Caption-Matching: A Multimodal Approach for Cross-Domain Image Retrieval

Lucas Iijima[1], Nikolaos Giakoumoglou[1] and Tania Stathaki[1]
[1]*Department of Electrical and Electronic Engineering, Imperial College London, London, UK*
lk422@ic.ac.uk, {n.giakoumoglou23, t.stathaki}@imperial.ac.uk

Keywords: Cross-domain image retrieval, Multimodal learning, Image captioning, Image-text matching

Abstract: Cross-Domain Image Retrieval (CDIR) is a challenging task in computer vision, aiming to match images across different visual domains such as sketches, paintings, and photographs. Existing CDIR methods rely either on supervised learning with labeled cross-domain correspondences or on methods that require training or fine-tuning on target datasets, often struggling with substantial domain gaps and limited generalization to unseen domains. This paper introduces a novel CDIR approach that incorporates textual context by leveraging publicly available pre-trained vision-language models. Our method, Caption-Matching (CM), uses generated image captions as a domain-agnostic intermediate representation, enabling effective cross-domain similarity computation without the need for labeled data or further training. We evaluate our method on standard CDIR benchmark datasets, demonstrating state-of-the-art performance in plug-and-play settings with consistent improvements on Office-Home and DomainNet over previous methods. We also demonstrate our method's effectiveness on a dataset of AI-generated images from Midjourney, showcasing its ability to handle complex, multi-domain queries.

1 INTRODUCTION

Image Retrieval (IR) is a fundamental computer vision task that involves finding images similar to a provided query from a large database. The applications of IR range from content-based search engines to visual question answering systems [Antol et al., 2015]. As the field has progressed, the challenge has expanded beyond single-domain retrieval to Cross-Domain Image Retrieval (CDIR), where the query and target images may belong to different visual domains such as sketches, paintings, or photographs.

Cross-domain image retrieval (CDIR) poses significant challenges due to the large domain gap between query and target images. Traditional image retrieval methods often fail in cross-domain settings because discriminative features within a single domain may not generalize across domains [Wang et al., 2017]. This domain gap, manifested as substantial differences in visual characteristics between domains, frequently results in poor retrieval performance.

Prior CDIR approaches have largely relied on supervised learning, requiring extensive labeled data and cross-domain correspondences [Sangkloy et al., 2016, Yu et al., 2016, Song et al., 2017b], but collecting such annotations is labor-intensive and limits scalability. Domain adaptation techniques attempt to align feature spaces across domains but often struggle when the visual domains are highly disparate [Tzeng et al., 2017]. Deep learning approaches that learn domain-invariant features can improve alignment, yet they typically demand significant computational resources and may fail to generalize to unseen domains [Ganin et al., 2017]. Unsupervised and self-supervised methods have been proposed to mitigate the reliance on labeled data [Kim et al., 2021, Li et al., 2021, Hu and Lee, 2022, Wang et al., 2023], but generally require training or fine-tuning on the target datasets and may not fully overcome the domain gap.

Recent advancements in vision-language models, such as CLIP [Radford et al., 2021], ALIGN [Jia et al., 2021], and BLIP-2 [Li et al., 2023], have opened new possibilities to tackle CDIR tasks. These models, pre-trained on large datasets of image-text pairs, have demonstrated remarkable zero-shot generalization across various vision tasks. However, their potential for CDIR remains largely unexplored.

In this paper, we introduce a novel approach to CDIR that leverages the capabilities of Large Language Models (LLMs) and Vision Transformers (ViTs) without the need for domain-specific annotations or fine-tuning. Our method, termed Caption-

Matching (CM), introduces caption-based semantic alignment, a conceptual shift that treats text as a domain-agnostic representation for cross-domain retrieval. Unlike prior methods that operate solely in the visual embedding space by learning domain-invariant features or aligning feature distributions, CM employs natural language descriptions to bridge domain gaps in a fundamentally different manner. By mapping images from various domains to a shared semantic space defined by captions, CM not only facilitates effective cross-domain retrieval without explicit domain alignment but also captures high-level semantic similarities across domains. This approach leverages the combined semantic understanding of LLMs and visual comprehension of ViTs, allowing our system to handle multiple domains simultaneously without requiring separate models or extensive fine-tuning for each domain pair, setting it apart from conventional techniques. The key contributions of our work are as follows:

- We introduce a novel CDIR framework that operates entirely without additional training, leveraging pre-trained vision-language models to eliminate the need for labeled cross-domain data or correspondence supervision. Our method uses generated image captions as a domain-agnostic intermediate representation, enabling effective cross-domain similarity computation.
- We demonstrate the effectiveness of our approach on DomainNet and Office-Home datasets, surpassing state-of-the-art performance in CDIR tasks.We further validate our method on a diverse dataset of Midjourney-generated images, showing its robustness in complex, multi-domain retrieval scenarios.

2 RELATED WORK

2.1 Image Retrieval

Image retrieval (IR) is a fundamental task in computer vision, typically involving the retrieval of images given an image query (*image-image* retrieval) [Noh et al., 2017]. Traditional approaches often follow a coarse-to-fine strategy, combining global retrieval with local feature aggregation and spatial verification. Deep learning and CNN-based methods have largely replaced hand-crafted features, yielding significant performance improvements [Noh et al., 2017, Radenović et al., 2019]. Notable methods include DELF and DELG [Noh et al., 2017], which learn local and global features jointly, GeM pooling [Radenović et al., 2019] for attentive aggregation, SOLAR [Ng et al., 2020] for self-attentive feature aggregation, and DOLG [Yang et al., 2021], which fuses local and global cues. Despite these advances, most IR systems remain sensitive to visually similar distractors, as global embeddings often rely on geometric consistency during re-ranking [Lee et al., 2023].

Retrieval paradigms have progressively evolved to embed natural language as a core interface for query formulation. Text-based image retrieval (TBIR) [Vendrow et al., 2024] uses textual queries to find images (*text-image* retrieval), while other methods [Song et al., 2017a, Sangkloy et al., 2022, Jia et al., 2021] combine an image with its description as a query (*[image+text]-image* retrieval) to improve similarity comparisons. In contrast, our method introduces a novel paradigm in which an image query retrieves semantically associated captions, each linked to an image, effectively performing image retrieval while leveraging a domain-agnostic, text-based intermediate (*image-text* retrieval). Importantly, it operates without explicit domain alignment or labeled cross-domain pairs and can handle multiple domains simultaneously, unlike traditional methods that require separate models or fine-tuning for each domain combination.

2.2 Cross-Domain Image Retrieval

Cross-domain image retrieval (CDIR) extends the challenge of image retrieval by searching across diverse domains, such as sketches, cartoons, paintings, and photographs. The primary challenge in CDIR is bridging the domain gap between query and database images. Early approaches leveraged category information for discriminative feature extraction or minimized losses like triplet [Yu et al., 2016] and HOLEF [Song et al., 2017b] for cross-domain pairing. However, these methods often require labor-intensive cross-domain correspondence annotations, limiting their practical applications [Sangkloy et al., 2016].

Recent efforts have explored unsupervised or self-supervised strategies to reduce reliance on labeled data. CDS [Kim et al., 2021] combines in-domain instance discrimination with cross-domain matching to learn domain-invariant representations. ProtoNCE [Li et al., 2021] introduces prototypes as latent semantic anchors within a prototypical contrastive framework. PCS [Yue et al., 2021] extends prototypical learning to few-shot domain adaptation through instance–prototype alignment. CCL [Hu and Lee, 2022] uses cluster-wise contrastive learning with a distance-to-distance objective for semantic align-

ment. CoDA [Wang et al., 2023] projects images into a shared subspace via correspondence-free alignment guided by self-matching and classifier-level consistency. While diverse in approach, all of these methods require training or fine-tuning on the target CDIR datasets, even when initialized with large pretrained backbones. In contrast, our caption-matching framework advances CDIR by operating entirely without training or fine-tuning on the target datasets.

2.3 Vision-Language Foundation Models

Vision-language models, such as CLIP [Radford et al., 2021] and ALIGN [Jia et al., 2021], have emerged as powerful foundation models by pretraining image and language encoder pairs on large-scale image-caption datasets. These models have demonstrated remarkable zero-shot generalization capabilities across various tasks, including image retrieval, classification, and visual question answering [Zhou et al., 2022, Song et al., 2022].

Following CLIP's success, numerous vision-language foundation models have been developed, incorporating larger datasets, novel architectures, and advanced training objectives [Alayrac et al., 2022, Li et al., 2022, Singh et al., 2022]. BLIP-2 [Li et al., 2023], which introduces a querying transformer to mediate between frozen pre-trained image encoders and LLMs, achieving state-of-the-art performance on various vision-language tasks. Although CLIP and BLIP-2 have been widely used for retrieval and captioning, their joint use as an intermediate semantic space for cross-domain retrieval remains unexplored.

In this work, we present a novel approach to adapt vision-language foundation models for cross-domain image retrieval tasks without requiring a dedicated CDIR dataset. Our caption-matching method uniquely combines the strengths of LLMs and ViTs, allowing it to capture high-level semantic similarities across diverse visual domains.

3 METHODOLOGY

Here, we present our method, Caption-Matching (CM), which utilizes pre-trained vision-language models to facilitate cross-domain image retrieval through a domain-agnostic, caption-based semantic embedding. This approach allows CM to effectively bridge the visual domain gap, enhancing retrieval accuracy without the need for extensive labeled datasets or domain-specific tuning. Figure 1 depicts an overview of the proposed approach.

3.1 Problem Statement

Consider a dataset distributed across a set of domains S, where $A \subset S$ and $B \subset S$ represent specific subsets associated with the query and target domains, respectively. The task of cross-domain image retrieval involves querying each image from subset D_A in domain A and retrieving a selection of the top-k similar images from subset D_B in domain B. This operation is formally denoted as $A \rightarrow B$. Evaluation is conducted at the category level, where retrieved images are deemed correct if they match the category of the query image. The challenge lies in effectively bridging the visual and semantic gaps between domains A and B without relying on labeled correspondences, requiring robust feature extraction that can generalize across these varied domains.

3.2 The Caption-Matching Method

Aiming to implement a domain-agnostic retrieval system, we introduce the Caption-Matching (CM) method, which leverages the advanced capabilities of LLMs and ViTs to analyze and synchronize data extracted from both image and text modalities. This method employs two principal models: f, a CLIP-based model that comprises f_{image} for image encoding and f_{text} for text encoding, and g, a BLIP-2-based model for generating textual captions from images.

Initially, each image x_i from the target domain B is processed by g to generate a descriptive caption t_i. These images are transformed into textual representations $t_i = g(x_i)$, converting visual data into a semantically enriched text format. Subsequently, in the comparison phase, each image x_j from the query domain A is encoded into its visual embedding $v_j = f_{\text{image}}(x_j)$ by the CLIP model. Simultaneously, the text encoder f_{text} is used to encode textual captions into embeddings. The matching process then computes the dot-product similarity scores $S_{ij} = v_j \cdot u_i$ between v_j and each u_i, where u_i is derived from $f_{\text{text}}(t_i)$, following CLIP's standard formulation. This metric assesses the relevance between the query images and the captions of the target images, facilitating a ranked retrieval of the most semantically aligned images. The final output is a sorted list of images from domain B that best correspond to the query image in domain A, thus bridging the semantic gap across domains via an effective integration of visual and linguistic analysis.

Figure 1 shows the diagram for the method proposed, where a BLIP-2 model (g) acts as an image captioning model and a CLIP model (f) acts as an image-text matching model. The combined implementation of an image captioning and an image-text

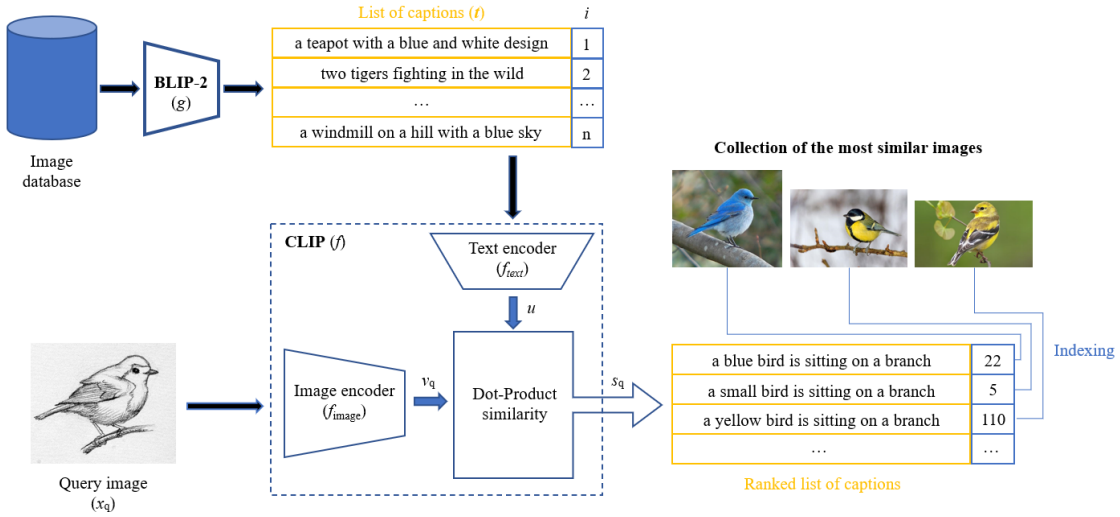

Figure 1: Overview of the proposed Caption-Matching (CM) method. The process begins with the BLIP-2 model g generating captions for each image in the database, where $g(x_i) = t_i$ produces textual captions t_i from images x_i. These captions are then processed through CLIP's text encoder to obtain text embeddings $u_i = f_{\text{text}}(t_i)$. Simultaneously, the query image x_q is encoded by CLIP's image encoder f_{image} to produce a visual embedding $v_q = f_{\text{image}}(x_q)$. A dot-product similarity score $S_{iq} = v_q \cdot u_i$ is calculated between v_q and each u_i, resulting in a list of similarity scores s_q. The relevance of each caption to the query image is ranked by sorting s_q and the corresponding images can be retrieved based on the highest scores.

matching model allows image representation using natural language, facilitating the seamless embedding of contextual information.

In the context of CDIR, a notable feature of CLIP when pre-trained on a sufficiently large and diverse dataset is its ability to bridge the domain gap without requiring modifications to the embedding space. Figure 2 illustrates this with a t-SNE [van der Maaten and Hinton, 2008] projection of text embeddings from selected examples. It shows that CLIP's text encoder clusters captions based on their semantic context while disregarding domain-specific terms such as "photo", "painting", and "sketch." This enables the corresponding image embeddings to be accurately matched with the correct image category across different domains. Hence, we choose captions as intermediates because text abstracts away stylistic domain characteristics while preserving semantic identity, making it suitable for CDIR.

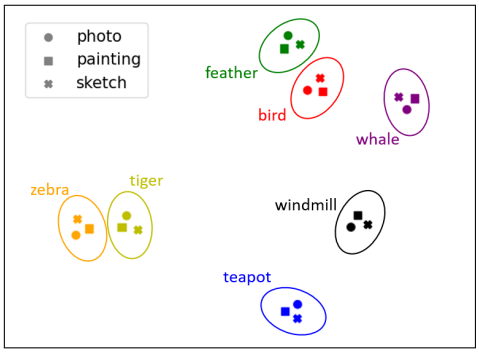

Figure 2: t-SNE visualization of CLIP text embeddings illustrating clustering capability across domains. In this experiment, captions were artificially constructed using the template "{Domain} of a {Class}" (e.g., "painting of a bird"), showing how semantic similarities are preserved while domain-specific identifiers are abstracted away, which facilitates domain-agnostic categorization.

4 EXPERIMENTS

4.1 Implementation Details

Models. We use two publicly available pre-trained models: BLIP-2[1] for image captioning and CLIP[2] for image-text matching. BLIP-2 combines a frozen ViT image encoder with a pre-trained OPT language model and was trained on a mixture of human-annotated datasets (COCO, Visual Genome) and large-scale web datasets (CC3M, CC12M, SBU, LAION-400M). CLIP uses a ViT-Huge image encoder with a GPT-2 text encoder and was trained on the LAION-2B-en dataset. Both models are used with default inference parameters and their publicly available pre-trained weights without modification.

Datasets. We evaluate our method using two diverse datasets. (1) **Office-Home** [Venkateswara et al.,

[1]https://huggingface.co/Salesforce/blip2-opt-2.7b

[2]https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K

2017] comprises approximately 15,500 manually annotated images sourced from the web, featuring four distinct domains: *art* (**Ar**), *clipart* (**Cl**), *product* (**Pr**), and *real world* (**Rw**)—spanning across 65 categories of common everyday objects such as chairs, keyboards, and bikes. (2) **DomainNet** [Peng et al., 2019] contains around 600,000 manually annotated images collected from the web, including six diverse domains: *clipart* (**C**), *infograph* (**I**), *painting* (**P**), *quickdraw* (**Q**), *real* (**R**), and *sketch* (**S**), spanning across 345 categories ranging from tools and clothing to natural objects and human body categories.

Baselines. We compare our method to state-of-the-art CDIR approaches: CDS [Kim et al., 2021], ProtoNCE [Li et al., 2021], PCS [Yue et al., 2021], CCL [Hu and Lee, 2022], and CoDA [Wang et al., 2023]. These methods require training on cross-domain target datasets, whereas CM requires no task-specific training or fine-tuning, yielding a direct comparison between training-free and training-based approaches. Furthermore, while all methods require feature extraction backbones, our goal is not to compare raw model capacity, but rather to evaluate whether caption-based representations provide an alternative paradigm for cross-domain retrieval. Accordingly, our results highlight methodological differences rather than variations in model scale.

Evaluation. We follow the experimental setup and metrics of [Hu and Lee, 2022, Wang et al., 2023]. On Office-Home, we evaluate all query-target domain pairs using precision scores P@1, P@5, P@15, and mAP@All. On DomainNet, we consider categories with over 200 samples and evaluate all domain pairs using P@50, P@100, and P@200.

4.2 Results

Table 1 summarizes the experimental results and comparisons with state-of-the-art methods. Our approach, which combines BLIP-2 for image captioning and CLIP for image-text matching, substantially outperforms existing techniques. On DomainNet and Office-Home, it nearly doubles the precision of prior methods across P@50, P@100, P@200 (DomainNet) and P@1, P@5, P@15 (Office-Home). It also consistently achieves higher mAP scores across all Office-Home domain pairs. These large gains arise because captions often encode semantic cues that remain invariant across visual domains, enabling CLIP's text encoder to perform more stable matching than visual features alone. Complementing the quantitative analysis, Figure 3 presents a qualitative comparison, illustrating the superiority of CM and revealing insights not captured by standard precision metrics.

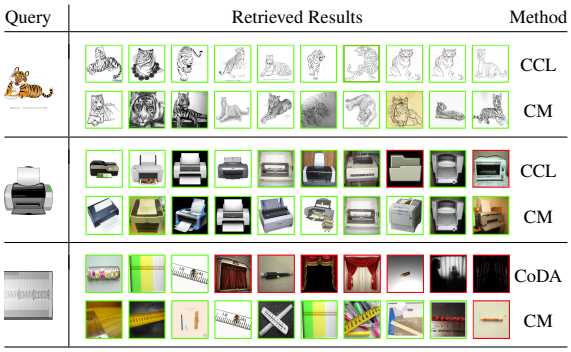

Figure 3: Qualitative comparison of the top-10 retrieved results. From top to bottom, the domain pair are C-S (DomainNet), Cl-Rw (Office-Home) and Cl-Ar (Office-Home). Correctly retrieved images are outlined in green, while incorrectly retrieved images are outlined in red.

4.3 Qualitative Evaluation with Multi-Domain Dataset

The proposed method is also qualitatively evaluated in an additional experiment with a multi-domain dataset[3] containing AI-generated images. This experiment used only the first 25k images generated with the model's latest versions, namely Midjourney v5.0, v5.1, and v5.2. Since the images are unlabeled, collections of retrieved images are presented as qualitative results. The aim of this experiment is to show how the CM method performs with complex images from previously unseen categories and ascertain its potential utility as a tool in generative applications.

The results presented in Figure 4 demonstrate the CM method's robust performance on highly detailed, AI-generated images. Notably, the method exhibits the ability to retrieve images across multiple domains for a single query, for example, pairing realistic images with anime-style illustrations (third query) or drawings with comic-style renderings (fourth query). This capability further emphasizes the domain-agnostic nature of the CM method.

5 DISCUSSION

Our Caption-Matching (CM) approach delivers strong improvements in cross-domain image retrieval while requiring no training or domain-specific adaptation. By leveraging BLIP-2 for caption generation and CLIP for image–text matching, CM inherits rich semantic priors learned from large-scale multimodal datasets. As illustrated in Figure 2, CLIP captures both content and contextual cues, enabling it to

[3]Available at https://www.kaggle.com/datasets/iraklip/modjourney-v51-cleaned-data.

Table 1: Comparison of retrieval performance between our method (CM) and state-of-the-art methods on the DomainNet and Office-Home datasets. Domain pairs are denoted by abbreviations (e.g., C-S for Clipart to Sketch). Bold values indicate the best performance for each metric and domain pair.

	DomainNet												
	C-S	**S-C**	**I-R**	**R-I**	**I-S**	**S-I**	**P-C**	**C-P**	**P-Q**	**Q-P**	**Q-R**	**R-Q**	**Avg**
P@50													
CDS [Kim et al., 2021]	0.458	0.591	0.285	0.567	0.306	0.463	0.632	0.378	0.188	0.214	0.193	0.154	0.369
ProtoNCE [Li et al., 2021]	0.468	0.545	0.284	0.570	0.282	0.398	0.554	0.391	0.216	0.240	0.264	0.251	0.372
PCS [Yue et al., 2021]	0.510	0.597	0.306	0.554	0.303	0.426	0.635	0.488	0.251	0.240	0.348	0.290	0.412
CCL [Hu and Lee, 2022]	0.563	0.631	0.355	0.577	0.313	0.437	0.664	0.526	0.397	0.334	0.428	0.419	0.470
CM (ours)	**0.968**	**0.939**	**0.693**	**0.884**	**0.722**	**0.879**	**0.971**	**0.931**	**0.832**	**0.546**	**0.522**	**0.792**	**0.807**
P@100													
CDS [Kim et al., 2021]	0.424	0.488	0.279	0.398	0.295	0.361	0.473	0.352	0.189	0.214	0.191	0.156	0.318
ProtoNCE [Li et al., 2021]	0.427	0.450	0.285	0.418	0.268	0.320	0.437	0.359	0.212	0.228	0.257	0.248	0.326
PCS [Yue et al., 2021]	0.469	0.507	0.303	0.421	0.284	0.341	0.532	0.462	0.246	0.232	0.339	0.289	0.369
CCL [Hu and Lee, 2022]	0.527	0.573	0.352	0.467	0.293	0.361	0.568	0.501	0.386	0.338	0.428	0.421	0.435
CM (ours)	**0.962**	**0.919**	**0.694**	**0.797**	**0.711**	**0.796**	**0.961**	**0.921**	**0.812**	**0.535**	**0.501**	**0.793**	**0.784**
P@200													
CDS [Kim et al., 2021]	0.372	0.374	0.275	0.264	0.270	0.273	0.329	0.328	0.179	0.195	0.187	0.158	0.267
ProtoNCE [Li et al., 2021]	0.364	0.351	0.285	0.303	0.242	0.248	0.326	0.321	0.206	0.216	0.244	0.238	0.279
PCS [Yue et al., 2021]	0.402	0.394	0.297	0.308	0.254	0.259	0.417	0.421	0.238	0.221	0.317	0.282	0.317
CCL [Hu and Lee, 2022]	0.474	0.482	0.344	0.355	0.265	0.281	0.467	0.461	0.376	0.343	0.427	0.416	0.391
CM (ours)	**0.935**	**0.869**	**0.690**	**0.652**	**0.684**	**0.658**	**0.921**	**0.899**	**0.751**	**0.522**	**0.503**	**0.747**	**0.736**
	Office-Home												
	Ar-Cl	**Ar-Pr**	**Ar-Rw**	**Cl-Ar**	**Cl-Pr**	**Cl-Rw**	**Pr-Ar**	**Pr-Cl**	**Pr-Rw**	**Rw-Ar**	**Rw-Cl**	**Rw-Pr**	**Avg**
P@1													
CDS [Kim et al., 2021]	0.256	0.328	0.451	0.224	0.272	0.325	0.358	0.377	0.540	0.447	0.389	0.494	0.372
ProtoNCE [Li et al., 2021]	0.290	0.295	0.405	0.213	0.212	0.252	0.357	0.361	0.538	0.445	0.412	0.477	0.355
PCS [Yue et al., 2021]	0.312	0.333	0.417	0.245	0.264	0.291	0.392	0.395	0.564	0.450	0.406	0.499	0.381
CCL [Hu and Lee, 2022]	0.327	0.354	0.451	0.273	0.278	0.333	0.425	0.423	0.574	0.480	0.447	0.517	0.407
CM (ours)	**0.783**	**0.718**	**0.807**	**0.678**	**0.711**	**0.729**	**0.843**	**0.883**	**0.924**	**0.845**	**0.848**	**0.869**	**0.803**
P@5													
CDS [Kim et al., 2021]	0.238	0.315	0.412	0.203	0.265	0.303	0.325	0.350	0.501	0.408	0.365	0.473	0.346
ProtoNCE [Li et al., 2021]	0.262	0.279	0.364	0.174	0.206	0.227	0.306	0.340	0.482	0.393	0.377	0.448	0.321
PCS [Yue et al., 2021]	0.287	0.315	0.385	0.213	0.259	0.261	0.348	0.375	0.508	0.399	0.381	0.471	0.350
CCL [Hu and Lee, 2022]	0.308	0.347	0.423	0.239	0.273	0.306	0.379	0.374	0.527	0.437	0.415	0.485	0.376
CM (ours)	**0.768**	**0.723**	**0.791**	**0.637**	**0.700**	**0.712**	**0.791**	**0.855**	**0.878**	**0.801**	**0.829**	**0.855**	**0.778**
P@15													
CDS [Kim et al., 2021]	0.224	0.289	0.387	0.173	0.249	0.278	0.268	0.304	0.456	0.355	0.332	0.440	0.313
ProtoNCE [Li et al., 2021]	0.230	0.258	0.340	0.145	0.205	0.208	0.246	0.282	0.422	0.330	0.320	0.412	0.283
PCS [Yue et al., 2021]	0.261	0.295	0.362	0.175	0.249	0.240	0.288	0.328	0.454	0.340	0.341	0.437	0.314
CCL [Hu and Lee, 2022]	0.287	0.326	0.401	0.205	0.260	0.281	0.314	0.337	0.479	0.384	0.374	0.450	0.341
CM (ours)	**0.728**	**0.694**	**0.756**	**0.558**	**0.656**	**0.671**	**0.694**	**0.804**	**0.837**	**0.721**	**0.788**	**0.816**	**0.727**
mAP@All													
PCS [Yue et al., 2021]	0.343	0.463	0.516	0.323	0.405	0.406	0.470	0.421	0.613	0.516	0.428	0.601	0.459
CoDA [Wang et al., 2023]	0.347	0.496	0.532	0.332	0.429	0.447	0.504	0.452	0.652	0.531	0.460	0.652	0.486
CM (ours)	**0.528**	**0.539**	**0.590**	**0.452**	**0.509**	**0.523**	**0.571**	**0.596**	**0.658**	**0.596**	**0.600**	**0.662**	**0.569**

bridge domain gaps that challenge traditional visual-only methods.

The use of captions as a domain-agnostic intermediate representation allows CM to encode high-level semantics that remain stable across diverse visual styles. This leads to more meaningful retrieval results, as shown in Figure 3, where CM returns images that align with both the category and contextual attributes of the query. Unlike existing CDIR approaches that rely solely on visual features, CM benefits from the complementary strengths of vision-language modeling.

CM also demonstrates strong scalability, supporting multiple domains without requiring separate models or fine-tuning. Its ViT-based components accept variable input resolutions, avoiding distortions common in CNN-based pipelines. Importantly, in contrast to existing CDIR methods, which are trained directly on the evaluation datasets to learn domain-invariant visual features, CM relies solely on off-the-shelf pre-trained models and achieves superior retrieval performance. This is further reflected in qualitative results such as those in Figure 3, where CM retrieves images that match both the category and pose of the query (e.g., a tiger laying down), capturing contextual semantics that visual-only methods fail to model. These findings highlight the advantages of multimodal representations for CDIR and suggest that further advances in caption generation and large-scale pre-training could yield even greater improvements.

6 CONCLUSIONS

We proposed a novel caption-matching approach for cross-domain image retrieval that leverages both language and vision features for the task. The CM

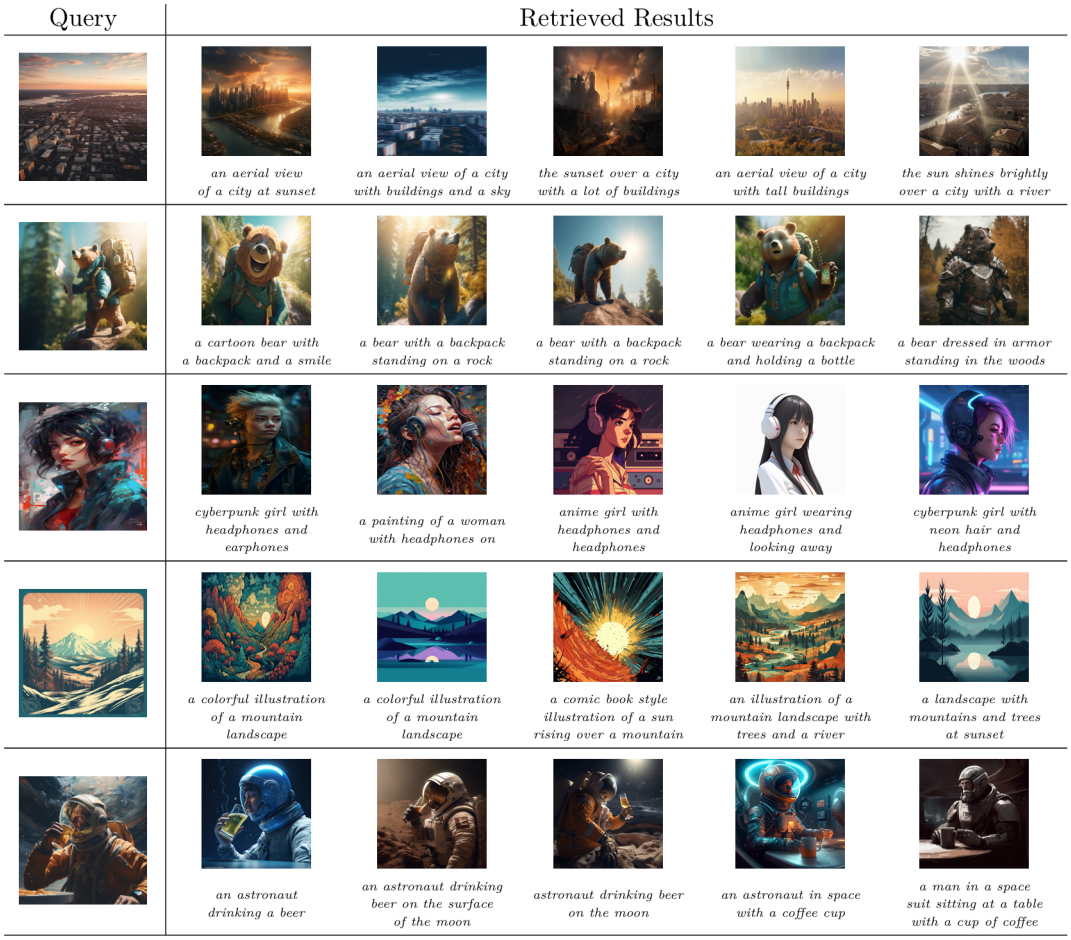

Figure 4: Top-5 retrieval results on Midjourney's database. Descriptions were generated by BLIP-2 as part of the CM method.

method matches a query image with the most suitable descriptions, which are concurrently associated with images from the target database. While most approaches in the literature are limited to operations with visual embeddings, the CM method overcomes the domain gap by strategically integrating CLIP, whose text encoder is able to cluster text descriptions based on image context. It achieves state-of-the-art performance in CDIR and performs remarkably well on AI-generated images, without the need for fine-tuning on specific datasets. Our work highlights the potential of using language as a bridge across visual domains, suggesting new research directions where captions or textual prompts serve as structured intermediates.

REFERENCES

Alayrac, J.-B. et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2017). Domain-adversarial training of neural networks. In *Domain Adaptation in Computer Vision Applications*, pages 189–209. Springer.

Hu, C. and Lee, G. H. (2022). Feature representation learning for unsupervised cross-domain image retrieval. In *European Conference on Computer Vision (ECCV)*.

Jia, C. et al. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Kim, D., Saito, K., Oh, T.-H., Plummer, B. A., Sclaroff, S., and Saenko, K. (2021). Cds: Cross-domain self-supervised pre-training. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9123–9132.

Lee, S., Lee, S., Seong, H., and Kim, E. (2023). Revisiting self-similarity: Structural embedding for image retrieval. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23412–23421.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Li, J., Zhou, P., Xiong, C., and Hoi, S. (2021). Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations (ICLR)*.

Ng, T., Balntas, V., Tian, Y., and Mikolajczyk, K. (2020). Solar: Second-order loss and attention for image retrieval. In *European Conference on Computer Vision (ECCV)*, pages 253–270. Springer.

Noh, H., Araujo, A., Sim, J., Weyand, T., and Han, B. (2017). Large-scale image retrieval with attentive deep local features. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3456–3465.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. (2019). Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1406–1415.

Radenović, F., Tolias, G., and Chum, O. (2019). Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Sangkloy, P., Burnell, N., Ham, C., and Hays, J. (2016). The sketchy database: learning to retrieve badly drawn bunnies. In *ACM SIGGRAPH 2016 Papers*, pages 1–12.

Sangkloy, P., Jitkrittum, W., Yang, D., and Hays, J. (2022). A sketch is worth a thousand words: Image retrieval with text and sketch. In *European Conference on Computer Vision (ECCV)*, pages 251–267.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. (2022). Flava: A foundational language and vision alignment model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650.

Song, H., Dong, L., Zhang, W., Liu, T., and Wei, F. (2022). Clip models are few-shot learners: Empirical studies on vqa and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6088–6100.

Song, J., Song, Y.-Z., Xiang, T., and Hospedales, T. (2017a). Fine-grained image retrieval: the text/sketch input dilemma. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 45.1–45.12.

Song, J., Yu, Q., Song, Y.-Z., Xiang, T., and Hospedales, T. M. (2017b). Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5551–5560.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Vendrow, E. et al. (2024). Inquire: A natural world text-to-image retrieval benchmark. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 126500–126514. Curran Associates, Inc.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5018–5027.

Wang, B., Yang, Y., Xu, X., Hanjalic, A., and Shen, H. T. (2017). Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 154–162.

Wang, X., Peng, D., Yan, M., and Hu, P. (2023). Correspondence-free domain alignment for unsupervised cross-domain image retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):10200–10208.

Yang, M., He, D., Fan, M., Shi, B., Xue, X., Li, F., Ding, E., and Huang, J. (2021). Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11772–11781.

Yu, Q., Liu, F., Song, Y.-Z., Xiang, T., Hospedales, T. M., and Loy, C.-C. (2016). Sketch me that shoe. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 799–807.

Yue, X. et al. (2021). Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348.