

Composed Image Retrieval for Training-FREE Domain Conversion

Nikos Efthymiadis^{1*} Bill Psomas^{1,2} Zakaria Laskar¹ Konstantinos Karantzalos²
 Yannis Avrithis³ Ondřej Chum¹ Giorgos Tolias¹

¹VRG, FEE, Czech Technical University in Prague ²National Technical University of Athens
³Institute of Advanced Research in Artificial Intelligence (IARAI), Austria

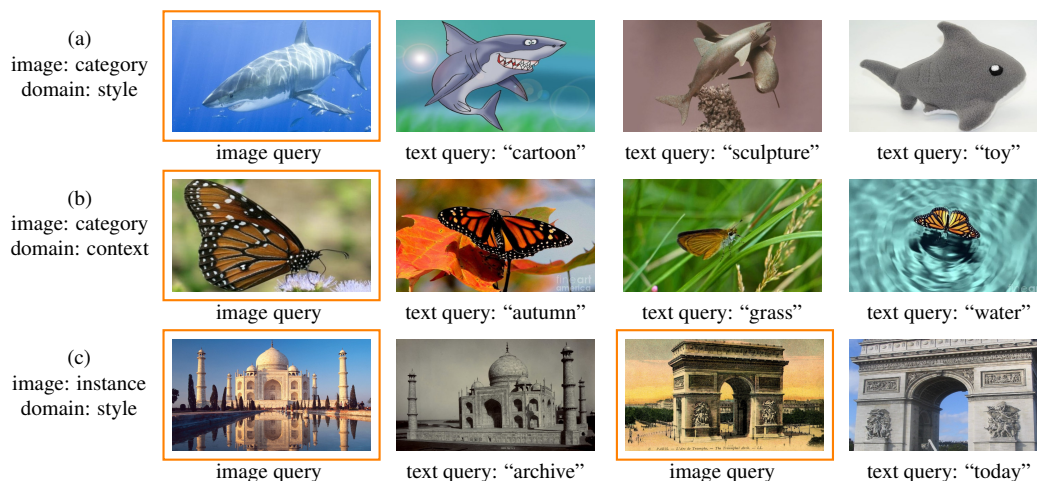


Figure 1. We introduce FREEDOM, a training-free, composed image retrieval method for domain conversion based on CLIP [68]. Given an *image query* (framed) and a *text query* that names a *domain*, images are retrieved having the class of the image query and the domain of the text query. A range of applications is targeted, where classes can be defined at *category level* (a,b) [31, 93] or *instance level* (c) [21], and domains can be defined as *styles* (a, c), or *context* (b). For each image query, retrieved images are shown for different text queries.

Abstract

This work addresses composed image retrieval in the context of domain conversion, where the content of a query image is retrieved in the domain specified by the query text. We show that a strong vision-language model provides sufficient descriptive power without additional training. The query image is mapped to the text input space using textual inversion. Unlike common practice that invert in the continuous space of text tokens, we use the discrete word space via a nearest-neighbor search in a text vocabulary. With this inversion, the image is softly mapped across the vocabulary and is made more robust using retrieval-based augmentation. Database images are retrieved by a weighted ensemble of text queries combining mapped words with the domain text. Our method outperforms prior art by a large margin on standard and newly introduced benchmarks. Code: <https://github.com/NikosEfth/freedom>

*Correspondence: efthynik@fel.cvut.cz

1. Introduction

Image-to-image retrieval is a task that involves applications to landmarks [67], fashion products [56], face recognition [14], remote sensing [7], and medical images [61], among others. The retrieval is performed according to the visual content of the query [25, 63]. On the other hand, if the object can be described with text, then text-to-image retrieval [22, 76, 92] applies. The most flexible way to express the user intent is a query comprising both an image and a text description. This is explored in *composed image retrieval* (CIR) [3, 11, 33, 47, 74, 82], which aims to retrieve target images that are not only visually similar to the query image but also modified by the text query.

Traditionally, CIR methods are supervised by triplets [10, 82, 87]. However, the labor-intensive process of labeling confined early research to specific applications in fashion [4, 29, 84], physical states [39], object attributes, and object composition [54, 62, 82]. The emergence of vision-language models (VLM) [42, 50, 68] led to their integration into CIR. Initially, this has been achieved by

fine-tuning using triplets [3]. More recently, *zero-shot composed image retrieval* (ZS-CIR) [2, 74] significantly increased the spectrum of applications. Most existing methods use *textual inversion*, i.e., mapping the query image to text, thus allowing query composition purely by means of text. Also, most methods are trained using unlabeled images [2, 74] or are not trained at all [44], but they require the use of large language models [6] (LLM).

In this paper, we focus on a specific variant of composed image retrieval, namely *domain conversion*, where the text query defines the target domain [74]. Unlike conventional cross-domain retrieval [37], where models are trained to use queries of a source domain and retrieve items from another target domain, we address a more practical, *open-domain* setting, where the query and database may be from any unseen domain. We target different variants of this task, where the class of the query object is defined at *category-level* or *instance-level*. At the same time, the domain corresponds to descriptions of *style* or *context*, as shown in Figure 1. Even though domain conversion is a subset of the tasks handled by existing CIR methods, the variants considered in our work reflect a more comprehensive set of applications than what was encountered in prior art [74].

Large pre-trained VLMs provide powerful representations of objects, domains, and their combinations. Our approach is *training-free* by using a frozen VLM and performing textual inversion in a *non-parametric* way, assuming access to an external large memory of words. Inversion maps images to the discrete input space of text instead of the continuous latent space of word tokens as in prior work [2]. Compared to such an alternative, our memory-based inversion is more efficient and intuitive and comes with significant performance benefits. While our emphasis lies in domain conversion, the proposed approach is versatile and applicable to various composed image retrieval tasks, where its performance is competitive to the state-of-the-art approaches. We make the following contributions:

1. We are the first to focus on composed image retrieval in the context of domain conversion, and we introduce three new benchmarks to the task while we extend an existing one to more source domains.
2. We introduce FREEDOM, a training-free CIR method for domain conversion that operates in an open world by inheriting the capabilities of a frozen CLIP model.
3. We demonstrate that textual inversion performs better in the discrete input space of known words than in the continuous latent space of pseudo-words.
4. FREEDOM outperforms all methods by a large margin on four benchmarks.
5. Our experimental results form a testbed for future comparisons in this task.

2. Related work

Composed image retrieval (CIR). Image-to-image [25, 63] and text-to-image [22, 76, 92] retrieval provide useful ways to explore large image collections. Nevertheless, composed image retrieval offers more flexible ways to express the query and enables novel applications. TIRG [82] is the first to *compose* image and text as a search query, where text serves as a modification of the image to refine the retrieval results. Training is supervised with cross-entropy loss, using triplets of the form *reference image*, *query text*, *target image*. Following the same setting, JVSM [10] learns image-text compositional embeddings in a unified space using multiple matching losses.

Other methods exploit attention in the form of multi-modal disentangled non-local blocks [87] to correlate text with image regions [33] extracted by an RPN [70], to perform the composition at multiple depths [11], to modulate content [47], and to discover the relation between composed query and target image [13]. DRA [43] learns a dual relation, implicit/explicit, alignment network, while MPC [62] introduces a task variant with multiple queries. All these methods perform training from scratch and rely on triplets related to fashion [4, 29, 84], physical states [39], object attributes, and object composition [54, 62, 82]. Labeling triplets is expensive and limits the broader use of CIR.

Inspired by vision-language foundation models [5, 42, 68], recent work builds upon them in different ways. CIR-PLANT [57] and FashionVLP [24] extract features from a reference image as well as text features using a tokenizer [15, 52] and fine-tune the VLM [52, 68, 91] using triplets. CLIP4CIR [3] fine-tunes CLIP [68] and trains a small network to combine image and text features using triplets. BLIP4CIR [58] builds upon CLIP4CIR with BLIP [50] and trains using reversed triplets, along with the original ones. SPRC [1] uses train triplets and trains a transformer model to combine the image and the text query.

To ensure scalability, datasets are often created through crowd-sourcing human text [57], by exploiting LAION-5B [55, 77] or VQA v2.0 [26, 48], or by automatically synthesizing millions of high-quality triplets [27, 40] using generative models [71]. All these methods benefit from the compositional ability of vision-language models [50, 52, 68, 91] but still rely on triplets. Chen and Lai [8] train only with image-caption pairs and a masking technique that improves a simple inference-time baseline.

Pic2Word [74] relies on a VLM and is the first to avoid triplets and to evaluate on the domain conversion task. It follows self-supervised training to invert the query image to a text token. Thus, query composition takes place in the text domain by combining this token with the query text. KEDs [80] extends Pic2Word by a retrieval augmentation process that uses an external image-caption database, while ISA [19] maps images to sentences in the form of multiple

tokens. LinCIR [28] efficiently learns the inversion used by Pic2Word but with language-only self-supervised training. In contrast to using an image-level representation in Pic2Word, Context-I2W [81] performs localization of the relevant image region to improve the inversion.

SEARLE [2] performs textual inversion with test-time optimization per query image. Then, a network is trained to imitate the result of such optimization to perform inference more efficiently. Our memory-based textual inversion avoids pre-training or test-time optimization.

CIReVL [44] composes image and text queries solely in the language domain. It uses CLIP [68] as an image and text encoder, BLIP-2 [49] to caption the reference image, and GPT-3.5 turbo [6] to recompose the generated caption based on the query text. All backbones are pre-trained.

MagicLens [90] is a co-current method that uses the implicit relations of images found on the same website to create triplets of query-image, instruction, and target-image. Then, they fine-tune a VLM that is expanded with four extra attention layers and one attention pooling layer.

Training-free use of VLMs. The emergence of vision-language models (VLMs) [42, 49, 50, 68] revolutionized the field of multimodal learning. Trained on massive datasets [77], these models have instrumental abilities to map images and text into a shared embedding space and are successful in training-free scenarios. MaskCLIP [94] and CLIP-DIY [85] demonstrate the intrinsic potential of CLIP for semantic segmentation, while FLDM [60] highlights its effectiveness in text-guided video editing. The training-free paradigm extends to text-guided image editing [32] and layout control [9] using cross-attention. VLMs are also promising in specialized applications, such as deepfake detection [69], cross-domain image composition [59], and phrase localization [51]. Related to our training-free approach, CIReVL [44] uses VLMs and LLMs to compose image and text queries in the language domain.

Cross-domain image retrieval (CDIR). In this task, the query and database images come from different domains, and the challenge is bridging the domain gap [96]. As a visual domain, one might consider style [79], color [23], texture [79], context [37], lighting conditions [37], or images captured using different sensors [30]. One main line of research is in sketch-based image retrieval [36, 53, 72, 73, 75, 79, 88], and another on consumer scenarios [37, 38, 41].

Early methods do not generalize to new object classes or domains. This is the goal of zero-shot sketch-based retrieval [16, 20, 86]. More recent methods dispense with the need for labeled cross-modal pairs and are unsupervised [34, 35, 45, 83]. Generalization to an unseen domain is only demonstrated by UCDR [64]. Nevertheless, no CDIR method includes the domain of the query image in the database, which becomes meaningful in our task, *i.e.*, domain conversion with image-text queries.

3. Method

3.1. Preliminaries

Composed image retrieval is the task of retrieving images using a query that combines both a *visual* component, represented by a query image, and a *textual* component, represented by a query text. This work focuses on a specific variant of composed image retrieval that targets *domain conversion* [74].

In particular, the query image y depicts an object of *class* $C(y)$ in the *source domain* $D(y)$, while the query text t represents the *target domain*, $D(t)$. The two elements are jointly referred to as the *composed query*, $q = (y, t)$. Given an image dataset X , the goal is to retrieve images from X whose class is the same as that of the query image, $C(y)$, and whose domain is specified by the query text, $D(t)$. Retrieval amounts to ranking images $x \in X$ according to their *composed similarity* $s(q, x) \in \mathbb{R}$ to the query q .

We rely on a pre-trained vision-language model that consists of a *visual encoder* $f : \mathcal{I} \rightarrow \mathbb{R}^d$ and a *text encoder* $g : \mathcal{T} \rightarrow \mathbb{R}^d$, which map respectively input images from the image space $\mathcal{I}$ and words¹ from the text space $\mathcal{T}$ to the same embedding space of dimension d . Using those encoders, a visual embedding $\mathbf{y} = f(y) \in \mathbb{R}^d$ and a text embedding $\mathbf{t} = g(t) \in \mathbb{R}^d$ are extracted for the query. Similarly, the embedding of an image $x \in X$ or a word w are denoted by $\mathbf{x} = f(x) \in \mathbb{R}^d$ and $\mathbf{w} = g(w) \in \mathbb{R}^d$, respectively. All embeddings are ℓ_2 -normalized.

Given the visual-language model, this work aims to represent the composed query in the same embedding space as images and text. That is, define a *composed encoder* $h : \mathcal{I} \times \mathcal{T} \rightarrow \mathbb{R}^d$ such that the composed query q is mapped to $\mathbf{q} = h(q) = h(y, t) \in \mathbb{R}^d$, again ℓ_2 -normalized. For any image $x \in X$, this allows us to express the composed similarity as the cosine similarity

$$s(q, x) := h(q)^\top f(x) = h(y, t)^\top f(x). \quad (1)$$

Thus, given the encoders f, g , the goal is to define h .

3.2. Expanded textual inversion

Textual inversion. The ability of the text encoder, by design, to combine different concepts at its input and map them jointly to the embedding space motivates us to represent the composed query $q = (y, t)$ entirely in the text space $\mathcal{T}$ and then map it to the embedding space, using g . The query text t is already in $\mathcal{T}$; but to map the query image y to $w^* \in \mathcal{T}$, it must first be embedded to $\mathbf{y} = f(y)$ and then mapped from the embedding space back to $\mathcal{T}$:

$$w^* = g^{-1}(f(y)). \quad (2)$$

¹With the term *words*, we shall refer to both words and sentences.

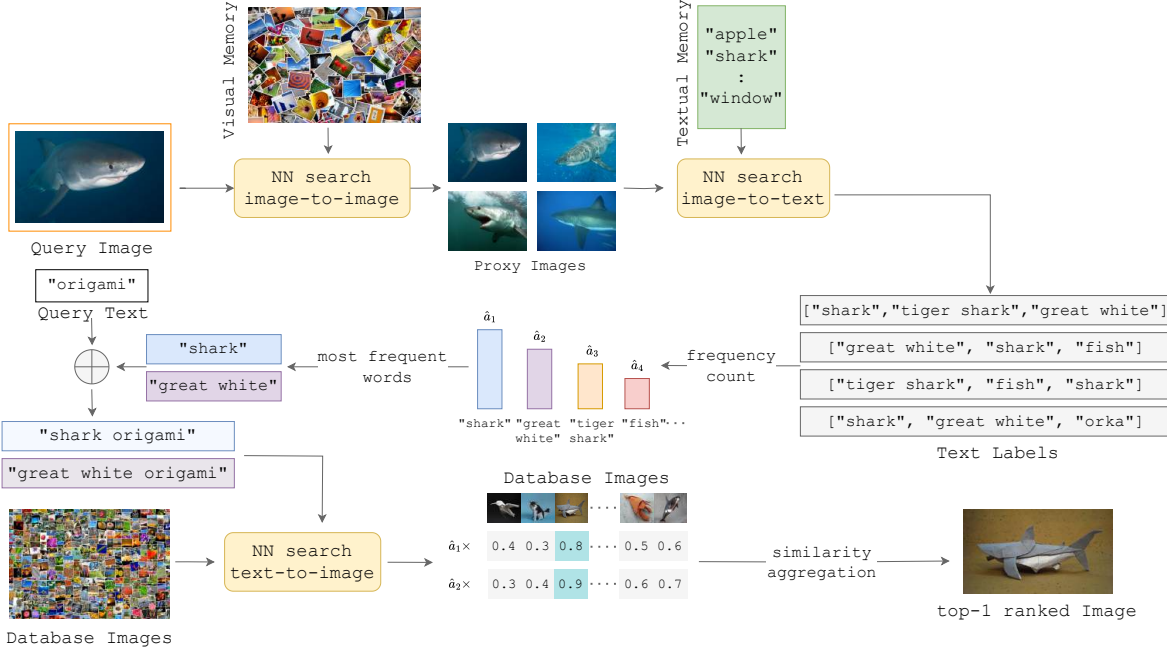


Figure 2. Overview of FREEDOM. Given a query image and a query text indicating the target domain, proxy images are first retrieved from the query through an image-to-image search over a visual memory. Then, a set of text labels is associated with each proxy image through an image-to-text search over a textual memory. Each of the most frequent text labels is combined with the query text in the text space, and images are retrieved from the database by text-to-image search. The resulting sets of similarities are linearly combined with the frequencies of occurrence as weights. Here: $k = 4$ proxy images, $n = 3$ text labels per proxy image, $m = 2$ most frequent text labels.

The process of mapping the query image y to text in $\mathcal{T}$ is called *textual inversion*, and the challenge is that the inverse mapping g^{-1} is unknown.

Common approaches are *pre-training* [74] and *test-time optimization* [2], both representing w^* in the latent space of the vector tokens. The former defines a decoder and trains it on a dataset to *learn* the inverse mapping g^{-1} of the text encoder. Its challenge is the sheer scale of training required to reach anywhere close to the quality of the pre-trained text encoder g . The latter defines a variable at the input of g and finds the optimal solution w^* such that $g(w^*) = f(y)$. Its challenge is that there is a multitude of locally optimal solutions. Thus, the approach overly relies on the initialization of the variable, which remains unknown.

Memory-based inversion. Contrary to existing approaches, the inversion is achieved by nearest neighbor search over an external vocabulary $V \subset \mathcal{T}$ of words [46], without training or optimization, and the w^* is found in the discrete text space $\mathcal{T}$ rather than the continuous latent space of vector tokens.

In particular, if $V = \{v_1, \dots, v_N\}$, vocabulary words v_i are mapped to text embeddings $\mathbf{v}_i = g(v_i)$ for $i = 1, \dots, N$. The *text memory* of words v_i and associated embeddings $\mathbf{v}_i$ is the restriction $g|_V : V \rightarrow \mathcal{T}$ of g to V :

$$g|_V : \{v_1, \dots, v_N\} \rightarrow \{\mathbf{v}_1, \dots, \mathbf{v}_N\}. \quad (3)$$

Given an embedding $\mathbf{v}_i \in \mathcal{V} = g(V) = \{\mathbf{v}_1, \dots, \mathbf{v}_N\}$, the associated word v_i can be instantly determined. This process essentially defines the restriction $g^{-1}|_{\mathcal{V}} : \mathcal{V} \rightarrow V$ of the otherwise unknown inverse g^{-1} to $\mathcal{V}$. For brevity, we refer to $g^{-1}|_{\mathcal{V}}$ as g^{-1} in the following.

What remains is, given the query embedding $\mathbf{y} \in \mathbb{R}^d$, to approximate it by one or more vectors in $\mathcal{V}$. This is done by finding the m nearest neighbors of $\mathbf{y}$ in $\mathcal{V}$,

$$\mathcal{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_m\} = \text{NN}_m(\mathbf{y}; \mathcal{V}), \quad (4)$$

given by descending order of (cosine) similarity. Since $\mathcal{W} \subset \mathcal{V}$, an embedding $\mathbf{w}_i \in \mathcal{W}$ can be mapped by g^{-1} back to the associated word $w_i = g^{-1}(\mathbf{w}_i)$ in V . Thus, all neighbors are mapped to words

$$W = \{w_1, \dots, w_m\} = g^{-1}(\mathcal{W}). \quad (5)$$

Putting everything together, this set of words is given by $W = \phi_V(y)$, where *NN-inversion*

$$\phi_V(y) := g^{-1}(\text{NN}_m(f(y); g(V))) \quad (6)$$

is an approximation of (2) by the vocabulary V . The larger the vocabulary, the better the quality of approximation—but the more expensive the process. The function ϕ_V (6) is used for defining different versions of composed encoder h below.

Memory-based inversion is similar to zero-shot recognition, where the query image y is represented by a set of words W from the vocabulary V . We call the words found by $\phi_V(y)$ the *text labels* or *labels* of y .

Single-word inversion. The closest word w_1 to the query image is merged with the query text t to form a composed query $w_1 \oplus t$ in the text space alone, where $\oplus$ denotes space-delimited string concatenation. Thus, h becomes

$$h_1(y, t) := \beta_1(\phi_V(y), t), \quad (7)$$

where ϕ_V is given by (6) and

$$\beta_1(W, t) := g(w_1 \oplus t). \quad (8)$$

A single word often works well, but using more words may help when the correct class is not top-ranked or when a collection of words represents better a particular image query.

Multi-word inversion: early fusion. We take advantage of the ability of the text encoder to combine several words in its input. Now, a composed query $w_1 \oplus \dots \oplus w_m \oplus t$ is formed in the text space and h becomes

$$h_E(y, t) := \beta_E(\phi_V(y), t), \quad (9)$$

where ϕ_V is given by (6) and

$$\beta_E(W, t) := g(w_1 \oplus \dots \oplus w_m \oplus t). \quad (10)$$

We refer to this approach as *early fusion* since the words w_i are combined at the earliest possible stage.

Multi-word inversion: late fusion. Early fusion may be sensitive to words assigned incorrectly by NN-inversion ϕ_V (6). The other extreme is *late fusion*, whereby words are composed at the latest possible stage. In particular, one composed query $w_i \oplus t$ is formed in the text space for each word w_i , is embedded separately, and a linear combination of these embeddings is formed. Thus, h becomes

$$h_L(y, t) := \beta_L(\phi_V(y), t), \quad (11)$$

where ϕ_V is given by (6),

$$\beta_L(W, t, \mathbf{a}) := \sum_{i=1}^m a_i g(w_i \oplus t) \quad (12)$$

and $a_i \in \mathbb{R}$ is a weight associated with word w_i , by default uniform $\mathbf{a} = \mathbf{1} \in \mathbb{R}^m$. Because of the linearity of (1), this is equivalent to m independent queries followed by a linear combination of the resulting similarities.

Memory-based expansion. Even if multiple nearest neighbors are used, the underlying cross-modal (image-to-text) similarity of (4) remains challenging. A retrieval-based augmentation mechanism is employed to achieve more reliable zero-shot recognition. A *visual memory* of images z_i and the associated embeddings $\mathbf{z}_i = f(z_i)$ from an external

image set Z are used. First, the query image is expanded through a set of k *proxy images* (including the query image), found as nearest neighbors of $\mathbf{y}$ in the embeddings $\mathcal{Z} = f(Z)$ based on unimodal (image-to-image) similarity:

$$\mathcal{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_k\} = \text{NN}_k(\mathbf{y}; \mathcal{Z}). \quad (13)$$

Then, following (4), for each proxy image $\mathbf{y}_j$ in the embedding space, the n nearest neighbors in $\mathcal{V}$ and the associated words in the text space are found

$$W_j = \{w_{j1}, \dots, w_{jn}\} = g^{-1}(\text{NN}_n(\mathbf{y}_j; \mathcal{V})). \quad (14)$$

From the union $W^+ = \cup_{j=1}^k W_j$ with $|W^+| \leq nk$ because of repeating words w_{ji} , the m most frequent words $\hat{W} = \{\hat{w}_1, \dots, \hat{w}_m\}$ are selected and the $\hat{a}_i$ is defined as the frequency associated with word $\hat{w}_i$. We write this filtered set of words as $\hat{W} = \phi_{X,V}^+(y)$ as a function of the image query y , where $\phi_{X,V}^+$ is called *expanded NN-inversion*. Finally, h can be defined via either early fusion

$$h_{E^+}(y, t) := \beta_E(\phi_{X,V}^+(y), t), \quad (15)$$

where β_E is given by (10), or by late fusion

$$h_{L^+}(y, t) := \beta_L(\phi_{X,V}^+(y), t, \mathbf{1}) \quad \text{and} \quad (16)$$

$$h_{L^+_\alpha}(y, t) := \beta_L(\phi_{X,V}^+(y), t, \hat{\mathbf{a}}), \quad (17)$$

with uniform and frequency weights $\mathbf{1}, \hat{\mathbf{a}} \in \mathbb{R}^m$, respectively, and β_L is given by (12). The last expression (17) is the complete FREEDOM method, summarized in Figure 2.

4. Experiments

4.1. Datasets, networks and evaluation protocol

We target a range of domain conversion applications where classes can be defined at the *category level* [31, 93, 95] or *instance level* [21], and domains can be defined as *styles* [21, 31, 95] or *context/environment* [93].

Datasets for domain conversion. ImageNet-R [31] has renditions of 200 ImageNet-1k [17] classes comprising 30,000 images. Following Pic2Word, we consider four domains: *cartoon*, *origami*, *sculpture*, and *toy*. We extend the benchmark by using every domain as the source.

MiniDomainNet [95] is a subset of DomainNet [65] with about 140,000 images, 126 classes, and four domains: *clipart*, *painting*, *real*, and *sketch*. Although this is a classification dataset, we adapt it for retrieval by using the official test set as our query set and the rest as the database.

Nico++ [93] is an out-of-distribution classification dataset with 88,866 real photographs, 60 categories, and six domains: *autumn*, *dim light*, *grass*, *outdoor*, *rock*, and *water*. We adapt this dataset for retrieval by using a random 10% as a query set and the rest as the database.

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.82	0.63	0.68	0.78	0.78	0.74
Image	4.27	3.12	0.84	5.86	5.08	3.84
Text $\times$ Image	8.21	5.62	6.98	8.95	9.41	7.83
Text + Image	6.61	4.45	2.17	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE	18.11	9.02	9.94	17.26	15.83	14.04
MagicLens	7.79	6.33	11.02	9.94	10.57	9.13
FREEDOM	35.97	11.80	27.97	36.58	37.21	29.91

METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text $\times$ Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.46	6.58	9.22	11.91	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE	13.49	13.73	17.91	17.99	15.79	11.84	15.13
MagicLens	18.76	15.17	22.14	23.61	21.99	16.30	19.66
FREEDOM	24.35	24.41	30.06	30.51	26.92	20.37	26.10

METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.38	7.78	6.66
Text $\times$ Image	9.01	8.66	15.87	5.90	9.86
Text + Image	9.59	9.97	9.22	8.53	9.33
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE	25.04	18.72	23.75	19.61	21.78
MagicLens	24.40	17.54	28.59	9.71	20.06
FREEDOM	41.96	31.65	41.12	34.36	37.27

METHOD	TODAY	ARCHIVE	AVG
Text	5.28	6.16	5.72
Image	8.47	24.51	16.49
Text $\times$ Image	16.42	29.90	23.16
Text + Image	9.60	26.13	17.86
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE	20.82	30.10	25.46
MagicLens	33.77	14.65	24.21
FREEDOM	30.95	35.52	33.24

Table 1. *Domain conversion mAP (%)* on four datasets. For each source domain (columns), we report the average mAP over every target domain. All methods are run by us. AVG: average mAP over all source-target domain combinations. **Bold**: best, **magenta**: second-best.

Large time lags location (LTLL) [21] contains images of 25 locations taken over a range of more than 150 years: 225 historical and 275 modern. We repurpose the dataset for retrieval by defining two domains: *today* and *archive*.

Network. We use the OpenAI pre-trained CLIP with a ViT-L/14 image encoder [18].

Evaluation Protocol. Unlike Precision@ k and Recall@ k , commonly used in literature and focusing on specific points in a ranked list, mean Average Precision (mAP) provides a more comprehensive assessment by considering precision across the entire ranking. Thus, we choose mAP as our evaluation metric. We use Recall@ k only to compare our method with existing methods whose implementation or models are not publicly available.

Text/visual memory and FREEDOM hyper-parameters. We use the 20k words of the Open Images V7 dataset [46] as our text memory. It is sufficiently large and is used in zero-shot recognition [12] and by SEARLE [2], allowing direct comparison between the methods. Unless otherwise stated, we use $k = 20$, $n = 7$, $m = 7$. We use the image database as visual memory. See the Supplementary for experiments with an external visual memory.

4.2. Simple Baselines

We show the following simple baselines for comparison. Advanced baselines can be found in the Supplementary.

Unimodal. Unimodal-query baselines rely only on similarity using one of the query modalities: *text-only* by $h_T(y, t) := g(t)$ and *visual-only* by $h_V(y, t) := f(y)$, referred to as “Text” and “Image” respectively in Table 1. Both are expected to fail as the final similarity misses as-

METHOD	# PARAM	AVG	
		R@10	R@50
Pic2Word [74]	429M	10.05	23.23
Pic2Word (CC-3M) [74]	429M	10.19	22.70
Pic2Word (LAION 2B-en) [74]	429M	10.35	22.91
ARTEMIS w/ CompoDiff dataset [13]	33M	12.24	23.58
CLIP4Cir w/ CompoDiff dataset [3]	178M	12.11	23.79
CompoDiff (T5-XL) [27]	3B	11.14	21.76
CompoDiff (CLIP+T5-XL) [27]	3.6B	13.31	24.86
CompoDiff (CLIP) [27]	568M	13.18	24.86
Context-I2W [81]	496M	12.90	27.60
KEDs [80]	428M	18.00	35.40
MagicLens ² (original prompt) [†] [90]	465M	8.61	21.60
MagicLens [†] [90]	465M	11.34	26.78
WeiCom [†] [66]	428M	12.17	23.04
SEARLE (default) [†] [2]	428M	2.36	14.24
SEARLE [†] [2]	428M	11.76	30.96
CIReVL [44]	12.5B	23.75	43.05
FREEDOM	428M	27.54	47.05

Table 2. *Domain conversion Recall@ k (%)* on ImageNet-R. AVG: average performance over all target domains and “photo” as the source. Top: methods that require training. Bottom: training-free methods. [†]: run by us. **Bold**: best, **magenta**: second-best.

pects of the composed query.

Product. This baseline combines the two unimodal approaches by using the product of the corresponding similarities. It is referred to as “Text $\times$ Image” in Table 1.

Sum. A common baseline in the literature that combines the two unimodal approaches by summation, i.e. $h_S(y, t) := g(t) + f(y)$, referred to as “Text + Image” in Table 1. The problem is that the text and image embeddings follow very different distributions.

²The original work reports results in an incorrect way (recall defined differently [78]: equal to 1 if at least one relevant example is retrieved), which we confirm via code inspection and communication with the author.

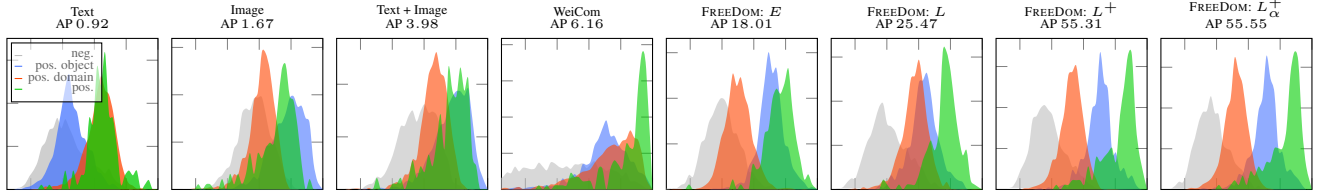


Figure 3. *Histogram of similarities* between a query and database images: **negative** (wrong object and domain); **positive** only w.r.t. the **object** (correct object, wrong domain); **positive** only w.r.t. the **domain** (wrong object, correct domain); **positive** (correct object and domain). *E*: early fusion; *L*: late fusion; *L⁺*: late fusion with memory-based expansion; *L_α⁺*: late fusion with memory-based expansion and frequencies as weights; AP: average precision. For better visualization, we sample an equal number of negatives, positives w.r.t. object, and positives w.r.t. domain, while the values in the histogram of positives are multiplied by 10. MiniDomainNet; text query: “clipart”.

4.3. Competitors

We extensively compare FREEDOM to recently proposed zero-shot composed image retrieval methods: Pic2Word [74], WeiCom [66], CompoDiff [27], SEARLE [2], and MagicLens [90]. The same ViT-L/14 [18] image encoder is used for all competitors. CompoDiff uses the text encoder of a ViT-G/14 [18]. Although FREEDOM uses the target domain’s name as a textual query for every dataset, we tailor the textual query (prompt) per dataset for many of the competing methods. All methods are run and evaluated by us. Details for these methods are provided in the Supplementary.

4.4. Experimental results

Comparison with SoTA. FREEDOM outperforms all baselines and competitors (Table 1) by a large margin. In particular, FREEDOM outperforms the second-best method on ImageNet-R by 15.87% mAP, on MiniDomainNet by 14.32%, on NICO++ by 6.44%, and on LTLL by 6.64%. On ImageNet-R, the second-best method is SEARLE. On MiniDomainNet, CompoDiff and SEARLE are the second and third best, respectively, while WeiCom performs lower than the simple baselines “Text × Image” and “Text + Image”. On NICO++, SEARLE and MagicLens are the only competitors that beat the simple baselines for more than 1%, while Pic2Word performs lower than the simple baseline “Text × Image”. On LTLL, WeiCom is the second-best method. Interestingly, MagicLens is the best method for “today” as the source, but for “archive”, it performs worse than the unimodal baseline “Image”. The baseline “Text × Image” performs higher than Pic2Word and CompoDiff.

We conduct an additional experiment using SigLIP [89] as a backbone. We get an mAP increase of 11.95, 5.72, 16.36, and 4.13 on ImageNet-R, NICO++, MiniDomainNet, and LTLL, respectively; details in the Supplementary.

Additionally, in Table 2, we compare FREEDOM with previous methods that evaluate with Recall@*k* on ImageNet-R. FREEDOM is the top-performing method among competitors that either requires training or is formed via an ensemble of several foundational models.

Qualitative analysis. Figure 3 shows the histogram of sim-

ilarities between a single query and the positives and negatives of different kinds of images. The unimodal baselines fail as expected. The sum baseline gives high importance to the image query. The normalized WeiCom improves that to a small extent but performs poorly as well. Early fusion improves further but gives too much importance to the object since several text labels are merged with a single word for the domain. Late fusion significantly improves this imbalance. The use of proxy images and weighting further boosts the performance, visualized by the blue histogram moving to the right relative to the orange.

4.5. Ablation study

Impact of hyper-parameters. Table 3 shows the impact of the number *k* of proxy images and the number of nearest words *n* from the vocabulary. The experiment is performed for a fixed number of text labels, *m* = 7. None of the combinations with *k* = 1 or *n* = 1 is the best. Therefore, the two steps of the nearest neighbor search are meaningful. Note that *k* = 1 corresponds to no expansion. The values *k* = 20 and *n* = 7 perform well across all datasets. In Table 4, we show the impact of varying *m*, demonstrating the benefits of going beyond one text label. The performance is stable for a large range of values.

Method components. In Table 4, we show the impact of our components by adding them one by one, leading to the final method. Additionally, we compare to SEARLE, which performs test-time optimization of a learnable word token. This is a direct comparison between inversion in the discrete input space of text and the continuous latent space of word tokens. SEARLE performs best if a single concept is used; concepts are similar to the text labels of our method. Our simplest variant with a single text label performs better than SEARLE on three out of four datasets, showcasing the benefit of our textual inversion in discrete words. FREEDOM benefits by using additional text labels and outperforms SEARLE on all datasets. Late fusion significantly boosts performance compared to early fusion, and so does expansion to proxy images. Using weights is beneficial on average and only slightly harmful on LTLL.

$k \backslash n$	AVG					IMAGENET-R					MINIDN					NICO++					LTLL				
	1	7	15	30	45	1	7	15	30	45	1	7	15	30	45	1	7	15	30	45	1	7	15	30	45
1	25.1	28.0	28.0	28.0	28.0	26.2	25.8	25.8	25.8	25.8	30.2	32.1	32.1	32.1	32.1	19.4	23.2	23.2	23.2	23.2	24.5	30.8	30.8	30.8	30.8
10	29.6	31.5	31.2	30.9	30.1	29.1	30.1	29.7	28.1	26.9	35.4	36.7	36.2	35.4	34.7	23.6	25.7	25.8	25.6	25.3	30.1	33.5	33.1	34.3	33.6
20	29.6	31.6	31.4	30.4	29.3	29.2	29.9	29.3	27.4	25.9	36.2	37.3	36.8	35.8	35.2	24.2	26.1	26.2	26.1	25.7	28.6	33.2	33.2	32.4	30.3
30	29.5	30.7	30.7	29.5	28.1	29.0	29.8	29.1	27.1	25.5	36.6	37.6	37.1	36.2	35.4	24.5	26.3	26.4	26.3	25.9	27.9	29.1	30.0	28.2	25.7
40	28.9	29.4	29.3	27.9	26.8	28.6	29.5	28.9	26.7	24.8	36.7	37.7	37.2	36.2	35.3	24.7	26.4	26.5	26.4	26.1	25.4	24.1	24.7	22.1	20.8
50	27.9	28.7	28.6	26.9	25.9	28.1	29.2	28.6	26.2	24.1	36.8	37.8	37.4	36.3	35.4	24.8	26.5	26.6	26.5	26.1	21.8	21.2	21.7	18.5	18.1

Table 3. The effect of the number of proxy images k vs. the number of labels n for each proxy in FREEDOM, measured in domain conversion mAP. Values $k \in \{10, 20\}$ and $n \in \{7, 15\}$ are better on average performance and competitive across datasets. Fixed $m = 7$ used.

m	AVG					IMAGENET-R					MINIDN					NICO++					LTLL				
	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
E	25.1	25.5	21.3	19.4	-	26.2	24.2	17.5	15.1	-	30.2	31.5	26.2	23.0	-	19.4	16.3	12.2	11.1	-	24.5	29.8	29.3	28.3	-
E+	26.9	28.0	23.2	20.6	-	28.5	27.2	19.2	16.4	-	34.9	35.4	28.6	24.8	-	22.3	18.1	13.3	12.0	-	22.0	31.3	31.6	29.2	-
L	25.1	28.1	28.0	27.4	26.3	26.2	27.3	25.8	24.7	23.0	30.2	33.0	32.1	30.9	28.9	19.4	22.7	23.2	22.8	21.9	24.5	29.3	30.8	31.2	31.6
L+	26.9	30.7	31.3	30.5	28.4	28.5	29.8	29.1	27.8	25.3	34.9	37.6	36.3	34.9	32.0	22.3	25.5	25.7	25.2	23.9	22.0	29.9	34.2	34.2	32.5
W	26.9	30.6	31.6	31.1	28.5	30.2	29.9	29.4	28.4	34.9	37.7	37.3	36.8	36.2	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.2	34.0	33.8	

Table 4. The impact of the number of selected labels m on each FREEDOM component, measured in domain conversion mAP. SRL: comparison to SEARLE that uses m words from the vocabulary to guide the textual inversion optimization. Components: E : early fusion; L : late fusion; E^+ : early fusion with memory-based expansion; L^+ : late fusion with memory-based expansion; W : late fusion with memory-based expansion and frequency-based weighting. Experiments with $k = 20$ (if memory-based expansion is used) and $n = 7$.

4.6. Oracle experiments

Information injection. In the inversion problem, a common challenge arises when the source domain appears within the text labels, causing conflicting domains in the query composition (e.g., “cartoon origami”). Conversely, the correct query class or the source domain may not be found in the m text labels. To study the impact of each, we conduct an oracle experiment and summarize the results in Table 5. In the first two columns (Object Gain), we compare early and late fusion with memory-based inversion after adding the name of the ground-truth class of the query to the m text labels. Late fusion achieves almost twice the gain compared to early fusion, showing that it can benefit more from the correct information. In the last two columns (Domain Loss), we compare the performance after including the source domain as a distractor in the text labels. Late fusion suffers almost half the loss compared to early fusion, showing more robust behavior to incorrect information.

Sensitivity to the vocabulary. We explore whether FREEDOM strongly depends on having the most appropriate word for each query class in the vocabulary. To reflect that, we perform an oracle experiment where the name of the ground truth class of the query image is used to remove its ℓ nearest words from the vocabulary. After removing the $\ell = 5$ most relevant words for each image query, FREEDOM performs 23.52, 31.95, 23.58, and 30.81 mAP on ImageNet-R, MiniDomainNet, NICO++, and LTLL, respectively. Despite the lack of the most appropriate words, FREEDOM is still the best method. More details are in the Supplementary.

Performance upper-bound. We use the ground-truth class of each image-query as a single text label. In this case, FREEDOM achieves 46.58, 34.00, 46.06, and 31.18 mAP on ImageNet-R, MiniDomainNet, NICO++, and LTLL, respectively, indicating space for improvement. However, this oracle experiment underperforms on LTLL, as class names do not always represent the object in this benchmark.

DATASET	OBJECT GAIN		DOMAIN LOSS	
	E^+	L^+	E^+	L^+
ImageNet-R	+0.72	+1.16	-4.30	-1.64
NICO++	+0.21	+0.40	-0.66	-1.29
MiniDN	+0.72	+0.59	-7.43	-2.79
LTLL	+1.56	+4.00	-2.86	-1.58
Avg	+0.80	+1.54	-3.81	-1.83

Table 5. Oracle experiment to study the impact of inliers and outliers in the m text labels. The inlier represents the query object name, and the outlier represents the source domain name. Gain or loss by adding inliers or outliers in text labels, respectively, is reported. Under memory-based expansion (E^+ and L^+), late fusion benefits more from inliers and is more robust to outliers.

5. Conclusions

We introduced FREEDOM, a training-free composed image retrieval method for domain conversion based on a pre-trained CLIP model. The key component is the textual inversion of the query image based on soft assignment to a sparse vocabulary of words. Our detailed ablations show the importance of every component of the method and its robustness to the choice of hyper-parameters. We also introduced three new benchmarks with different domain types, providing a broad testbed for future research. Despite its zero supervision, data, or training requirements, FREEDOM outperforms the state-of-the-art methods by a large margin.

Acknowledgments: This work was supported by the Junior Star GACR GM 21-28830M, the Czech Technical University in Prague grant No. SGS23/173/OHK3/3T/13, the programme Johannes Amos Comenius CZ.02.01.01/00/22_010/0003405, the CTU institutional support (Future fund RVO13000), the RAMONES H2020 project (grant: 101017808), and the HFRI under the BiCUBES project (grant: 03943). The access to the computational infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics” is gratefully acknowledged.

References

- [1] Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. *ICLR*, 2024. 2
- [2] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 2023. 2, 3, 4, 6, 7
- [3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, 2022. 1, 2, 6
- [4] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 1, 2
- [5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 2, 3
- [7] Ushasi Chaudhuri, Biplab Banerjee, and Avik Bhattacharya. Siamese graph convolutional network for content based remote sensing image retrieval. *Computer vision and image understanding*, 2019. 1
- [8] Junyang Chen and Hanjiang Lai. Pretrain like your inference: Masked tuning improves zero-shot composed image retrieval. In *arXiv*, 2023. 2
- [9] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *WACV*, 2024. 3
- [10] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *ECCV*, 2020. 1, 2
- [11] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *CVPR*, 2020. 1, 2
- [12] Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. “this is my unicorn, fluffy”: Personalizing frozen vision-language representations. In *ECCV*, 2022. 6
- [13] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *ICLR*, 2022. 2, 6
- [14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1
- [15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 2018. 2
- [16] Sounak Dey, Pau Riba, Anjan Dutta, Josep Lladós, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, 2019. 3
- [17] Wei Dong, Richard Socher, Li Li-Jia, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 5
- [18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6, 7
- [19] Yongchao Du, Min Wang, Wengang Zhou, Shuping Hui, and Houqiang Li. Image2sentence based asymmetrical zero-shot composed image retrieval. In *ICLR*, 2023. 2
- [20] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019. 3
- [21] Basura Fernando, Tatiana Tommasi, and Tinne Tuytelaars. Location recognition over large time lags. *Computer Vision and Image Understanding*, 2015. 1, 5, 6
- [22] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NeurIPS*, 2013. 1, 2
- [23] Anibal Fuentes and Jose M Saavedra. Sketch-qnet: A quadruplet convnet for color sketch-based image retrieval. In *CVPR*, 2021. 3
- [24] Sonam Goenka, Zhaozheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *CVPR*, 2022. 2
- [25] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, 2016. 1, 2
- [26] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 2
- [27] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yooheon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *TMLR*, 2024. 2, 6, 7
- [28] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yooheon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. In *CVPR*, 2024. 3
- [29] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *ICCV*, 2017. 1, 2
- [30] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *PAMI*, 2018. 3
- [31] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 1, 5
- [32] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. 3

- [33] Mehrdad Hosseinzadeh and Yang Wang. Composed query image retrieval using locally bounded features. In *CVPR*, 2020. 1, 2
- [34] Conghui Hu and Gim Hee Lee. Feature representation learning for unsupervised cross-domain image retrieval. In *ECCV*. Springer, 2022. 3
- [35] Conghui Hu, Yongxin Yang, Yunpeng Li, Timothy M Hospedales, and Yi-Zhe Song. Towards unsupervised sketch-based image retrieval. *BMVC*, 2021. 3
- [36] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, 2013. 3
- [37] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015. 2, 3
- [38] Sarah Ibrahimi, Nanne van Noord, Zeno Geradts, and Marcel Worring. Deep metric learning for cross-domain fashion instance retrieval. In *ICCVW*, 2019. 3
- [39] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 1, 2
- [40] Young Kyun Jang, Donghyun Kim, Zihang Meng, Dat Huynh, and Ser-Nam Lim. Visual delta generator with large multi-modal models for semi-supervised composed image retrieval. In *CVPR*, 2024. 2
- [41] Xin Ji, Wei Wang, Meihui Zhang, and Yang Yang. Cross-domain image retrieval with attention modeling. In *ACM Multimedia*, 2017. 3
- [42] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 2, 3
- [43] Xintong Jiang, Yaxiong Wang, Yujiao Wu, Meng Wang, and Xueming Qian. Dual relation alignment for composed image retrieval. *ICLR*, 2024. 2
- [44] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *arXiv*, 2023. 2, 3, 6
- [45] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cds: Cross-domain self-supervised pre-training. In *ICCV*, 2021. 3
- [46] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 4, 6
- [47] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *CVPR*, 2021. 1, 2
- [48] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and early fusion for composed image retrieval. *arXiv*, 2023. 2
- [49] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023. 3
- [50] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1, 2, 3
- [51] Jiahao Li, Greg Shakhnarovich, and Raymond A Yeh. Adapting clip for phrase localization without further training. *arXiv*, 2022. 3
- [52] Xiujuan Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2
- [53] Yi Li, Timothy M Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*, 2014. 3
- [54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2
- [55] Yikun Liu, Jiangchao Yao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Zero-shot composed text-image retrieval. *BMVC*, 2023. 2
- [56] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 1
- [57] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2021. 2
- [58] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *WACV*, 2024. 2
- [59] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *ICCV*, 2023. 3
- [60] Tianyi Lu, Xing Zhang, Jiaxi Gu, Hang Xu, Renjing Pei, Songcen Xu, and Zuxuan Wu. Fuse your latents: Video editing with multi-source latent diffusion models. *ACM Multimedia*, 2024. 3
- [61] Lakshmi R Nair, Kamalraj Subramaniam, and GKD Prasannan. A review on multiple approaches to medical image retrieval system. *Intelligent Computing in Engineering*, 2019. 1
- [62] Andrei Neculai, Yanbei Chen, and Zeynep Akata. Probabilistic compositional embeddings for multimodal image retrieval. In *CVPR*, 2022. 1, 2
- [63] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017. 1, 2
- [64] Soumava Paul, Titir Dutta, and Soma Biswas. Universal cross-domain retrieval: Generalizing across classes and domains. In *ICCV*, 2021. 3
- [65] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 5
- [66] Bill Psomas, Ioannis Kakogeorgiou, Nikos Efthymiadis, Giorgos Toliás, Ondrej Chum, Yannis Avrithis, and Konstantinos Karantzas. Composed image retrieval for remote sensing. In *IGARSS*, 2024. 6, 7

- [67] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *PAMI*, 2019. 1
- [68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3
- [69] Tal Reiss, Bar Cavia, and Yedid Hoshen. Detecting deepfakes without seeing any. *arXiv*, 2023. 3
- [70] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 2
- [71] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2
- [72] Jose M Saavedra. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (shelo). In *ICIP*, 2014. 3
- [73] Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based image retrieval using learned keyshapes (lks). In *BMVC*, 2015. 3
- [74] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*, 2023. 1, 2, 3, 4, 6, 7
- [75] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *TOG*, 2016. 3
- [76] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *ICCV*, 2019. 1, 2
- [77] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 2, 3
- [78] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016. 6
- [79] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017. 3
- [80] Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *CVPR*, 2024. 2, 6
- [81] Yuanmin Tang, Jing Yu, Keke Gai, Zhuang Jiamin, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. *AAAI*, 2024. 3, 6
- [82] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, 2019. 1, 2
- [83] Xu Wang, Dezhong Peng, Ming Yan, and Peng Hu. Correspondence-free domain alignment for unsupervised cross-domain image retrieval. In *AAAI*, 2023. 3
- [84] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *CVPR*, 2021. 1, 2
- [85] Monika Wysoczańska, Michaël Ramamonjisoa, Tomasz Trzciński, and Oriane Siméoni. Clip-diy: Clip dense inference yields open-vocabulary semantic segmentation for-free. In *WACV*, 2024. 3
- [86] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 3
- [87] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *ECCV*, 2020. 1, 2
- [88] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016. 3
- [89] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 7
- [90] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. MagicLens: Self-supervised image retrieval with open-ended instructions. In *ICML*, 2024. 3, 6, 7
- [91] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 2
- [92] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *CVPR*, 2020. 1, 2
- [93] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyang Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *CVPR*, 2023. 1, 5
- [94] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 3
- [95] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *ICLR*, 2021. 5
- [96] Xiaoping Zhou, Xiangyu Han, Haoran Li, Jia Wang, and Xun Liang. Cross-domain image retrieval: methods and applications. *International Journal of Multimedia Information Retrieval*, 2022. 3