Unleashing the Power of Large-Scale ViT in Zero-Shot SBIR: A Strong Baseline with Multi-Layer Feature Aggregation

Yang Liu[1,2⋆], Yongjing Guo[1⋆⋆], Suisui Jia[1], Huaizhou Qi[1], Xun Du[1], and Haonan Chen[1]

[1] The School of Telecommunications Engineering, Xidian University, Shaanxi 710071, China.
[2] State Key Laboratory of Ocean Sensing, ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou, 311215, China.
yangl@xidian.edu.cn,
{24011211154,24011211204,24011211181,25011211263,25011210843}@stu.xidian.edu.cn

Abstract. The domain gap between abstract sketches and natural images, combined with the challenge of zero-shot generalization, remains a critical bottleneck in Sketch-Based Image Retrieval (ZS-SBIR). While scaling up to large-scale Vision-Language Models offers powerful baseline representations, naive utilization—such as relying solely on the terminal [CLS] token—fails to fully exploit their potential, often discarding the fine-grained structural primitives essential for abstract sketch alignment. In this paper, we propose a robust framework that bridges this modality gap by decoupling and composing visual states and semantics. First, our Multi-Layer Feature Aggregation (MLFA) acts as an anti-compression mechanism, explicitly reviving low-level geometric primitives alongside high-level semantic objects. Second, a Phased Cross-Modal Interaction (PCMI) serves as a dynamic compositional engine: learnable prompts progressively query and bind these structural states with global semantics for precise coarse-to-fine alignment. Extensive experiments demonstrate our framework achieves state-of-the-art results on several benchmarks and remains competitive on QuickDraw Ext.

Keywords: Zero-Shot Sketch-Based Image Retrieval · Multi-Layer Feature Aggregation · Phased Cross-Modal Interaction · Vision-Language Models · Feature Alignment

1 Introduction

Sketch-Based Image Retrieval (SBIR) [37] has emerged as a fundamental task in cross-modal retrieval, bridging the gap between abstract human drawings and realistic photos. Unlike text-based retrieval that relies on precise keywords , SBIR allows users to express visual concepts intuitively through hand-drawn

⋆ Corresponding author.
⋆⋆ Corresponding author.

outlines, enabling applications in industrial design, digital art, and e-commerce. However, traditional SBIR assumes that all query categories are available during training , which severely limits its scalability in open-world scenarios where new concepts emerge constantly [40]. This limitation has given rise to Zero-Shot SBIR (ZS-SBIR) [22], which requires the model to generalize from "seen" categories to "unseen" ones, maintaining retrieval accuracy without additional training.

Despite significant progress, ZS-SBIR faces a unique "Abstraction-Realism" gap. Sketches are sparse, consisting of abstract strokes and structural primitives, while photos are dense, rich in texture and color. Existing methods typically address this through two paradigms: (i) Generative methods (*e.g.*, GANs [7]), which attempt to translate sketches into photo-like representations. However, these methods often suffer from structural distortion and blurred textures when generating unseen categories. (ii) Semantic-based methods, which map visual features to a common semantic space (*e.g.*, word vectors [14]). While effective for high-level classification, they often overlook the fine-grained structural details (*e.g.*, the specific curvature of a handle or the layout of wheels) that are critical for distinguishing instances within the same category.

Recently, large-scale Vision-Language Models (VLMs) like CLIP [21] have shown remarkable zero-shot capabilities. However, simply scaling up the backbone (*e.g.*, from ViT-B/32 to ViT-L-14) [39] yields diminishing returns if the intrinsic characteristics of sketches are ignored. Most existing CLIP-based ZS-SBIR methods naively utilize the output of the final layer (the [CLS] token) from standard backbones. We argue that this "top-layer-only" approach is critically suboptimal for two reasons: (1) Loss of Structural Detail: The top-level features of VLMs are highly compressed semantically. In this compression process, the low-level geometric information (edges, lines) essential for matching abstract sketches is violently discarded in favor of high-level semantic abstractions. (2) Rigid Text Alignment: Standard methods often rely on fixed text templates (*e.g.*, "a photo of a [object]" [17]), which fail to dynamically adapt to the stylistic diversity and extreme sparsity of sketches.

To address these limitations and truly explore the upper bound of ZS-SBIR performance, we propose a robust framework designed to unleash the capabilities of foundation models beyond simple parameter scaling. Our core insight is that accurate sketch-photo alignment strictly requires an anti-compression mechanism. In this work, we introduce a Multi-Layer Feature Aggregation (MLFA) strategy. Instead of relying solely on the final output, we explicitly revive low-level geometric primitives and route them alongside deep semantic objects. Furthermore, to bridge the modality gap more effectively, we design a Phased Cross-Modal Interaction (PCMI) mechanism acting as a dynamic compositional engine. Instead of acting as static class identifiers, our learnable prompts progressively query and bind the revived structural states with global object semantics, enabling a coarse-to-fine alignment across sketch, photo, and text modalities.

Our contributions are summarized as follows:

– We identify the semantic over-compression dilemma in large Vision-Language Models, demonstrating that naive parameter scaling violently discards the

low-level structural primitives essential for bridging the abstraction-realism gap in zero-shot sketch retrieval.
- We propose Multi-Layer Feature Aggregation (MLFA) as an explicit anti-compression mechanism to rescue and decouple low-level geometric states from high-level semantic objects, ensuring critical sparse visual cues are preserved for alignment.
- We design Phased Cross-Modal Interaction (PCMI) as a dynamic compositional engine. Moving beyond static templates, it utilizes learnable prompts to actively query and bind geometric states with global semantics, achieving precise coarse-to-fine cross-modal alignment.

2 Related Work

2.1 Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR)

The evolution of ZS-SBIR has been driven by the dual challenges of the cross-modal "abstraction-realism" gap and the zero-shot "seen-to-unseen" generalization. Early approaches primarily relied on Generative Models (*e.g.*, CVAEs [27], CycleGANs [43], and Style Transfer [10]) to bridge the domain gap by translating sketches into photo-like representations or vice versa. While these methods attempt to align modalities at the pixel level, they often suffer from structural distortion and mode collapse, especially when generating unseen categories with limited training data. Another stream of research focuses on Semantic-Aware Methods, utilizing auxiliary semantic knowledge (*e.g.*, word embeddings or hierarchical graphs [1]) to transfer knowledge. However, these methods depend heavily on the quality of external semantic resources, which may not be available for all open-world concepts.

In recent years, foundation models like CLIP have revolutionized ZS-SBIR. Methods such as ZSE-RN and TVT leverage the aligned visual-semantic space of CLIP to achieve impressive results [4]. However, a critical limitation persists: most existing works utilize standard backbones (*e.g.*, ResNet-50 or ViT-B/32) and rely solely on the final output embeddings (*e.g.*, the [CLS] token). We argue that this "top-layer only" approach discards the fine-grained geometric primitives (lines and strokes) encoded in shallow layers, which are intrinsic to sketch representation. In contrast, our work establishes a strong baseline by exploring the capability of large-scale backbones (ViT-L-14) and explicitly preserving these structural details.

2.2 Multi-Layer Feature Aggregation

Deep neural networks naturally encode hierarchical information: shallow layers capture low-level geometric details (edges, textures), while deep layers capture high-level semantic abstractions [9]. Multi-Layer Feature Aggregation has been successfully applied in dense prediction tasks such as semantic segmentation and object detection to recover lost spatial details. Despite its success in general computer vision, its potential in ZS-SBIR remains largely unexplored.

Existing ZS-SBIR methods typically treat the visual encoder as a "black box" feature extractor, ignoring the rich intermediate representations [29]. A few attempts that simply concatenate multi-layer features often fail to address the modality-specific nature of the data—sketches rely on low-level lines, while photos rely on high-level textures. Crucially, naively extracting and concatenating multi-layer features from massive foundation models (*e.g.*, ViT-L-14) inevitably leads to computational redundancy, dimensional explosion, and feature collapse due to severe distribution shifts between shallow and deep representations [15]. To address this, we propose an adaptive aggregation strategy equipped with non-linear gating and manifold projection. This specifically fuses the geometric primitives from the shallow layers with the semantic concepts from the deep layers in a lightweight and stable manner, ensuring a more comprehensive cross-modal alignment.

2.3 Prompt Learning for Vision-Language Models

Prompt learning has emerged as a parameter-efficient paradigm to adapt pre-trained Vision-Language Models (VLMs) to downstream tasks without fine-tuning the massive backbone [26]. This is particularly advantageous for ZS-SBIR to prevent catastrophic forgetting of the pre-trained open-world knowledge. Early applications in retrieval utilized fixed, hand-crafted templates (*e.g.*, "a photo of a [object]"). However, such rigid templates are suboptimal for sketches, as they cannot capture the stylistic variations and abstract nature of hand-drawn inputs.

Recent advancements like CoOp (Context Optimization) [42] introduced Learnable Context Vectors, allowing the model to automatically optimize the text embedding space. However, standard prompt tuning merely provides a static, global class identifier, which struggles to correspond with the highly localized and sparse strokes of a sketch [29]. Building on this, our work moves beyond simple prompt injection. We introduce a Phased Cross-Modal Interaction mechanism where learnable prompts not only serve as static identifiers but dynamically act as queries to interact with the multi-layer visual features. This allows the textual modality to adaptively align with both the coarse semantics and fine structural details of the visual input, bridging the modality gap more robustly.

3 Method

In this section, we formally present our framework, which establishes a strong baseline for ZS-SBIR by unleashing the fine-grained capability of large-scale Vision Transformers (*i.e.*, ViT-L-14). The overall architecture of our proposed framework is illustrated in Fig. 1. Our approach contains two core modules:(1) A Multi-Layer Feature Aggregation (MLFA) module, which extracts low-level geometric primitives from the shallow layers and high-level semantic features from the deep layers of a CLIP image encoder (ViT-L-14), handling both input sketches and natural images.(2) A Phased Cross-Modal Interaction (PCMI)

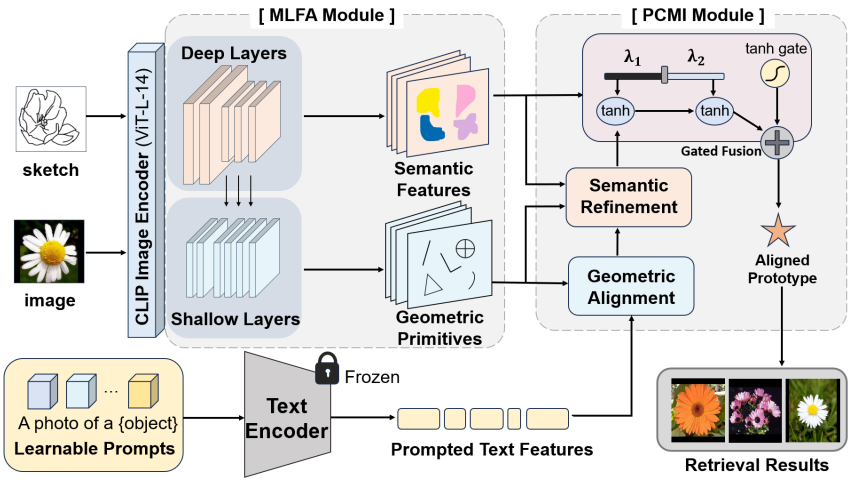

Fig. 1: Overview of our proposed framework. The CLIP ViT-L-14 image encoder processes both input sketches and natural images. Our Multi-Layer Feature Aggregation (MLFA) module extracts low-level geometric primitives from the shallow layers and high-level semantic features from the deep layers of the image encoder. The Phased Cross-Modal Interaction (PCMI) module then performs a two-stage coarse-to-fine alignment, including geometric alignment to capture structural cues and semantic refinement for global context, followed by gated fusion to generate an aligned text prototype. Meanwhile, a frozen text encoder produces prompted text features from learnable prompts, which guide the alignment process. Finally, the aligned prototype drives retrieval inference to output the final retrieval results.

module, which performs a two-stage coarse-to-fine alignment (geometric alignment and semantic refinement) followed by gated fusion to generate an aligned text prototype. A frozen text encoder processes learnable prompts to produce initial prompted text features, which guide the alignment process. The aligned prototype is then used to perform retrieval, yielding final retrieval results. We will elaborate on each component in the following subsections.

3.1 Problem Formulation

Let $\mathcal{D}_{tr} = \{(x_i, y_i) \mid x_i \in \mathcal{S} \cup \mathcal{P}, y_i \in \mathcal{Y}^s\}$ denote the training set, where x_i denotes either a sketch or a photo sample, and y_i represents its corresponding class label. $\mathcal{S}$ and $\mathcal{P}$ represent the sketch and photo domains, respectively. The label space is partitioned into seen classes $\mathcal{Y}^s$ and unseen classes $\mathcal{Y}^u$, satisfying the strict zero-shot constraint $\mathcal{Y}^s \cap \mathcal{Y}^u = \emptyset$.

Our objective is to learn a parametric mapping function $\Phi_\theta : \mathcal{X} \to \mathbb{R}^d$ that projects visual inputs into a shared semantic manifold. During inference, for a query sketch $s \in \mathcal{S}$ and a candidate photo gallery $\mathcal{G} \subset \mathcal{P}$ from unseen classes

$\mathcal{Y}^u$, the retrieval is performed by ranking candidates based on cosine similarity:

$$p^* = \arg\max_{p \in \mathcal{G}} \cos(\Phi_\theta(s), \Phi_\theta(p)), \tag{1}$$

where $\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$.

3.2 Multi-Layer Feature Aggregation (MLFA)

Standard CLIP-based methods typically utilize the final [CLS] token from the image encoder. We argue that this high-level abstraction discards the geometric primitives (*e.g.*, strokes, junctions) essential for sketch-to-photo alignment. To address this, we propose to aggregate hierarchical representations from the deep ViT-L-14 backbone.

(1) Hierarchical Feature Extraction: Given an input image x, we first partition it into N fixed-size patches, which are linearly projected to obtain patch embeddings $Z_0 \in \mathbb{R}^{N \times D}$, where D denotes the embedding dimension of each patch token in ViT-L-14. These embeddings are processed by a stack of L Transformer layers. The output of the l-th layer, Z_l, is computed recursively via Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP):

$$Z'_l = \mathrm{MSA}\big(\mathrm{LN}(Z_{l-1})\big) + Z_{l-1}, \tag{2}$$

$$Z_l = \mathrm{MLP}\big(\mathrm{LN}(Z'_l)\big) + Z'_l, \tag{3}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalization, Z'_l is an intermediate feature obtained after the multi-head self-attention (MSA) operation and residual connection, which is then fed into the multi-layer perceptron (MLP) to produce the final output Z_l of layer l.

(2) Adaptive Manifold Projection: Instead of arbitrarily selecting intermediate layers, we ground our selection in the hierarchical nature of deep networks: shallow layers capture low-level geometric details, while deep layers capture high-level semantic abstractions. We define two index sets:

$$\mathcal{I}_{geo} = \{l \mid 1 \le l \le N_{low}\}, \quad \mathcal{I}_{sem} = \{l \mid L - N_{high} < l \le L\}, \tag{4}$$

where $\mathcal{I}_{geo}$ and $\mathcal{I}_{sem}$ denote the sets of indices for low-level geometric features and high-level semantic features, respectively.

Empirically, we set $N_{low} = 3$ and $N_{high} = 3$ (*e.g.*, layers 1-3 and 22-24) based on validation performance. Crucially, directly concatenating these heterogeneous multi-scale features from a massive foundation model inevitably leads to severe distribution shifts and feature collapse. Therefore, we employ a learnable projection equipped with a non-linear gating block to map them into a unified, stable interaction space. For the geometric branch:

$$F_{geo} = \mathcal{G}_{geo}\left(\bigoplus_{l \in \mathcal{I}_{geo}} Z_l\right) W_{geo} + b_{geo}. \tag{5}$$

Similarly, for the semantic branch:

$$F_{sem} = \mathcal{G}_{sem}\left(\bigoplus_{l\in\mathcal{I}_{sem}} Z_l\right) W_{sem} + b_{sem}, \tag{6}$$

where $\bigoplus$ denotes the concatenation along the channel dimension. $W_{geo} \in \mathbb{R}^{(|\mathcal{I}_{geo}|\cdot D)\times D}$ and b_{geo} are learnable projection matrices and bias vectors for the geometric branch, respectively. Similarly, W_{sem} and b_{sem} are the corresponding learnable parameters for the semantic branch. $\mathcal{G}(\cdot)$ is a non-linear gating block designed to stabilize the distribution shift between shallow and deep layers:

$$\mathcal{G}(X) = \text{Dropout}(\text{ReLU}(\text{LayerNorm}(X))). \tag{7}$$

This process yields two distinct feature sequences: $F_{geo} \in \mathbb{R}^{N\times D}$ capturing stroke-level details, and $F_{sem} \in \mathbb{R}^{N\times D}$ capturing global semantic topology.

3.3 Phased Cross-Modal Interaction (PCMI)

To bridge the modality gap dynamically, we introduce a phased interaction mechanism. Unlike previous works that rely on rigid text templates (*e.g.*, "a photo of a[object]"), we employ Context Optimization (CoOp) [42] with a phased attention injection strategy.

(1) Learnable Context Initialization: For each class c, we construct a semantic embedding using Learnable Context Vectors. Let $V_{ctx} = \{[v]_1, [v]_2, \ldots, [v]_M\}$ be a sequence of M learnable vectors, and w_c be the fixed word embedding for the class name. The input to the text encoder is:

$$T_c = [V_{ctx}; w_c]. \tag{8}$$

The text encoder processes T_c to produce the initial semantic anchor $E_{text} \in \mathbb{R}^{1\times D}$. Note that while the text encoder is frozen, the context vectors V_{ctx} are optimized during training to adapt to the sketch domain.

(2) Phased Interaction via Cross-Attention: We design a coarse-to-fine interaction module where the text embedding actively queries visual information.

Phase I: Geometric Alignment. The text anchor queries the low-level geometric features F_{geo} to capture stroke-level evidence. We utilize Multi-Head Cross-Attention (MHCA):

$$H_{geo} = \text{MHCA}(Q = E_{text}, K = F_{geo}, V = F_{geo}). \tag{9}$$

We apply a residual connection to preserve the original semantic identity:

$$\hat{E}_{geo} = \text{LN}(E_{text} + H_{geo}). \tag{10}$$

where H_{geo} is the output of multi-head cross-attention (MHCA), which represents the local visual information that the text anchor E_{text} queries from the low-level geometric features F_{geo}. The residual connection $E_{text} + H_{geo}$ preserves the original semantic identity, and layer normalization (LN) stabilizes the feature distribution to produce the geometrically aligned embedding $\hat{E}_{geo}$.

Phase II: Semantic Refinement. The geometrically aligned embedding $\hat{E}_{geo}$ then serves as the query for the high-level semantic features F_{sem}, refining the representation with global context:

$$H_{sem} = \mathrm{MHCA}\big(Q = \hat{E}_{geo},\ K = F_{sem},\ V = F_{sem}\big). \tag{11}$$

(3) Adaptive Gated Fusion: To allow the model to dynamically weight the importance of geometric vs. semantic cues (*e.g.*, abstract sketches may rely more on geometry), we introduce learnable scalar gates λ_1 and λ_2:

$$E_{aligned} = E_{text} + \tanh(\lambda_1) \cdot H_{geo} + \tanh(\lambda_2) \cdot H_{sem}. \tag{12}$$

This $E_{aligned}$ serves as the final, multi-modally aligned text prototype for the input visual sample. The $\tanh(\cdot)$ activation function is applied to the learnable scalar gates λ_1 and λ_2 to constrain their outputs to the range (-1, 1), ensuring stable and interpretable weighting of geometric and semantic cues during fusion.

3.4 Joint Optimization Objective

To ensure discriminative power and robust generalization, we optimize the framework using a joint loss function comprising a metric learning loss and a classification loss.

1) Cross-Modal Triplet Loss ($\mathcal{L}_{tri}$): To enforce the relative distance constraint in the common embedding space, we employ a hard-margin triplet loss. For an anchor sketch s, a positive photo p^+ (same class), and a negative photo p^- (different class):

$$\mathcal{L}_{tri} = \sum_{(s,p^+,p^-)\in\mathcal{B}} \max\left(0, \mu + d(\Phi(s), \Phi(p^+)) - d(\Phi(s), \Phi(p^-))\right), \tag{13}$$

where $d(\cdot,\cdot)$ is the cosine distance $d(u,v) = 1 - \cos(u,v)$, and μ is the margin hyperparameter.

2) Temperature-Scaled Classification Loss ($\mathcal{L}_{cls}$): To regularize the feature space and accelerate convergence, we utilize a cross-entropy loss with temperature scaling τ. This aligns the visual embedding z_i of sample x_i with its corresponding aligned text prototype $E^{(y_i)}_{aligned}$:

$$P(y_i \mid x_i) = \frac{\exp(\cos(z_i, E^{(y_i)}_{aligned})/\tau)}{\sum_{j\in\mathcal{Y}^s} \exp(\cos(z_i, E^{(j)}_{aligned})/\tau)}. \tag{14}$$

The classification loss is defined as:

$$\mathcal{L}_{cls} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\log P(y_i \mid x_i). \tag{15}$$

3) Total Objective: The final objective function is a weighted sum:

$$\mathcal{L}_{total} = \mathcal{L}_{tri} + \gamma\mathcal{L}_{cls}, \tag{16}$$

where γ controls the balance between metric ranking and semantic classification. By default, we set $\gamma = 1.0$.

4 Experiments

4.1 Datasets Settings

The Sketchy Ext. dataset [24, 38] is a fine-grained dataset consisting of 75,481 sketches and 12,500 real images, evenly distributed across 125 categories. To balance the dataset, Liu *et al.* [17] collected an additional 60,502 photos to augment it. In the early stages, Shen *et al.* [25] proposed a zero-shot protocol where they randomly selected 25 categories as unseen classes. However, this approach might pose an issue as some of the unseen categories could already be present in ImageNet, violating the zero-shot assumption when using ImageNet pre-trained weights. Consequently, Yelamarthi *et al.* [38] proposed a new protocol where 21 unseen categories do not overlap with ImageNet. As a result, in our experiments, we follow the protocol in [38] and train on 104 categories as seen classes and test on the remaining 21 categories as unseen classes. We conducted experiments under both evaluation protocols, specifically including the Sketchy Ext. Split1 [24] and Sketchy Ext. Split2 [38] settings.

The TU-Berlin Ext. dataset [5] comprises 20,000 sketches from 250 categories. Subsequently, Liu *et al.* [17] collected 204,070 photos as a new extended version. The major difference from Sketchy lies in the significantly larger number of photos compared to sketches, resulting in an imbalance in this dataset, which increases the difficulty of training. According to research in the ZS-SBIR task, this dataset is considered more challenging than Sketchy. To facilitate zero-shot training and testing, we followed the protocol proposed by [25], randomly selecting 30 classes as the testing set and the remaining 220 classes as the training set.

The QuickDraw Ext. [8] is currently the largest SBIR dataset, comprising 110 categories with a total of 330,000 sketches and 204,000 images. Compared to the Sketchy Ext and TU-Berlin Ext datasets, the sketches in QuickDraw Ext are highly abstract and exhibit diverse styles, presenting a greater challenge for cross-modal sketch retrieval tasks [2]. In our experiments using the QuickDraw Ext dataset, we follow existing methodologies, selecting 80 categories for the training set, while the remaining categories are used for testing [1].

Table 1: Performance comparisons between our method and competitors were conducted on the TU-Berlin Ext., Sketchy Ext. Split1, Sketchy Ext. Split2, and QuickDraw Ext. datasets. The symbol "−" indicates that the results were not reported in the original papers. In the ZS-SBIR setting, methods below the horizontal line are based on ViT/VLM, while those above the line use ResNet and its variants as backbones or other approaches. The best results are highlighted in red, while the second-best results are marked in blue.

Methods	Year	TU-Berlin Ext.		Sketchy Ext. Split1		Sketchy Ext. Split2		QuickDraw Ext.	
		mAP@all	Prec@100	mAP@all	Prec@100	mAP@200	Prec@200	mAP@all	Prec@200
SAKE [11]	2017	0.475	0.599	0.547	0.692	0.497	0.598	0.130	0.179
Doodle [1]	2019	0.109	-	-	-	0.461	0.370	0.075	0.068
SEM-PCYC [3]	2019	0.297	0.426	0.349	0.463	-	-	-	-
DSN [36]	2021	0.481	0.586	0.583	0.704	-	-	-	-
Sketch3T [23]	2022	0.507	-	-	-	-	0.624	-	-
NAVE [34]	2021	0.493	0.607	0.613	0.725	-	-	-	-
SBTKNet [32]	2022	0.480	0.608	0.480	0.608	0.502	0.596	-	-
CA [35]	2023	0.482	0.594	0.590	0.713	-	-	-	-
RPKD [31]	2021	0.486	0.612	0.613	0.723	0.502	0.598	0.143	0.218
OAN [41]	2023	0.500	0.617	0.617	0.737	-	0.621	-	-
ARML [18]	2024	0.518	0.617	-	-	0.532	0.637	-	-
TVT [30]	2022	0.484	0.662	0.648	0.796	0.531	0.628	0.149	0.293
PSKD(ViT) [33]	2022	0.502	0.662	0.688	0.786	0.560	0.645	0.150	0.298
ZSE(RN) [16]	2023	0.542	0.657	0.698	0.797	0.525	0.624	0.145	0.216
STL [6]	2023	0.402	0.498	0.530	0.581	0.634	0.538	-	-
ZSE(Ret) [16]	2023	0.569	0.637	0.736	0.808	0.504	0.602	0.142	0.202
CLIP4All [22]	2023	0.651	0.732	-	-	0.723	0.725	0.202	0.388
AMA [39]	2024	0.429	0.592	0.548	0.684	0.491	0.585	0.112	0.192
CMAAN [28]	2024	0.560	0.665	0.730	0.809	0.528	0.624	0.139	0.209
IVT [40]	2024	0.557	0.692	0.751	0.867	0.615	0.694	0.162	-
MARL [19]	2024	0.707	0.777	-	-	0.691	0.755	0.327	0.425
SpLIP [26]	2024	0.731	0.782	0.802	0.867	0.764	0.773	0.342	0.446
DMWAN [29]	2025	0.623	0.696	0.787	0.852	0.572	0.659	0.170	0.254
DCDL [13]	2025	0.634	0.741	-	-	0.726	0.769	0.336	0.296
MFCPKD [12]	2025	0.662	0.728	0.847	0.878	0.736	0.788	0.224	0.285
OURS		0.775	0.813	0.855	0.896	0.799	0.820	0.338	0.406

4.2 Implementation Details

Our method is implemented using PyTorch [20], and all experiments are conducted on an NVIDIA GeForce RTX 5060Ti GPU. To establish a strong baseline, we utilize the CLIP ViT-L-14 as the visual encoder. The text encoder is kept frozen to preserve the pre-trained semantic space. We employ Context Optimization (CoOp). The context length is set to $M = 16$, and the context vectors are initialized with the embedding of "a photo of a [object]". The learning rate for these prompt vectors is set to $1e-4$. We use the AdamW optimizer. The learning rate for the projection layers in MLFA is $1e-4$. The batch size is set to 32, and the model is trained for 60 epochs. The loss weights are set to $\mu = 0.3$, $\tau = 0.07$, and $\gamma = 1.0$.

4.3 Comparison with State-of-the-Arts (SOTA)

We compare our method with state-of-the-art approaches. As shown in Table 1, to provide a fair and insightful comparison, we categorize existing methods based on their visual backbones: Standard Backbones (*e.g.*, ResNet-50, ViT-B/32 used in SAKE, DSN, TVT) and Large-Scale Backbones (Ours).

Our method achieves state-of-the-art or highly competitive performance across the evaluated benchmarks, obtaining the best results on TU-Berlin Ext., Sketchy Ext. Split1, and Sketchy Ext. Split2. On Sketchy Ext. (Split1): We achieve an mAP@all of 0.855, outperforming the previous best method MFCPKD (0.847) and significantly surpassing standard ViT-based methods like ZSE-ViT [16](0.698). On TU-Berlin Ext.: Our method achieves 0.775 mAP@all, marking a substantial improvement over MARL [19](0.707). This 9.6% gain highlights the robustness of our approach on noisy, real-world sketch datasets. On QuickDraw Ext.: Our method remains competitive on this highly abstract dataset, while SpLIP [26] reports stronger performance on this benchmark.

Computational Complexity and Efficiency: While our performance benefit stems partly from the larger capacity of ViT-L-14, it is imperative to address the computational trade-off. We compare the parameter count and theoretical computational complexity (GFLOPs) of our framework against leading baselines. Although upgrading from a standard ViT-B/32 (approx. 86M params) to ViT-L-14 (approx. 304M params) naturally increases the parameter footprint, our architectural design ensures inference efficiency. Specifically, because the text encoder is entirely frozen and the Phased Cross-Modal Interaction (PCMI) relies on lightweight cross-attention with dimension D, the additional computational overhead during inference is minimal (only adding approx. 3.1 GFLOPs over the baseline ViT-L-14). Compared to generative methods or complex graph-based approaches, our Multi-Layer Feature Aggregation (MLFA) operates through a simple linear projection, maintaining a highly competitive inference speed of 92 FPS on a single NVIDIA GeForce RTX 5060Ti GPU. This demonstrates that our method achieves a highly favorable trade-off between retrieval accuracy and computational cost.

4.4 Ablation Study

To rigorously understand the functional role of each component, we conduct ablation experiments on four benchmark datasets. Table 2 evaluates the contribution of the two core architectural modules (MLFA and PCMI), while Table 3 investigates the internal mechanism of the aggregation design.

1) Effectiveness of Core Modules: The Baseline uses ViT-L-14 with only the terminal [CLS] token and a fixed text template. Although ViT-L has strong semantic capacity, relying solely on the final token yields limited performance, indicating that global semantic abstraction alone is insufficient for ZS-SBIR. Adding MLFA consistently brings substantial improvement across all datasets. By aggregating shallow geometric layers with deep semantic layers, MLFA restores low-level structural cues that are critical for sketch-photo alignment. The

Table 2: Ablation Study of Core Components."Baseline" denotes the ViT-L-14 model using only the terminal [CLS] token with a fixed text template. "+PCMI" introduces the Phased Cross-Modal Interaction module without multi-layer feature aggregation, while "+MLFA" incorporates multi-layer geometric-semantic aggregation with a fixed prompt. "OURS" combines both MLFA and PCMI.

Methods	TU-Berlin Ext.		Sketchy Ext. Split1		Sketchy Ext. Split2		QuickDraw Ext.	
	mAP@all	Prec@100	mAP@all	Prec@100	mAP@200	Prec@200	mAP@all	Prec@200
Baseline (ViT-L [CLS] only)	0.572	0.612	0.612	0.699	0.591	0.646	0.227	0.289
Baseline + PCMI (No MLFA)	0.634	0.680	0.682	0.748	0.656	0.695	0.284	0.325
Baseline + MLFA (Fixed Prompt)	0.694	0.748	0.750	0.821	0.716	0.745	0.314	0.355
Baseline + MLFA + PCMI (OURS)	**0.775**	**0.813**	**0.855**	**0.896**	**0.799**	**0.820**	**0.338**	**0.406**

Table 3: Micro-Ablation on Aggregation Design. "w/o LayerNorm" removes the LayerNorm module, and "w/o Classification" removes the classification loss.

Methods	TU-Berlin Ext.		Sketchy Ext. Split1		Sketchy Ext. Split2		QuickDraw Ext.	
	mAP@all	Prec@100	mAP@all	Prec@100	mAP@200	Prec@200	mAP@all	Prec@200
w/o LayerNorm	0.527	0.563	0.587	0.669	0.547	0.596	0.207	0.259
w/o Classification	0.744	0.795	0.802	0.859	0.765	0.806	0.324	0.365
OURS	**0.775**	**0.813**	**0.855**	**0.896**	**0.799**	**0.820**	**0.338**	**0.406**

results confirm that hierarchical geometric information plays a fundamental role in bridging the abstraction gap. Introducing PCMI to the baseline also improves performance, showing that dynamic cross-modal interaction is more effective than fixed prompts. However, without multi-layer visual features, PCMI operates on a restricted representation, limiting its impact. Combining MLFA and PCMI achieves the best performance. MLFA enriches visual representations with multi-scale structure, while PCMI enhances cross-modal alignment. Their collaboration enables effective coarse-to-fine matching.

2) Micro-Ablation on Aggregation Design: Removing LayerNorm leads to a sharp performance drop, even below the baseline. This indicates that shallow and deep ViT layers follow different feature distributions, and normalization is essential for stable multi-scale fusion. Removing the classification loss causes a moderate decrease, suggesting that it mainly regularizes semantic topology during prompt optimization. The results show that structural aggregation is the primary driver of improvement, while auxiliary supervision further refines semantic consistency.

Table 4: This table reports the sensitivity analysis results of geometric (shallow) and semantic (deep) layer selection in MLFA. Columns 2–5 show the performance of different geometric layer ranges when the semantic layers are fixed to 22–24, and Columns 6–9 present the performance of different semantic layer ranges when the geometric layers are fixed to 1–3. Experiments are conducted on TU-Berlin Ext. (evaluated by mAP@all) and Sketchy Ext. Split2 (evaluated by mAP@200), with bold values indicating the optimal performance on each dataset.

Datasets	Geo Layers				Sem Layers			
	1–3	4–6	7–9	10–12	19–21	20–22	21–23	22–24
TU-Berlin Ext. (mAP@all)	**0.775**	0.764	0.752	0.731	0.750	0.763	0.770	**0.775**
Sketchy Ext. Split2 (mAP@200)	**0.799**	0.782	0.765	0.753	0.769	0.773	0.785	**0.799**

4.5 Layer Selection Sensitivity Analysis

To justify the selection of $\mathcal{I}_{geo}$ and $\mathcal{I}_{sem}$, we conduct layer sensitivity analysis on two representative datasets (TU-Berlin Ext. and Sketchy Ext. Split2). For clarity, we only report mAP and focus on the performance trend rather than exhaustive combinations.

As shown in Table 4, when fixing $\mathcal{I}_{sem}$ = 22-24 and shifting the geometric layer indices from early layers (1–3) to intermediate layers (4-6, 7-9, 10-12), mAP consistently decreases on both datasets. This trend indicates that the earliest transformer blocks preserve fine-grained structural primitives that are crucial for sketch-photo alignment. As the selected layers move deeper into the network, geometric information becomes increasingly abstracted, leading to reduced compatibility with sparse sketch representations. When evaluating different semantic layer choices while fixing $\mathcal{I}_{geo}$ = 1–3. We observe a gradual performance improvement as the selected layers approach the top of the network, with the best results achieved at layers 22–24. This confirms that the final transformer blocks provide the most stable and discriminative semantic representations for zero-shot retrieval.

4.6 Visualization Analysis of the Results

1) Visualization of Feature Embedding: To intuitively show the cross-modal features learned by the model, we use t-SNE to visualize the feature distribution of 8 categories. Figure 2 shows that different categories form clear clusters, demonstrating that our model effectively captures category-specific features. Due to high abstraction, the clustering compactness of QuickDraw Ext. is relatively poor, which provides guidance for future improvements.

2) Visualization of Retrievals: To intuitively showcase the retrieval performance of the proposed method, we present the top 7 real-image retrieval results for sketches across the TU-Berlin Ext. and Sketchy Ext. Split2 datasets.

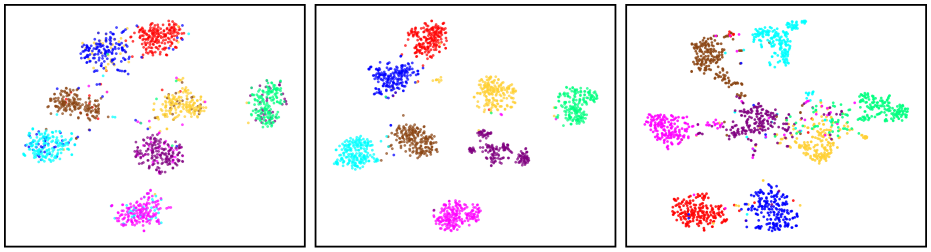

Fig. 2: T-SNE visualization of embeddings for sketches and photos was conducted on three dataset. For visualization, we randomly sampled examples from 8 test categories. Different colors represent different categories. (a) shows the visualization results for the TU-Berlin Ext. dataset, (b) presents the results for the Sketchy Ext. Split2, dataset, and (c) displays the results for the QuickDraw Ext. dataset.

As illustrated in Figure 3, each row starts with a sketch, followed by the retrieved real images. For the "Windmill" query, the top retrieved images share not only the semantic category but also the specific orientation of the blades and the viewpoint of the tower. This fine-grained alignment capability is a direct result of the Phased Cross-Modal Interaction, where the text context dynamically attends to visual details during the retrieval process.

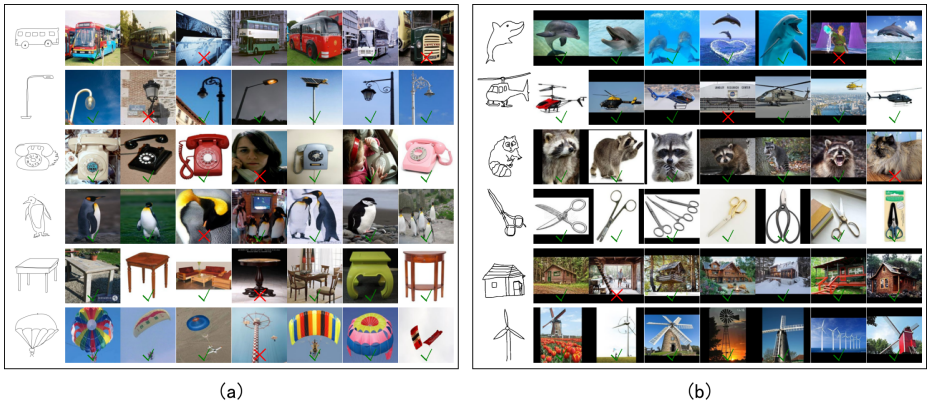

Fig. 3: The figure shows the top 7 real images retrieved by sketch-based search on the TU-Berlin Ext. dataset (a) and Sketchy Ext. Split2 dataset (b), with all samples from unseen categories. We use ticks and crosses to mark correct and incorrect retrieval results, respectively, and distinguishing between similar objects during retrieval remains a challenge for our model.

5 Conclusion

In this paper, we establish a theoretically grounded baseline for ZS-SBIR that moves beyond the naive parameter scaling of large foundation models. Recognizing that massive Vision-Language Models suffer from severe semantic over-compression—which violently discards the geometric primitives essential for abstract sketch matching—we propose the MLFA strategy. MLFA serves as an explicit anti-compression mechanism to rescue and decouple low-level structural states from deep semantic objects. To align these signals, we introduce the PCMI module, a dynamic compositional engine that utilizes learnable prompts to actively query and bind these decoupled structural states with global semantic entities for precise coarse-to-fine cross-modal alignment. Extensive experiments demonstrate that our framework achieves state-of-the-art results on several benchmarks and remains competitive on QuickDraw Ext. Ultimately, our findings highlight a fundamental insight: to fully unlock the zero-shot capabilities of vision-language models in sparse and abstract domains, explicitly counteracting semantic over-compression through compositional decoupling is paramount.

Acknowledgements

This research was funded by the National Natural Science Foundation of China under Grant 62376207, in part by the Open Project of the State Key Laboratory of Ocean Sensing (No. OSKF-2025Z06), in part by the Xidian University Specially Funded Project for Interdisciplinary Exploration, China under Grant TZJH2024045, in part by the Innovation Fund of Xidian University (No. YJSJ26022), and in part by the Fundamental Research Funds for the Central Universities, China.

References

1. Dey, S., Riba, P., Dutta, A., Llados, J., Song, Y.Z.: Doodle to search: Practical zero-shot sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2179–2188 (2019)
2. Du, J., Liu, Y., Gao, X., Han, J., Zhang, L.: Zero-shot sketch-based image retrieval with teacher-guided and student-centered cross-modal bidirectional knowledge distillation. Pattern Recognition **164**, 111529 (2025)
3. Dutta, A., Akata, Z.: Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In: CVPR. pp. 5089–5098 (2019)
4. Dutta, T., Singh, A., Biswas, S.: Styleguide: zero-shot sketch-based image retrieval using style-guided image generation. IEEE Transactions on Multimedia **23**, 2833–2842 (2020)
5. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? ACM Transactions on graphics (TOG) **31**(4), 1–10 (2012)
6. Ge, C., Wang, J., Qi, Q., Sun, H., Xu, T., Liao, J.: Semi-transductive learning for generalized zero-shot sketch-based image retrieval. In: AAAI. vol. 37, pp. 7678–7686 (2023)

7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
8. Ha, D., Eck, D.: A neural representation of sketch drawings. In: International Conference on Learning Representations (ICLR) (2018)
9. Jiang, H., Li, L.F., Yang, X., Wang, X., Luo, M.X.: Bsnet: a boundary-aware medical image segmentation network. The European Physical Journal Plus **139**(4), 1–14 (2024). https://doi.org/10.1140/epjp/s13360-024-05960-z
10. Kim, T., Cha, M., Kim, H.J., Lee, J.K., Kim, J.: Learning what and where to draw. CoRR **abs/1710.09698** (2017), http://arxiv.org/abs/1710.09698
11. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: CVPR. pp. 3174–3183 (2017)
12. Li, C.X., Zhang, D., Hu, Z., Wu, X.J.: Modality fused class-proxy with knowledge distillation for zero-shot sketch-based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology **35**(6), 6158–6169 (2025). https://doi.org/10.1109/TCSVT.2025.3530248
13. Li, Q., Wang, S., Zhang, W., Bai, S., Nie, W., Liu, A.: Dcdl: Dual causal disentangled learning for zero-shot sketch-based image retrieval. IEEE Transactions on Multimedia **27**, 5575–5590 (2025). https://doi.org/10.1109/TMM.2025.3543035
14. Li, S., Wang, L., Wang, S., Kong, D., Yin, B.: Hierarchical coupled discriminative dictionary learning for zero-shot learning. IEEE Transactions on Circuits and Systems for Video Technology (2023)
15. Li, X., Yu, J., Jiang, S., Lu, H., Li, Z.: Msvit: Training multiscale vision transformers for image retrieval. Trans. Multi. **26**, 2809–2823 (Jan 2024). https://doi.org/10.1109/TMM.2023.3304021
16. Lin, F., Li, M., Li, D., Hospedales, T., Song, Y.Z., Qi, Y.: Zero-shot everything sketch-based image retrieval, and in explainable style. In: CVPR. pp. 23349–23358 (2023)
17. Liu, L., Shen, F., Shen, Y., Liu, X., Shao, L.: Deep sketch hashing: Fast free-hand sketch-based image retrieval. In: CVPR. pp. 2862–2871 (2017)
18. Liu, Y., Dang, Y., Gao, X., Han, J., Shao, L.: Zero-shot sketch-based image retrieval via adaptive relation-aware metric learning. Pattern Recognition **152**, 110452 (2024)
19. Lyou, E., Lee, D., Kim, J., Lee, J.: Modality-aware representation learning for zero-shot sketch-based image retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5646–5655 (January 2024)
20. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. Journal of Machine Learning Research **22**(219), 1–28 (2021)
22. Sain, A., Bhunia, A.K., Chowdhury, P.N., Koley, S., Xiang, T., Song, Y.Z.: Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. arXiv preprint arXiv:2303.13440 (2023)
23. Sain, A., Bhunia, A.K., Potlapalli, V., Chowdhury, P.N., Xiang, T., Song, Y.Z.: Sketch3t: Test-time training for zero-shot sbir. In: CVPR. pp. 7462–7471 (2022)
24. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016)

25. Shen, Y., Liu, L., Shen, F., Shao, L.: Zero-shot sketch-image hashing. In: CVPR. pp. 3598–3607 (2018)
26. Singha, M., Jha, A., Gupta, D., Singla, P., Banerjee, B.: Elevating all zero-shot sketch-based image retrieval through multimodal prompt learning (2024), https://arxiv.org/abs/2407.04207
27. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in neural information processing systems. vol. 28 (2015)
28. Su, H., Song, G., Huang, K., Wang, J., Yang, M.: Cross-modal attention alignment network with auxiliary text description for zero-shot sketch-based image retrieval. In: Artificial Neural Networks and Machine Learning – ICANN 2024. pp. 52–65 (2024)
29. Su, H., Song, G., Wang, J., Zhu, Y.: Dynamic multi-level weighted alignment network for zero-shot sketch-based image retrieval (2025), https://arxiv.org/abs/2511.00925
30. Tian, J., Xu, X., Shen, F., Yang, Y., Shen, H.T.: Tvt: Three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval. In: AAAI. vol. 36, pp. 2370–2378 (2022)
31. Tian, J., Xu, X., Wang, Z., Shen, F., Liu, X.: Relationship-preserving knowledge distillation for zero-shot sketch based image retrieval. In: ACM MM. pp. 5473–5481 (2021)
32. Tursun, O., Denman, S., Sridharan, S., Goan, E., Fookes, C.: An efficient framework for zero-shot sketch-based image retrieval. Pattern Recognition **126**, 108528 (2022)
33. Wang, K., Wang, Y., Xu, X., Liu, X., Ou, W., Lu, H.: Prototype-based selective knowledge distillation for zero-shot sketch based image retrieval. In: ACM MM. pp. 601–609 (2022)
34. Wang, W., Shi, Y., Chen, S., Peng, Q., Zheng, F., You, X.: Norm-guided adaptive visual embedding for zero-shot sketch-based image retrieval. In: IJCAI. pp. 1106–1112 (2021)
35. Wang, X., Peng, D., Hu, P., Gong, Y., Chen, Y.: Cross-domain alignment for zero-shot sketch-based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology (2023)
36. Wang, Z., Wang, H., Yan, J., Wu, A., Deng, C.: Domain-smoothing network for zero-shot sketch-based image retrieval. arXiv preprint arXiv:2106.11841 (2021)
37. Yang, F., Ismail, N.A., Pang, Y.Y., Kebande, V.R., Al-Dhaqm, A., Koh, T.W.: A systematic literature review of deep learning approaches for sketch-based image retrieval: Datasets, metrics, and future directions. IEEE Access **12**, 14847–14869 (2024)
38. Yelamarthi, S.K., Reddy, S.K., Mishra, A., Mittal, A.: A zero-shot framework for sketch based image retrieval. In: ECCV. pp. 300–317 (2018)
39. Yin, Z., Yan, J., Xu, C., Deng, C.: Asymmetric mutual alignment for unsupervised zero-shot sketch-based image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 16504–16512 (2024)
40. Zhang, H., Cheng, D., Kou, Q., Asad, M., Jiang, H.: Indicative vision transformer for end-to-end zero-shot sketch-based image retrieval. Advanced Engineering Informatics **60**, 102398 (2024)
41. Zhang, H., Jiang, H., Wang, Z., Cheng, D.: Ontology-aware network for zero-shot sketch-based image retrieval. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023)

42. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (Jul 2022). https://doi.org/10.1007/s11263-022-01653-1, http://dx.doi.org/10.1007/s11263-022-01653-1
43. Zhu, J., Xu, X., Shen, F., Lee, R.K.W., Wang, Z., Shen, H.T.: Ocean: A dual learning approach for generalized zero-shot sketch-based image retrieval. In: ICME. pp. 1–6. IEEE (2020)