to answer the following research questions:

**RQ1:** How does CoLT perform in comparison with the state-of-the-art cross-modal image retrieval models in terms of both relevance and diversity?

**RQ2:** Can the proposed semantics-aware contrastive learning and the transformer-based token classifier effectively boost relevance and diversity?

**RQ3:** How do different components/parameters contribute to the effectiveness of CoLT?

### 4.2 Datasets and Metrics

Here we briefly summarize the datasets and metrics used in our paper. More details can be referred to [19]. Two datasets are used in our paper:

**Table 2: Performance comparison with the state-of-the-art methods on Div400. P@k and CR@k are evaluation metrics for relevance and diversity, respectively. F1@k evaluates the overall performance regarding both relevance and diversity.**

Method	P@10	P@20	CR@10	CR@20	F1@10	F1@20
IMRAM [5]	79.22%	76.98%	28.93%	33.79%	42.38%	46.96%
FCA-Net [13]	79.98%	78.42%	29.91%	34.90%	43.54%	48.30%
CLIP [36]	<b>90.17%</b>	<b>87.92%</b>	35.68%	52.97%	51.10%	65.60%
MMR [38]	<u>86.79%</u>	<u>84.52%</u>	36.96%	53.88%	51.85%	65.81%
UMONS [40]	79.28%	73.37%	<u>44.71%</u>	<u>63.24%</u>	<u>57.17%</u>	67.93%
DESA [35]	76.50%	73.82%	39.47%	52.82%	52.07%	61.58%
Su et al. [43]	77.55%	74.50%	35.66%	45.77%	48.85%	56.70%
VMIG [64]	82.31%	83.01%	40.37%	59.46%	54.17%	<u>69.28%</u>
CoLT (ours)	84.45%	<u>85.48%</u>	<b>47.06%</b>	<b>64.16%</b>	<b>60.44%</b>	<b>73.30%</b>

**Div400:** Div400<sup>1</sup> is collected by the MediaEval Workshop [18]. It contains 396 queries with 43,418 images. All queries are mainly related to tourist locations and the average length of queries is 3.7 words. On average, the ground truth of a query covers 11.8 semantic categories of images in the dataset. Each image has a coarse-grained textual description (e.g. “Big Ben”) and a fine-grained one (e.g. “partial view”).

**Div150Cred:** Div150Cred<sup>2</sup> is derived from the competition dataset for diverse social image retrieval in 2014 [16]. It has a total of 153 queries with 45,375 images. The ground truth of a query averagely covers 22.6 semantic categories of images in the dataset.

Three metrics are used to evaluate the performance, including *precision (P)* for measuring semantic relevance, *cluster recall (CR)* for measuring semantic diversity, and the *F1 score* of *P* and *CR* to measure the overall balanced performance. Specifically, we calculate the evaluation metrics of the top-*k* results, where *k* is set to 10 and 20 by following [64]. In the rest of this paper, we use P@k, CR@k, and F1@k to denote the *P*, *CR*, and *F1* value of the top-*k* results, respectively. Higher P@k indicates better relevance, and higher CR@k means richer semantic diversity.

### 4.3 Implementation Details

CoLT is implemented in PyTorch-1.10. All experiments are conducted on 4 NVIDIA 3090 GPUs with 24GB memory. The model is trained using the Adam [23] optimizer with a learning rate of  $10^{-5}$  for the visual feature re-encoder *g* and  $10^{-4}$  for the transformer-base token classifier  $\phi$ . The batch size is set to 32.  $\tau$ ,  $\alpha$ ,  $\beta$ , and  $\epsilon$  are set to small values:  $\tau = 0.2$ ,  $\alpha = 0.01$ ,  $\beta = 0.02$ , and  $\epsilon = 0.01$  by following [11, 24, 69, 72]. *X*, *N*, and *L* are set to 1, 200, and 8 through ablation study. The probabilities used for data augmentation are set by following [51]. In particular, we have  $p_q = 0.5$ ,  $p_o = 0.2$ ,  $p_d = 0.2$  and  $p_c = 0.2$ . All different semantic categories (or simply semantics) in each dataset are stored as prototypes. As a result, we store 629 prototypes for Div400 dataset while 725 for Div150Cred dataset.

### 4.4 Comparing with SOTA Methods (RQ1)

To demonstrate the effectiveness of our method CoLT, we compare it with several state-of-the-art approaches, including three typical cross-modal image retrieval methods: IMRAM [5], FCA-Net [13]

**Table 3: Performance comparison with the state-of-the-art methods on Div150Cred.**

Method	P@10	P@20	CR@10	CR@20	F1@10	F1@20
CLIP [36]	<b>96.02%</b>	<b>95.04%</b>	23.48%	35.32%	37.73%	51.51%
MMR [38]	<u>95.37%</u>	<u>94.23%</u>	23.50%	35.49%	37.71%	51.56%
UMONS [40]	77.40%	84.15%	<u>26.69%</u>	<b>40.10%</b>	<u>39.69%</u>	<u>54.32%</u>
VMIG [64]	90.81%	89.96%	23.83%	37.97%	37.75%	53.40%
CoLT (ours)	93.41%	<u>94.39%</u>	<b>27.53%</b>	<u>39.30%</u>	<b>42.52%</b>	<b>55.49%</b>

and CLIP [36], two post-processing-based diverse retrieval methods: MMR [38] and UMONS [40], and three learning-based diverse retrieval approaches: DESA [35], GRAPH4DIV [43] and VMIG [64]. Since MMR, UMONS and VMIG require a feature encoder, for fairness, we use the same feature encoder CLIP [36] to implement them and our method CoLT. For MMR and UMONS, we use grid search to obtain their best results. Generally, our results are higher than those in the original papers thanks to the strong feature encoder. For example, the P@20, CR@20, and F1@20 values of VMIG on the DIV400 dataset are lifted from 78.27%, 59.01% and 67.29% to 83.01%, 59.46% and 69.28%. Experimental results on Div400 and Div150Cred are given in Tab. 2 and Tab. 3, respectively. Here, the best values are bolded while the second-best results are underlined.

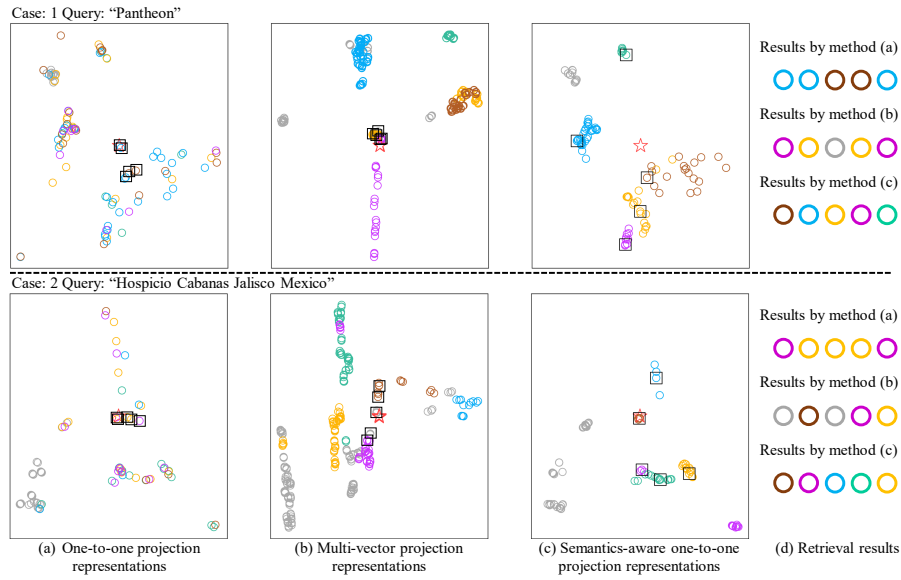
From Tab. 2 and Tab. 3, we can see that 1) typical cross-modal image retrieval methods including large-scale pre-trained encoder CLIP perform well in relevance-based retrieval but cannot do diverse retrieval well. For example, although CLIP achieves the best relevance performance, it is inferior to the others in diversity score. 2) Post-processing-based approaches can only moderately trade-off accuracy and diversity. For example, as can be seen in Tab. 2, the diversity improvement achieved by MMR is very limited (CR@10 increases from 35.68% to 36.96% on Div400). As for UMONS, its accuracy score is greatly degraded (P@10 decreases from 90.17% to 79.28% on Div400) though it obtains a relatively large diversity improvement. As a result, their *F1* scores are not satisfactory. 3) Recently proposed learning-based methods achieve balanced relevance and diversity scores. For instance, VMIG outperforms most existing methods in CR@10 and CR@20, and performs better than UMONS in relevance score. However, its relevance and diversity are both limited due to the weaknesses of the multi-vector projection. 4) Our method CoLT obtains the best diversity score, high precision, and obviously the highest overall *F1* score on both Div400 and Div150Cred. In particular, CoLT outperforms CLIP and VMIG by significant margins, i.e., 7.70% and 4.02% of F1@20 on Div400, respectively. This indicates that CoLT is able to get retrieval results of both high relevance and rich semantic diversity. Besides, we present a variant of CoLT that outperforms CLIP on both relevance and diversity, we will discuss the details in Sec. 4.6.

### 4.5 Visualization (RQ2)

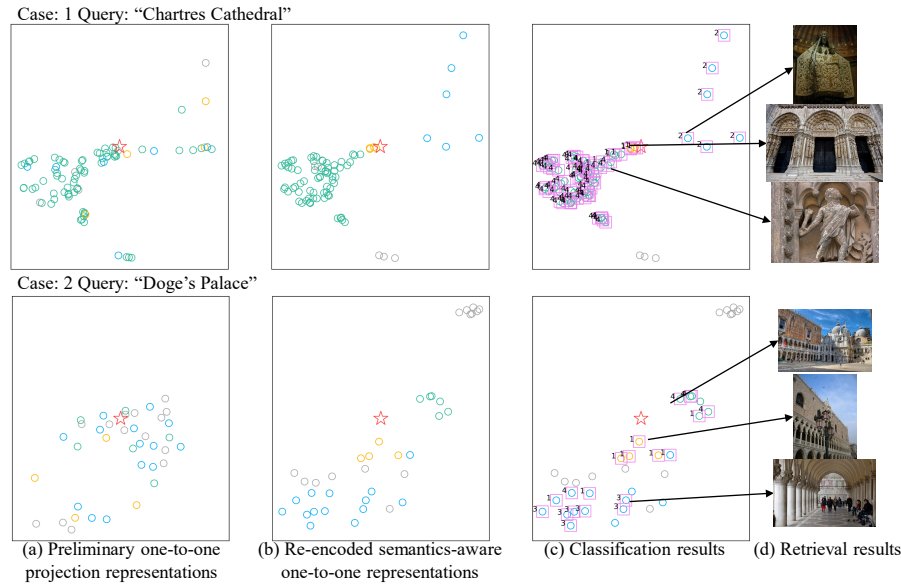
To better demonstrate the advantages of the proposed techniques and study how they lift the performance, we visualize some cases in the test set of the DIV400 dataset via UMAP [32]. The visualization comparison between our semantics-aware representation and existing methods’ representations are illustrated in Fig. 4.

<sup>1</sup><http://multimediaeval.org/mediaeval2014/diverseimages2014>

<sup>2</sup><http://campus.pub.ro/lab7/bionescu/Div150Cred.html>



**Figure 4: Visualization comparison of different representations. (a) One-to-one projection (OOP) representations generated by the cross-modal encoder  $f$ . (b) Multi-vector projection (MVP) representations. (c) Semantics-aware one-to-one projection (SA-OOP) representations re-encoded by  $g$ . Retrieved images are marked by black square. (d) The final retrieval results generated by different methods.**



**Figure 5: The visualization of CoLT results. (a) One-to-one projection (OOP) representations generated by a cross-modal encoder  $f$ . (b) Semantics-aware one-to-one projection (SA-OOP) representations generated by the re-encoder  $g$ . (c) Classification results of TTC  $\phi$ . To make the figures clear, irrelevant images are not marked. The numbers around the boxes are the category ID predicted by TTC. (d) The final retrieval results obtained by our post-processing algorithm.**

From Fig. 4(a), we can see that the preliminary OOP representations extracted by CLIP can distinguish some irrelevant images. However, its weaknesses are also evident: (1) The query is closer to some image features of common semantics (the blue and brown

points in the 1st case); (2) Images of various semantics are mixed. As a result, such representations are not suitable for mining diversity. Then, let us pay attention to multi-vector projection (MVP). As can be seen from Fig. 4(b), each image and query are projected

**Table 4: Ablation study of CoLT on Div400.**

ID	Variant	P@20	CR@20	F1@20
1	without SCL&TTC	87.92%	52.97%	65.60%
2	SCL + TTC	85.48%	64.16%	73.30%
3	without SCL	84.26%	62.94%	72.06%
4	UMONS	73.37%	63.24%	67.93%
5	SCL + DBSCAN	75.94%	63.90%	69.40%
6	SCL + top- $k$	89.06%	89.93%	67.79%
7	without DA	86.34%	62.19%	72.30%
8	unfixed $f$	76.52%	51.25%	61.38%

**Table 5: Effect of pair construction in SCL.**

Variant	P@20	CR@20	F1@20
All 4 pairs	85.48%	<b>64.16%</b>	<b>73.30%</b>
Without pair (2)	85.04%	63.78%	72.89%
Without pair (4)	<b>85.72%</b>	62.17%	72.07%

into multiple points to enrich diversity. However, on the one hand, some outliers are mistakenly projected into the neighborhood of the query feature to represent diversity (a grey point in the 1st case while two in the 2nd case). On the other hand, some image features of rare semantics are projected into remote regions (the green points in the 1st case) where the top- $k$  algorithm cannot reach. Thus, as shown in Fig. 4(d), some irrelevant images are selected while some images of rare semantics are not retrieved. Finally, we check the representations of our SCL and the images retrieved by TTC. From Fig. 4(c) we can see that (1) the representations of images of the same semantics are clustered and much more distinguishable compared with the typical OOP representations in Fig. 4(a); (2) Some irrelevant images are also pushed away. For instance, in the 2nd case, some grey points are pushed to the left-bottom corner. This demonstrates the advantages and effectiveness of our SCL. Then, TTC and a post-processing are employed to classify and select images from each category, including rare semantic categories like green points in the 1st case, to form the final results.

We also visualize the classification results of TTC to further demonstrate its effect. The visualization is shown in Fig. 5, from which we can see that (1) TTC is able to distinguish different semantics and irrelevant images. Taking the 1st case for example, the yellow points are classified into the 1st category, the majority of the green points are subsumed into the 4th category, and blue points to the 2nd category. Irrelevant images are also correctly classified. This demonstrates the effectiveness of the proposed TTC. (2) The classification performance of TTC can be further improved. For example, as can be seen in the 2nd case, TTC mistakenly classifies one of the green points into the 1st category. In summary, the power of TTC is demonstrated well via visualization.

#### 4.6 Ablation Study (RQ3)

Here we conduct ablation study on Div400 to demonstrate the contributions of different modules and the effect of some parameters in our method. The metrics P@20, CR@20 and F1@20 are used. Results are presented in from Tab. 4 to Tab. 10.

**Overall performance improvement.** As shown in the 1st row and 2nd row in Tab. 4, our method significantly boosts the diversity score and  $F1$  score from 52.97% to 64.16% and 65.60% and 73.30%,

**Table 6: Performance when using different feature encoders.**

Variant	P@20	CR@20	F1@20
ViT+BERT	87.75%	52.39%	65.60%
+CoLT	85.48%	64.16%	73.30%
R50+BERT	87.01%	52.10%	65.17%
+CoLT	85.62%	60.83%	72.94%
GroupViT	83.68%	52.02%	64.21%
+CoLT	82.62%	70.91%	70.91%

respectively, with only a slight decrease in relevance score. This supports the superiority of our method.

**Effect of SCL.** Here we check the effect of the proposed SCL. Specifically, we first design a variant that removes SCL and directly applies TTC. Obviously, as can be seen in the 2nd row and the 3rd row of Tab. 4, all metrics including relevance, diversity, and  $F1$  score are degraded. Besides, we also design a variant that combines SCL with the idea of UMONS to generate the final retrieval results via DBSCAN. Comparing the results of the 4th row and the 5th row, the performance with SCL is better than that of the original UMONS. The reason lies in that SCL is able to make the image features more distinguishable, and such representations are more suitable for existing post-processing schemes.

Then, we check the effect of the constructed pairs in SCL. As mentioned in Sec. 3.2, SCL uses 4 kinds of pairs. Among these pairs, (1) and (3) are common in contrastive learning [36] to align images and queries while (2) and (4) play important roles in distinguishing images of various semantics. Ergo, we remove (2) and (4) separately to examine their influence. As can be seen in Tab. 5, without pair (2) and pair (4), diversity score and  $F1$  score are degraded. On the contrary, their influence on relevance score is minor. This justifies the effectiveness of SCL – making the representations much more distinguishable for promoting diversity.

**Effect of TTC.** To check the effect of the proposed transformer-based token classifier, we design two variants that replace TTC by DSCAN (the 5th row of Tab. 4) or top- $k$  (the 6th row of Tab. 4) to generate the retrieval results. Obviously, such variants are inferior to our method (the 2nd row of Tab. 4). This demonstrates the advantage of TTC.

**Effect of token-wise data augmentation.** Here we implement a variant that removes the token-wise data augmentation module. Results of this variant are given in the 7th row of Tab. 4. Evidently, the resulting performance is inferior to ours (the 2nd row of Tab. 4).

**Why fix the cross-modal feature encoder?** In CoLT, we fix the cross-modal feature encoder  $f$  to better maintain the pre-trained knowledge. To support this design, we implement a variant that finetunes the feature encoder  $f$ . Experimental results are given in the 8th row of Tab. 4. Obviously, all metrics including relevance, diversity and  $F1$  are significantly degraded, comparing with ours (the 2nd row of Tab. 4). Possibly, finetuning too many parameters is prone to over-fitting.

**Can CoLT support various feature encoders?** As mentioned above, CoLT is general, i.e., it can work with various cross-modal encoders to do diverse image retrieval. To verify this point, we try three different encoder configurations, including ViT [8] and BERT [7] developed by [36], R50 [14] and BERT [7] implemented



**Table 7: Performance vs. the number of images selected from each category.**

$X$	P@20	CR@20	F1@20
1	85.48%	64.16%	73.30%
2	88.09%	58.54%	70.34%
3	88.45%	57.28%	69.54%

**Table 8: Time cost comparison. We report the result of one-time retrieval on a 3090 GPU.**

	CLIP	MMR	UMONS	VMIG	CoLT (Ours)
Time (ms)	18.06	24.10	22.65	86.77	30.23

**Table 9: Time cost comparison among major components. We report the result of one time retrieval on a 3090 GPU.  $f$  and  $g$  are tested in a parallel manner.**

	$f$	$g$	$\phi$
Time (ms)	18.06	0.37	11.80

by [36], and the encoders proposed in GroupViT [53]. The experimental results are given in Tab. 6, from which we can see that (1) all these pre-trained cross-modal encoders are good at relevance-based retrieval but perform poorly in terms of CR@20; (2) After applying our method CoLT, the diversity score is significantly boosted, with only slight decrease in precision. As a result, superior F1 score is achieved. This validates that CoLT can work well with various feature encoders to boost performance.

**Can CoLT flexibly balance accuracy and diversity?** As mentioned above, we can flexibly trade-off the relevance and diversity of the retrieval results without modifying network parameters. This is achieved by controlling the hyper-parameter  $X$ . As described in Sec. 3.3, the post-processing algorithm will select  $X$  images from each semantic category to form a retrieval list  $\mathcal{R}$  of length  $k$ . Thus, a smaller  $X$  will select fewer images from each category but can include more different categories (estimated by  $\lfloor k/X \rfloor$ ), which will benefit the diversity of the retrieval list  $\mathcal{R}$  but may hurt the relevance since classification accuracy on rare semantic categories is poor. On the contrary, a larger  $X$ , i.e., selecting more images from each category of common semantics will benefit the accuracy but limit the semantic diversity since fewer categories are exploited. We present the experimental results of how  $X$  impacts performance in Tab. 7. We can see that the best diversity is achieved when  $X = 1$  while the best accuracy is obtained when  $X=3$ . This indicates that CoLT can meet various retrieval settings, which demonstrates the flexibility of our approach. In this paper, we set  $X=1$  by default to obtain the best diversity and F1 score.

**Time cost.** We first compare the time cost of our method with that of various SOTA methods. The experimental results are given in Tab. 8. On the one hand, our method CoLT is 2.87 $\times$  faster than the state-of-the-art learning-based method VMIG. On the other hand, our method consumes moderately more time than the post-processing-based methods. For example, CoLT takes 6.23ms more than MMR. This justifies the efficiency of our method.

Then, we further check the time cost of each major module in CoLT: the fixed feature encoder  $f$ , the visual feature re-encoder  $g$ ,

**Table 10: The effect of parameter  $L$  in TTC.**

$L$	P@20	CR@20	F1@20	Time (ms)
6	85.82%	61.34%	71.54%	8.93
8	85.48%	64.16%	73.30%	11.80
10	85.29%	61.67%	71.58%	18.56

and TTC  $\phi$ . The experimental results are given in Tab. 9. We can see that  $g$  and  $\phi$  incur much less time than the feature encoder  $f$ . The reason lies in that  $g$  is a simple multi-layer perceptron while  $\phi$  consists of multiple transformer encoder layers that can run in parallel. It is worthy of mentioning that the image features generated by  $f$  and  $g$  can be cached offline in application. Hence, the main cost is from TTC  $\phi$ , which is very limited (11.80ms according to Tab. 9). This also verifies the efficiency of our method.

**Effect of the parameter  $L$ .** Here we study the effect of the number of transformer layers  $L$ . On the one hand, a larger  $L$  may result in over-fitting at a higher probability due to the limited training data. On the other hand, a smaller  $L$  cannot fully exploit the potential of TTC. Therefore, we conduct a grid search to determine the value of  $L$ . As can be seen in Tab. 10, the best performance is achieved when  $L=8$ .

**Effect of the parameter  $N$ .** Here we check how the number of images  $N$  fed to the transformer-based token classifier  $\phi$  impacts the performance. Intuitively, a large  $N$  will include images with more semantics. On the other hand, a large  $N$  will introduce more irrelevant images that may make token classification more difficult. On the contrary, a small  $N$  includes less irrelevant images but also fewer semantics. Therefore, both small  $N$  and large  $N$  are not appropriate for TTC. We conduct grid search to determine  $N$  on two datasets. Based on our results, we set  $N$  to 200 for the DIV400 dataset because the F1@20 scores of  $N = 150$  and  $N = 250$  are 72.10% and 72.13%, which is inferior to that of  $N = 200$  where F1@20 is 73.30%. While on the DIV150Cred dataset, the best performance is achieved when  $N = 200$  (an F1 of 55.49%) and 250 (an F1 of 55.32%). Ergo, we set this hyper-parameter to 200.

## 5 CONCLUSION

In this paper, we address keyword-based diverse image retrieval and propose a new method called Semantics-aware Classification Transformer (CoLT) to do this task. Different from existing works, CoLT first extracts highly representative images and query features via semantics-aware contrastive learning, then a transformer-based token classifier is employed to fuse these features and subsume them into their appropriate categories. Finally, a post-processing algorithm is applied to flexibly selecting images from each category to form the retrieval results. The advantages of CoLT are four-fold: *high semantic relevance*, *high semantic diversity*, *general* and *easy-to-use*, and *easy-to-control*. Extensive experiments on two datasets Div400 and Div150Cred demonstrate the superiority of our method.

## ACKNOWLEDGMENTS

Minyi Zhao was supported in part by the 2022 Tencent Rhino-Bird Research Elite Training Program. Shuigeng Zhou was supported by National Key R&D Program of China under grant No. 2021YFC3340302.

## REFERENCES

- [1] Yuan Bo and Xinbo Gao. 2019. Diversified textual features based image retrieval. *Neurocomputing* 357 (2019), 116–124.
- [2] Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural Machine Translation with Monolingual Translation Memory. In *ACL*. ACL, 7307–7318.
- [3] Da Cao, Yawen Zeng, Xiaochi Wei, Liqiang Nie, Richang Hong, and Zeng Qin. 2020. Adversarial Video Moment Retrieval by Jointly Modeling Ranking and Localization. In *Proceedings of the ACM MM*. ACM, 898–906.
- [4] Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. 2019. Cross-modal image-text retrieval with semantic consistency. In *Proceedings of the ACM MM*. ACM, 1749–1757.
- [5] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. IMRAM: Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. In *Proceedings of the CVPR*. IEEE, 12652–12660.
- [6] Laming Chen, Guoxin Zhang, and Eric Zhou. 2018. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in Neural Information Processing Systems* 31 (2018).
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiuhua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *kdd*, Vol. 96. 226–231.
- [10] Ifrah Gamzu, Marina Haikin, and Nissim Halabi. 2020. Query Rewriting for Voice Shopping Null Queries. In *Proceedings of the SIGIR*. ACM, 1369–1378.
- [11] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems* 33 (2020), 11309–11321.
- [12] Yun Gu, Khushi Vyas, Mali Shen, Jie Yang, and Guang-Zhong Yang. 2021. Deep Graph-Based Multimodal Feature Embedding for Endomicroscopy Image Retrieval. *IEEE Trans. Neural Networks Learn. Syst.* 33, 2 (2021), 481–492.
- [13] Ning Han, Jingjing Chen, Guangyi Xiao, Zhang Hao, Yawen Zeng, and Hao Chen. 2021. Fine-grained Cross-modal Alignment Network for Text-Video Retrieval. In *Proceedings of the ACM MM*. ACM, 3826–3834.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the CVPR*. IEEE, 770–778.
- [15] Bogdan Ionescu, Alexandru-Lucian Ginsca, Maia Zaharieva, Bogdan Boteanu, Mihai Lupu, and Henning Müller. 2016. Retrieving Diverse Social Images at MediaEval 2016: Challenge, Dataset and Evaluation. In *Working Notes Proceedings of the MediaEval 2016 Workshop*, Vol. 1739. CEUR-WS.org.
- [16] Bogdan Ionescu, Adrian Popescu, Mihai Lupu, Alexandru Lucian Ginsca, Bogdan Boteanu, and Henning Müller. 2015. Div150cred: A social image retrieval result diversification with user tagging credibility dataset. In *Proceedings of the 6th ACM Multimedia Systems Conference*. 207–212.
- [17] Bogdan Ionescu, Adrian Popescu, Anca-Livia Radu, and Henning Müller. 2016. Result diversification in social image retrieval: a benchmarking framework. *Multimedia Tools and Applications* 75, 2 (2016), 1301–1331.
- [18] Bogdan Ionescu, Anca-Livia Radu, Maria Menéndez, Henning Müller, Adrian Popescu, and Loni Babak. 2014. Div400: a social image retrieval result diversification dataset. In *Multimedia Systems Conference 2014*. ACM, 29–34.
- [19] Bogdan Ionescu, Maia Rohm, Bogdan Boteanu, Alexandru-Lucian Ginsca, Mihai Lupu, and Henning Müller. 2021. Benchmarking Image Retrieval Diversification Techniques for Social Media. *IEEE Trans. Multim.* 23 (2021), 677–691.
- [20] Bogdan Ionescu, Maia Rohm, Bogdan Boteanu, Alexandru Lucian Ginsca, Mihai Lupu, and Henning Müller. 2020. Benchmarking Image Retrieval Diversification Techniques for Social Media. *IEEE Transactions on Multimedia* 23 (2020), 677–691.
- [21] Zhong Ji, Yuxin Sun, Yunlong Yu, Yanwei Pang, and Jungong Han. 2020. Attribute-Guided Network for Cross-Modal Zero-Shot Hashing. *IEEE Trans. Neural Networks Learn. Syst.* 31, 1 (2020), 321–330.
- [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 *International Conference on Machine Learning*. PMLR, 4904–4916.
- [23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [24] Frank Klinker. 2011. Exponential moving average versus moving exponential average. *Mathematische Semesterberichte* 58, 1 (2011), 97–107.
- [25] Saar Kuzi, Abhishek Narwekar, Anusri Pampari, and ChengXiang Zhai. 2019. Help Me Search: Leveraging User-System Collaboration for Query Construction to Improve Accuracy for Difficult Queries. In *Proceedings of the SIGIR*. ACM, 1221–1224.
- [26] V. Quoc Le and Tomás Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the ICML*. JMLR.org, 1188–1196.
- [27] Dongliang Liao, Jin Xu, Gongfu Li, Huang Weijie, Liu Weiqing, and Li Jing. 2019. Popularity Prediction on Online Articles with Deep Fusion of Temporal Process and Content Features. In *Proceedings of the AAAI*. AAAI, 200–207.
- [28] Haoliang Liu, Tan Yu, and Ping Li. 2021. Inflate and shrink: Enriching and reducing interactions for fast text-image retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9796–9809.
- [29] Shenglan Liu, Muxin Sun, Lin Feng, Hong Qiao, Shuyuan Chen, and Yang Liu. 2021. Social Neighborhood Graph and Multigraph Fusion Ranking for Multi-feature Image Retrieval. *IEEE Trans. Neural Networks Learn. Syst.* 32, 3 (2021), 1389–1399.
- [30] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2125–2134.
- [31] Minnan Luo, Xiaojun Chang, Zhihui Li, Liqiang Nie, Alexander G. Hauptmann, and Qinghua Zheng. 2017. Simple to complex cross-modal learning to rank. *Computer Vision and Image Understanding* (2017), 67–77.
- [32] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
- [33] Liang Peng, Yi Bin, Xiyao Fu, Jie Zhou, Yang Yang, and Heng Tao Shen. 2017. CFM@MediaEval 2017 Retrieving Diverse Social Images Task via Re-ranking and Hierarchical Clustering. In *Proceedings of the Working Notes Proceedings of the MediaEval 2017 Workshop*, Vol. 1984.
- [34] Yuxin Peng, Xin Huang, and Yunzhen Zhao. 2018. An Overview of Cross-Media Retrieval: Concepts, Methodologies, Benchmarks, and Challenges. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 9 (2018), 2372–2385.
- [35] Xubo Qin, Zhicheng Dou, and Ji-Rong Wen. 2020. Diversifying Search Results using Self-Attention Network. In *Proceedings of the CIKM*. ACM, 1265–1274.
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
- [37] Nikhil Rasiwasia, José Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A New Approach to Cross-Modal Multimedia Retrieval. In *Proceedings of the ACM MM*. ACM, 251–260.
- [38] Jean-Michel Renders and Gabriela Csurka. 2017. NLE@MediaEval'17: Combining Cross-Media Similarity and Embeddings for Retrieving Diverse Social Images. In *Proceedings of the Working Notes Proceedings of the MediaEval 2017 Workshop*, Vol. 1984.
- [39] Mustafa Ilker Sarac and Pinar Duyugulu. 2014. Bilkent-RETINA at Retrieving Diverse Social Images Task of MediaEval 2014. In *Proceedings of the Working Notes Proceedings of the MediaEval 2014 Workshop*, Vol. 1263.
- [40] Omar Seddati, Nada Ben-Lhachemi, Stéphane Dupont, and Said Mahmoudi. 2017. UMONS @ MediaEval 2017: Diverse Social Images Retrieval. In *Proceedings of the Working Notes Proceedings of the MediaEval 2017 Workshop*, Vol. 1984.
- [41] Amit Singhal et al. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* 24, 4 (2001), 35–43.
- [42] Yale Song and Mohammad Soleymani. 2019. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. In *Proceedings of the CVPR*. IEEE, 1979–1988.
- [43] Zhan Su, Zhicheng Dou, Yutao Zhu, Xubo Qin, and Ji-Rong Wen. 2021. Modeling Intent Graph for Search Result Diversification. In *Proceedings of the SIGIR*. ACM, 736–746.
- [44] Hanghang Tong, Jingrui He, Mingjing Li, Changshui Zhang, and Wei-Ying Ma. 2005. Graph based multi-modality learning. In *Proceedings of the ACM MM*. ACM, 862–871.
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, N. Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the NeurIPS*. 5998–6008.
- [46] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the ACM MM*. ACM, 154–162.
- [47] Huanwen Wang, Yawen Zeng, Jianguo Chen, Zhouting Zhao, and Hao Chen. 2022. A Spatiotemporal Graph Neural Network for session-based recommendation. *Expert Systems with Applications* (2022).
- [48] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A Comprehensive Survey on Cross-modal Retrieval. *arXiv preprint arXiv:1607.06215* (2016).
- [49] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the CVPR*. IEEE, 5005–5013.
- [50] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5764–5773.
- [51] Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*

- (2019).
- [52] Jiaxin Wu and Chong-Wah Ngo. 2020. Interpretable Embedding for Ad-Hoc Video Search. In *Proceedings of the ACM MM*. ACM, 3357–3366.
- [53] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. 2022. GroupViT: Semantic Segmentation Emerges from Text Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18134–18144.
- [54] Jin Xu, Bo Tang, Haibo He, and Hong Man. 2017. Semisupervised Feature Selection Based on Relevance and Redundancy Criteria. *IEEE Trans. Neural Networks Learn. Syst.* 28, 9 (2017), 1974–1984.
- [55] Ruicong Xu, Li Niu, Jianfu Zhang, and Liqing Zhang. 2020. A Proposal-Based Approach for Activity Image-to-Video Retrieval. In *Proceedings of the AAAI*. AAAI Press, 12524–12531.
- [56] Caixia Yan, Qinghua Zheng, Xiaojun Chang, Minnan Luo, Chung-Hsing Yeh, and Alexander G. Hauptmann. 2020. Semantics-Preserving Graph Propagation for Zero-Shot Object Detection. *IEEE Transactions on Image Processing* (2020), 8163–8176.
- [57] Xiaojun Yang, Lunjia Liao, Qin Yang, Bo Sun, and Jianxiang Xi. 2021. Limited-energy output formation for multiagent systems with intermittent interactions. *Journal of the Franklin Institute* (2021), 6462–6489. <https://doi.org/10.1016/j.franklin.2021.06.009>
- [58] Yi Yang, Feiping Nie, Dong Xu, Jiebo Luo, Yueting Zhuang, and Yunhe Pan. 2011. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 4 (2011), 723–742.
- [59] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.
- [60] Tan Yu, Hongliang Fei, and Ping Li. 2022. U-BERT for Fast and Scalable Text-Image Retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*. 193–203.
- [61] Maia Zaharieva and Patrick Schwab. 2014. A Unified Framework for Retrieving Diverse Social Images. In *Proceedings of the Working Notes Proceedings of the MediaEval 2014 Workshop*, Vol. 1263.
- [62] Yawen Zeng, Da Cao, Shaofei Lu, Hanling Zhang, Jiao Xu, and Qin Zheng. 2022. Moment is Important: Language-Based Video Moment Retrieval via Adversarial Learning. *ACM Trans. Multim. Comput. Commun. Appl.* 18 (2022), 56:1–56:21.
- [63] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. 2021. Multi-Modal Relational Graph for Cross-Modal Video Moment Retrieval. In *Proceedings of the CVPR*. IEEE, 2215–2224.
- [64] Yawen Zeng, Yiru Wang, Dongliang Liao, Gongfu Li, Weijie Huang, Jin Xu, Da Cao, and Hong Man. 2022. Keyword-Based Diverse Image Retrieval With Variational Multiple Instance Graph. *IEEE Transactions on Neural Networks and Learning Systems* (2022).
- [65] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2012. Effective Heterogeneous Similarity Measure with Nearest Neighbors for Cross-Media Retrieval. In *Proceedings of the Advances in Multimedia Modeling*. Springer, 312–322.
- [66] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2014. Modeling Information Retrieval by Formal Logic: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 6 (2014), 965–978.
- [67] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).
- [68] Lu Zhang, Yang Wang, Jiaogen Zhou, Chenbo Zhang, Yinglu Zhang, Jihong Guan, Yatao Bian, and Shuigeng Zhou. 2022. Hierarchical Few-Shot Object Detection: Problem, Benchmark and Method. In *Proceedings of the 30th ACM International Conference on Multimedia*. 2002–2011.
- [69] Minyi Zhao, Yi Xu, and Shuigeng Zhou. 2021. Recursive fusion and deformable spatiotemporal attention for video compression artifact reduction. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5646–5654.
- [70] Wanqing Zhao, Ziyu Guan, Hangzai Luo, Jinye Peng, and Jianping Fan. 2017. Deep Multiple Instance Hashing for Object-based Image Retrieval. In *Proceedings of the IJCAI*. 3504–3510.
- [71] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep Supervised Cross-Modal Retrieval. In *Proceedings of the CVPR*. IEEE, 10394–10403.
- [72] Linchao Zhu and Yi Yang. 2020. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4344–4353.
- [73] Yueting Zhuang, Yi Yang, and Fei Wu. 2008. Mining Semantic Correlation of Heterogeneous Multimedia Data for Cross-Media Retrieval. *IEEE Transactions on Multimedia* 10, 2 (2008), 221–229.