Modality-Aware Representation Learning for Zero-shot Sketch-based Image Retrieval

Eunyi Lyou[1] Doyeon Lee[1] Jooeun Kim[1] Joonseok Lee[1,2*]
[1]Seoul National University [1,2]Google Research
{onlyou0416, kje980714, joonseok}@snu.ac.kr, omocomo83@gmail.com

Abstract

Zero-shot learning offers an efficient solution for a machine learning model to treat unseen categories, avoiding exhaustive data collection. Zero-shot Sketch-based Image Retrieval (ZS-SBIR) simulates real-world scenarios where it is hard and costly to collect paired sketch-photo samples. We propose a novel framework that indirectly aligns sketches and photos by contrasting them through texts, removing the necessity of access to sketch-photo pairs. With an explicit modality encoding learned from data, our approach disentangles modality-agnostic semantics from modality-specific information, bridging the modality gap and enabling effective cross-modal content retrieval within a joint latent space. From comprehensive experiments, we verify the efficacy of the proposed model on ZS-SBIR, and it can be also applied to generalized and fine-grained settings.

1. Introduction

Sketch-based image retrieval (SBIR) is a cross-view retrieval task, retrieving relevant natural images given an abstract and ambiguous sketch of an object. Typically, a machine learning model is trained to map a photo and its corresponding sketch closely in a common latent space, trained on a set of paired samples.

However, it is usually challenging and often infeasible to acquire paired samples of free-hand sketches and natural photos across all potential objects. For this reason, SBIR task has naturally adopted zero-shot learning, which has emerged to classify samples belonging to categories never seen at training. In Zero-Shot Sketch-based Image Retrieval (ZS-SBIR) [31], a model is trained on some collection of paired sketch-photo samples, and at testing time, it is asked to retrieve photos with an object unknown at training given its sketch.

Under this zero-shot setting, prior works have studied semantic transfer by utilizing external semantic knowledge, such as label embeddings or hierarchical graphs [19, 7, 45, 25, 2, 3, 9]. They provide insights for strategic handling of zero-shot scenarios in the presence of external supplementary information. Recently, CLIP-AT [26] proposes employing comprehensive visual and textual embeddings from CLIP [24], trained on large-scale vision-language data. Transferring the rich visual-linguistic relationship learned from the large data, this approach significantly improves the retrieval performance.

However, adapting off-the-shelf multimodal models such as CLIP [24] to the ZS-SBIR task is non-trivial. Modality gap [14, 32] is a known phenomenon that image and text embeddings in a common space still tend to be clearly separated, even though the model is trained to embed images and texts based on their semantics, not based on their modalities. Ideally, we would like to learn to map instances based on both their semantics and forms (modalities), where these two are disentangled. Then, a simple nearest neighbor search within the target modality space would retrieve relevant items regardless of their forms.

To bridge the modality gap and fully utilize a multimodal foundation model, we propose Modality-Aware encoders for Sketch-based Image Retrieval (MA-SBIR), which explicitly and separately learn both modality-agnostic semantics and modality-specific information. At a high level, our model is trained to transform a modality space into another within a shared latent space, by explicitly learning modality-specific nuance, separated from semantics.

A key advantage of this design is that the model can be trained without paired sketch-photo examples; it can be trained on a set of sketches and another set of photos labeled with a common vocabulary, without necessarily having one-to-one relationships. Instead of directly learning to locate photos and sketches, we adopt an indirect approach through the category annotation (with their names as text modality). Our model consumes an image (either a sketch or a photo) and its associated text at a time, and their semantic embeddings are aligned as previous models like CLIP. At the same time, the model learns to distinguish sketches and photos with a separate modality encoding. In this way, the

*Corresponding author

model learns to represent semantics and modality, image by image, without requiring sketch-photo pairs.

By design, we believe our indirect alignment is more effective than the conventional direct alignment for the categorical SBIR, where a photo is considered correct if it is in the same category with the queried sketch (*e.g.*, correct if both belong to clothes, regardless of their specific type or color). In addition to this, we apply our proposed approach on two other relevant settings as well. First, we test on the Generalized Zero-shot SBIR (GZS-SBIR) [45, 3, 9, 25, 15], which is a more realistic setting that the test set contains both seen and unseen classes. This is to better simulate a real situation where performance on both is vital due to greater prevalence of seen classes. Second, we test on a fine-grained (or instance-wise) SBIR setting (FG-SBIR), where only the one-to-one mapped photo is considered correct for each query sketch. As our model is not designed to directly learn sketch-photo alignment, we expect sub-optimal performance on this setting. For this reason, we adapt our model with a few modifications (Sec. 3.3).

We summarize our contributions as follows:

1. We propose a novel method to align a joint embedding space, disentangling semantics from modality-specific information.
2. Our proposed model indirectly aligns sketches and photos, removing the necessity of paired sketch-photo examples for training.
3. We verify that our proposed method achieves the state-of-the-art performance on diverse zero-shot sketch-based image retrieval tasks.

2. Related Work

Sketch-based Image Retrieval (SBIR) can be categorized into two based on retrieval granularity: Category-level and Instance-level (fine-grained SBIR). Category-level SBIR is to retrieve photos of an object in the same category among the candidate images covering multiple categories, based on a given sketch. For instance, given an image of a cat, category-level SBIR considers images of any kind of cats correct. In instance-level SBIR, on the other hand, a sketch query should yield an exactly matching instance from images, not just those belonging to the same category. In the above example, the images with the same cat instance are considered correctly retrieved.

Category-level SBIR. Recent SBIR works have focused on aligning the shared embedding space of sketches and images by employing triplet ranking methods [42, 16, 4, 13], re-ranking scenarios [11, 39], or efficient hashing optimization [17, 44, 20] for this. Furthermore, some approaches [44, 10] align the distributions of images and sketches to train a generative model, forcing sketch and image representations to preserve their shared semantics.

The zero-shot framework, pioneered by [31], is designed to facilitate the retrieval of categories that are unseen at training. ZS-SBIR often utilizes auxiliary information as a means of guiding previously unseen images into a common semantic space, *e.g.*, from predefined class labels [19], hierarchical graphs [7, 45, 25, 2, 3], and textual embeddings from a vision-language joint common space [26]. Adversarial training [45, 7, 2] also has been employed. [3] aligns features from the intermediate and concluding layers of dual backbones, and [12] employs cross-domain mix-up strategies. In order to disentangle domain and semantic features, gradient reversal layers [5] and separate modeling for each encoding [35]. Modern research has tackled catastrophic forgetting at fine-tuning to preserve the accumulated knowledge from the pre-training via knowledge distillation [19, 38, 36], backbone sharing [37], prompt learning [26], and test-time adaptation [28].

However, zero-shot models often overfits towards unknown categories. To address this, a more realistic Generalized Zero-Shot Sketch-Based Image Retrieval (GZS-SBIR) [7] setting is proposed, mixing seen and unseen classes in the test set. A few recent works [45, 3, 9, 25, 15] have explored this direction.

Instance-level SBIR. Unlike the categorical SBIR, fine-grained (FG) SBIR aims to retrieve the exact target item at the instance level given a sketch image. FG-SBIR models focus on enhancing pixel representations by incorporating spatial modules to account for detailed spatial positions [33], cross-interaction modules to calculate patch-level similarities [34, 15], and utilizing randomized patch shuffling techniques [22, 26].

Aligned with the zero-shot setting in SBIR, fine-grained task is also tackled in zero-shot settings [21]. Recent studies address cross-category generalization by leveraging the well-aligned CLIP [24] space [26], introducing cross-modal attention and patch-level matching [15], and knowledge distillation from additional unlabelled photos [27].

3. Method

3.1. Problem Formulation

Let $\mathcal{X}$ be a paired set of n samples of a sketch and its corresponding real photo, where each pair is annotated with a class. The i-th sample is denoted by $(\mathbf{S}^{(i)}, \mathbf{P}^{(i)}, c^{(i)}) \in \mathcal{X}$, where $\mathbf{S}^{(i)} \in \mathbb{R}^{M \times N \times 3}$ is a sketch image of size $M \times N$, $\mathbf{P}^{(i)} \in \mathbb{R}^{M' \times N' \times 3}$ is its corresponding ground truth photo of size $M' \times N'$. ($\mathbf{S}$ and $\mathbf{P}$ may have different size.) $c^{(i)} \in \mathcal{C}$ is the ground truth category of the pair, where $\mathcal{C}$ is a set of all categories. Each category $c \in \mathcal{C}$ is transformed into a textual caption, formatted as 'a photo of a c', denoted by $T \in |\mathcal{V}|^L$, where $\mathcal{V}$ is the vocabulary set and L is the maximum length of the sentence.

On zero-shot tasks, $\mathcal{C}$ is further split to the seen classes

($\mathcal{C}_\text{s}$) and unseen classes ($\mathcal{C}_\text{u}$), used for training and testing, respectively, where $\mathcal{C}_\text{s} \cap \mathcal{C}_\text{u} = \varnothing$ and $\mathcal{C} = \mathcal{C}_\text{s} \cup \mathcal{C}_\text{u}$. The sketches and photos are split into train and test sets according to their labels; that is, $\mathcal{X}_\text{train} = \{(\mathbf{S}^{(i)}, \mathbf{P}^{(i)}, T^{(i)}) \,|\, c^{(i)} \in \mathcal{C}_\text{s}\}$, $\mathcal{X}_\text{test} = \{(\mathbf{S}^{(i)}, \mathbf{P}^{(i)}) \,|\, c^{(i)} \in \mathcal{C}_\text{u}\}$.

In Sketch-based Image Retrieval (SBIR) task, a query sketch image $\mathbf{S}$ is given, and the model aims to retrieve relevant images from the candidate photos, either within the same category (category-level) or the exactly matched image instance (instance-level or fine-grained). For category-level SBIR, the photos are considered relevant with any sketch within the same category; that is, $\mathcal{X}_c = \{(\mathbf{S}^{(i)}, \mathbf{P}^{(j)}) \mid c^{(i)} = c^{(j)} = c\}$ is a set of correct pairs for a class $c \in \mathcal{C}$. In fine-graind SBIR, there always exists exactly one matched photo for each sketch; that is, $\mathcal{X}_i = \{(\mathbf{S}^{(i)}, \mathbf{P}^{(i)})\}$ for a sample datum i. For Generalized ZS-SBIR, randomly sub-sampled examples from $\mathcal{X}_\text{train}$ are added to $\mathcal{X}_\text{test}$, where $\mathcal{X}_\text{train}$ remains unchanged.

3.2. The Proposed Model

Overview. The overall workflow of our approach for the category-level SBIR is illustrated in Fig. 1. The first step comprises of modality-specific encoding of an input image, either a sketch $\mathbf{S}$ or a photo $\mathbf{P}$ (using the image encoder $\mathbf{E}_\text{img}$) and caption T (using the text encoder $\mathbf{E}_\text{txt}$), and further alignment within a batch using CLIP loss, $\mathcal{L}_\text{clip}$. In the second step, a learnable modality embedding vector is subtracted from the CLIP image ($\mathbf{z}_\text{img}$) and text ($\mathbf{z}_\text{txt}$) embeddings, to leave only their semantics, denoted by $\mathbf{s}_\text{img}$ and $\mathbf{s}_\text{txt}$, respectively. We apply another semantic alignment loss, $\mathcal{L}_\text{sem}$, between them. Lastly, the modality encoding is added back to the opposite modality, producing converted text ($\mathbf{z}'_\text{txt}$) and image ($\mathbf{z}'_\text{img}$) embeddings. They are trained to reconstruct the original embedding ($\mathcal{L}_\text{rec}$) and to classify the target modality correctly ($\mathcal{L}_\text{mc}$). This assists the model to disentangle common semantics from modality-specific information, eventually helping to transform one domain (*e.g.*, sketch) to another (*e.g.*, photo). Once trained, the original and transformed embeddings are within the same embedding space, applicable for retrieval tasks.

Inputs. The network receives a single image, either a sketch $\mathbf{S}$ or a photo $\mathbf{P}$, with a caption (T) as input. Adopting the Vision Transformer [6], the input image is divided into multiple $P \times P$ patches $\in \mathbb{R}^{n_\text{patch} \times (P^2 \cdot 3)}$, where n_patch is the resulting number of patches. Each image patch is then linearly projected to a d_img-dimensional space. Similarly, the caption is tokenized and vectorized, resulting in a sequence with n_seq token embeddings of size d_txt. An additional `[CLS]` token embedding is prepended to each sequence. We denote the input image sequence by $\mathbf{x} \in \mathbb{R}^{(n_\text{patch}+1) \times d_\text{img}}$ and the text sequence by $\mathbf{y} \in \mathbb{R}^{(n_\text{seq}+1) \times d_\text{txt}}$.

Indirect Alignment of Sketches and Photos. We use pretrained CLIP [24] image ($\mathbf{E}_\text{img}$) and text encoders ($\mathbf{E}_\text{txt}$), taking $\mathbf{x}$ and $\mathbf{y}$ as inputs, respectively. From the output sequences, $\mathbf{E}_\text{img}(\mathbf{x}) \in \mathbb{R}^{(n_\text{patch}+1) \times d_\text{img}}$ and $\mathbf{E}_\text{txt}(\mathbf{y}) \in \mathbb{R}^{(n_\text{seq}+1) \times d_\text{txt}}$, we take the representations corresponding to `[CLS]`, and linearly map them to a common embedding size d. We denote the resulting embeddings by $\mathbf{z}_\text{img} \in \mathbb{R}^d$ and $\mathbf{z}_\text{txt} \in \mathbb{R}^d$, respectively, and they are our base image and text representations.

Given $\mathbf{z}_\text{img}$ and $\mathbf{z}_\text{txt}$, we further align the embeddings by minimizing the following loss:

$$\mathcal{L}_\text{clip} = \frac{1}{2}\left(\text{CLIP}(\mathbf{z}_\text{img}, \mathbf{z}_\text{txt}) + \text{CLIP}(\mathbf{z}_\text{txt}, \mathbf{z}_\text{img})\right), \quad (1)$$

$$\text{where } \text{CLIP}(\mathbf{a}, \mathbf{b}) = -\log \frac{\exp\left(\mathbf{a} \cdot \mathbf{b}/\tau\right)}{\sum_{j=1}^{B} \exp\left(\mathbf{a} \cdot \mathbf{b}_j/\tau\right)}. \quad (2)$$

Our approach is distinguished from existing methods in that ours does not have a direct mechanism to align sketches and photos within the image modality, while the subtle differences between them are indirectly aligned through texts. This approach is opposed to the direct alignment via triplet learning in existing works [42, 16, 4, 13]. The biggest advantage of our approach over the previous direct triplet learning is that ours does not require paired training examples of sketches and photos, which are costly to collect. In order to directly train the model to distinguish the two, previous models have relied on positive pairs of a sketch and a photo. With our indirect approach, however, no explicit positive relationship between a sketch and a photo is utilized. Instead, both sketches and photos are embedded via a common image encoder and aligned with the associated text, taking advantage of the intricate representation capacity of the CLIP. This capability of leveraging unpaired datasets addresses the well-known issue of data scarcity in the community, aligning with the efforts of [1, 27] to tackle limited data availability. Our design also simplifies the contrastive loss term, removing the need for negative sampling.

Modality Encoding. Since our model uses a common visual encoder for sketches and photos, we need an additional mechanism to distinguish them. For this, we introduce modality encoder $\mathbf{E}_\text{mod}$, composed of learnable encoding ($\mathbf{m} \in \mathbb{R}^d$) for each modality. Specifically, we assign a unique index to each specific modality; for example, sketches, photos, and texts are assigned with 0, 1, and 2, respectively. Under our design, the model *learns* to represent the modality itself, regardless of its nature, in its modality encoding $\mathbf{m}$.

Under this setting, our model is trained to separate its latent space into its constituent semantic and modality components. Once this is achieved, one can convert an embedding from one domain to another simply by

$$\mathbf{x}_\text{trg} = \mathbf{x}_\text{src} - \mathbf{m}_\text{src} + \mathbf{m}_\text{trg}, \quad (3)$$

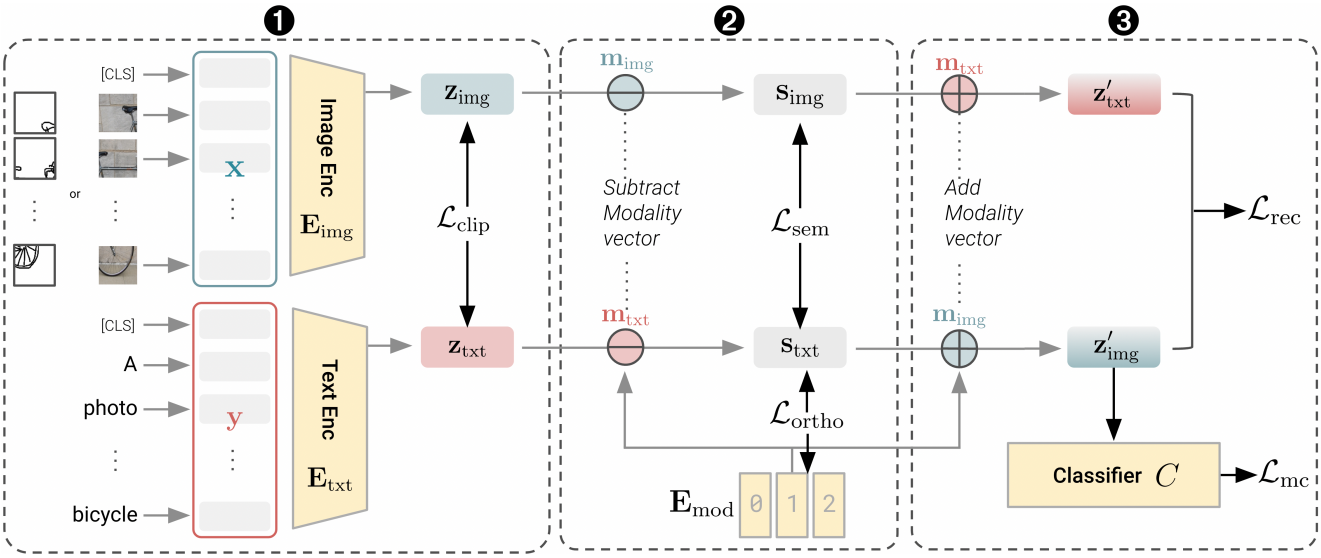

Figure 1. **Overview of our Architecture.** Given an image $\mathbf{x}$ (either a sketch or a photo) and a text $\mathbf{y}$, modality-specific encoders ($\mathbf{E}_{\text{img}}$ and $\mathbf{E}_{\text{txt}}$) embed them to $\mathbf{z}_{\text{img}}$ and $\mathbf{z}_{\text{txt}}$, respectively, where each $\mathbf{z}$ is a sum of semantic embedding $\mathbf{s}$ and modality encoding $\mathbf{m}$. We acquire semantic-only vectors ($\mathbf{s}_{\text{img}}$, $\mathbf{s}_{\text{txt}}$) from them by subtracting the modality encoding ($\mathbf{m}_{\text{img}}$, $\mathbf{m}_{\text{txt}}$). By adding the opposite modality encoding to the semantic embeddings, we reconstruct the image and text embeddings ($\mathbf{z}'_{\text{img}}$ and $\mathbf{z}'_{\text{txt}}$).

where $\mathbf{x} \in \mathbb{R}^d$ are the learned embedding (containing both semantics and modality), $\mathbf{m} \in \mathbb{R}^d$ are the modality encodings corresponding to each dataset, either source (src) and target (trg). We expect the converted representation in this way to yield improved performance on the retrieval tasks.

Visual-Text Alignment. From the CLIP embeddings, $\mathbf{z}_{\text{img}} \in \mathbb{R}^d$ and $\mathbf{z}_{\text{txt}} \in \mathbb{R}^d$, we subtract their modality encodings, $\mathbf{m}_{\text{img}} \in \mathbb{R}^d$ and $\mathbf{m}_{\text{txt}} \in \mathbb{R}^d$, respectively. We get image and text embeddings purely based on their semantics:

$$\mathbf{s}_{\text{img}} = \mathrm{N}(\mathbf{z}_{\text{img}} - \mathbf{m}_{\text{img}}),\ \mathbf{s}_{\text{txt}} = \mathrm{N}(\mathbf{z}_{\text{txt}} - \mathbf{m}_{\text{txt}}), \quad (4)$$

where $\mathrm{N}(x) = \mathbf{x}/\|\mathbf{x}\|_2$, indicating the normalization operator. For a positive paired example of an image and its associated text, we train the model to keep their semantic representations similar. Specifically, we consider the normalized average of image and text embeddings as its general semantic representations, denoted by $\mathbf{z}$, and both semantic representations are encouraged to be close to it. Formally, we minimize the following semantic loss $\mathcal{L}_{\text{sem}}$:

$$\mathcal{L}_{\text{sem}} = -\cos\left(\mathbf{s}_{\text{img}}, \mathbf{z}\right) - \cos\left(\mathbf{s}_{\text{txt}}, \mathbf{z}\right), \quad (5)$$

where $\mathbf{z} = \mathrm{N}((\mathbf{z}_{\text{img}} + \mathbf{z}_{\text{txt}})/2)$. We use cosine similarity, but other similarity or distance functions may be applicable.

Cross-modal Reconstruction. On the purely semantic representations, $\mathbf{s}_{\text{img}}$ and $\mathbf{s}_{\text{txt}}$, we add the modality encoding from the opposite modality, producing reconstructed text and image embeddings, denoted by $\mathbf{z}'_{\text{txt}}$ and $\mathbf{z}'_{\text{img}}$:

$$\mathbf{z}'_{\text{txt}} = \mathrm{N}(\mathbf{s}_{\text{img}} + \mathbf{m}_{\text{txt}}),\ \mathbf{z}'_{\text{img}} = \mathrm{N}(\mathbf{s}_{\text{txt}} + \mathbf{m}_{\text{img}}). \quad (6)$$

We apply the following loss to ensure the reconstructed embedding maintain similar directions to the original:

$$\mathcal{L}_{\text{rec}} = -\cos\left(\mathbf{z}'_{\text{txt}}, \mathbf{z}_{\text{txt}}\right) - \cos\left(\mathbf{z}'_{\text{img}}, \mathbf{z}_{\text{img}}\right). \quad (7)$$

In addition, to reconstruct more precisely, we introduce a modality classifier on the reconstructed embeddings, $\mathbf{z}'_{\text{img}}$ and $\mathbf{z}'_{\text{txt}}$. For each, we minimize a CE loss $\mathcal{L}_{\text{mc}}$ defined as

$$\mathcal{L}_{\text{mc}} = \sum_{j=1}^{C} m_j \log(\hat{m}_j), \quad (8)$$

where $\hat{m}_j \in \mathbb{R}^C$ is the predicted logits for the input embedding to belong to each modality, m_j is the one-hot encoding of its ground-truth modality, and C is the number of modalities. By minimizing the classification loss, we ensure that the reconstructed $\mathbf{z}'_{\{\text{txt,img}\}}$ originate from separate classes recognizable by the classifier, irrespective of their initial modality. This approach helps emphasize the distinctions between different modalities.

Orthogonal Regularization. While the previously introduced losses ensure the disentanglement of semantic and modality information, this can be enhanced by imposing orthogonality between two directions. Specifically, we design the orthogonality regularizer $\mathcal{L}_{\text{ortho}}$ as follows:

$$\mathcal{L}_{\text{ortho}} = \frac{1}{C} \sum_{j=1}^{C} |\mathbf{z} \cdot \mathbf{m}_j|, \quad (9)$$

where $\mathbf{m}_j$ is the j'th modality embedding, and $\mathbf{z} = \mathrm{N}(\mathbf{z}_{\text{img}} + \mathbf{z}_{\text{txt}}/2)$. This leads to the alignment of the dot products between all vectors in semantic matrix E_s and modality matrix E_m towards zero, consequently enforcing orthogonality. While a dot product of zero indicates either perpendicularity between non-zero vectors or one of the vectors being zero, we preemptively prevent the latter through other previously mentioned loss terms. For instance, the

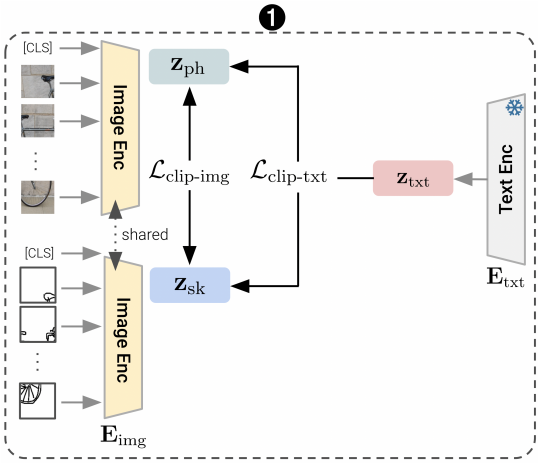

Figure 2. **Overview of FG-SBIR model architecture.** Unlike for categorical SBIR, fine-grained version takes paired sketches and photos as input to obtain latent vectors ($\mathbf{z}_{\{\text{ph,sk}\}}$), followed by step 2 and 3 described in Fig. 1.

uniqueness of modality vectors is ensured by minimizing $\mathcal{L}_{\text{mc}}$, and semantic vectors resist convergence to zero due to the preservation of uniformity and alignment in the latent space, via $\mathcal{L}_{\text{clip}}$, as observed in [40].

Overall Training Objective. Our model is trained to minimize the following:

$$\mathcal{L} = \mathcal{L}_{\text{clip}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{mc}}\mathcal{L}_{\text{mc}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}},$$

where all λs are hyperparameters to control relative importance between each loss. See Sec. 3.4 for more details.

3.3. Extension to Instance-level SBIR

Fine-grained SBIR (FG-SBIR) task aims to retrieve precisely matched photos from a sketch query. Unlike the category-level SBIR, merely retrieving images in the same category is insufficient. Intuitively, our proposed approach in Sec. 3.2 is not suitable for this task, as it mainly relies on the indirect connection between sketches and photos through coarse-grained textual descriptions, without direct enforcement based on paired examples. For this reason, our vanilla model is not suitable for finer-grained alignment.

However, we pose a question: can our model still perform well, if it is trained on the paired data? To answer this question, we adapt our model to the instance-level SBIR setting with some modifications guided by the inherent nature of the task. Specifically, we take a photo-sketch-text triplet as input, instead of an image-text pair, although we still use a common image encoder. It is inevitable in this fine-grained setting to take paired photo-sketch examples. Some loss functions are also modified, explained in detail subsequently. However, there is no major change to the model architecture, other than the straightforward extension to use three encoders depicted in Fig. 2.

Inputs and CLIP Loss. To learn more sophisticated semantic distribution in the sketch and photograph modalities, a training example is composed of a triplet $(\mathbf{S}, \mathbf{P}, T)$, where $\mathbf{S}$, $\mathbf{P}$, and T indicate a sketch, its corresponding photo, and their textual caption, respectively. Each of these three is considered a distinguished modality, assigned with its own modality index. Subsequently, we extract modality-specific features: $\mathbf{z}_{\text{sk}} = \mathbf{E}_{\text{img}}(\mathbf{S}) \in \mathbb{R}^d$, $\mathbf{z}_{\text{ph}} = \mathbf{E}_{\text{img}}(\mathbf{P}) \in \mathbb{R}^d$, and $\mathbf{z}_{\text{txt}} = \mathbf{E}_{\text{txt}}(T) \in \mathbb{R}^d$, respectively, where $\mathbf{E}_{\{\text{img,txt}\}}$ are pretrained modality-specific encoders.

The biggest change for fine-grained model is introduction of direct sketch-photo alignment. Specifically, we add another contrastive loss designed for sketch-photo alignment, defined as follows:

$$\mathcal{L}_{\text{clip-img}} = \frac{1}{2}\left(\text{CLIP}(\mathbf{z}_{\text{sk}}, \mathbf{z}_{\text{ph}}) + \text{CLIP}(\mathbf{z}_{\text{ph}}, \mathbf{z}_{\text{sk}})\right), \quad (10)$$

where $\text{CLIP}(\mathbf{a}, \mathbf{b})$ is defined in Eq. (2).We balance this direct alignment with the existing text-guided indirect alignment from Eq. (1) by convex combination:

$$\mathcal{L}_{\text{clip}} = \lambda_{\text{clip-img}}\mathcal{L}_{\text{clip-img}} + (1 - \lambda_{\text{clip-img}})\mathcal{L}_{\text{clip-txt}}, \quad (11)$$

where $\mathcal{L}_{\text{clip-txt}}$ is equivalent as Eq. (1) but we use different notation to ensure that notation $\mathcal{L}_{\text{clip}}$ continues to represent the overall contrastive loss. With a hyperparameter $\lambda_{\text{clip-img}}$, we adjust the relative influence from the two types of contrastive losses.

Additional Regularizations. Beyond comprehending the semantics of images, grasping the structural aspects, such as the position of an instance or edge detection within the image, is another crucial dimension in FG-SBIR. To effectively handle geometric attributes, we adopt patch shuffling, aligning with prior studies [34, 15, 22, 26] with similar objectives. Initially, non-overlapping patches from the sketch and the photo are randomly shuffled using a permutation function π, and then fed into the image encoder $\mathbf{E}_{\text{img}}$. We take the representation corresponding to the `[CLS]` token as the sketch and photo embedding, denoted by $\mathbf{z}_{\text{sk}}$ and $\mathbf{z}_{\text{ph}}$, respectively. We additionally minimize the following patch shuffling loss, allowing push-and-pull between these two:

$$\mathcal{L}_{\text{ps}} = \frac{1}{2}\left(\text{CLIP}(\mathbf{z}_{\text{sk}}, \mathbf{z}_{\text{ph}}) + \text{CLIP}(\mathbf{z}_{\text{ph}}, \mathbf{z}_{\text{sk}})\right). \quad (12)$$

Overall Training Objective. With λs as hyperparameters, the overall loss is given by

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{clip}} + \lambda_{\text{ps}}\mathcal{L}_{\text{ps}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} \\ + \lambda_{\text{mc}}\mathcal{L}_{\text{mc}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}}.\end{aligned} \quad (13)$$

3.4. Implementation Details

We set d to 512 and 768 for ZS-SBIR and FG-ZS-SBIR, respectively. We use pretrained CLIP [24] and ViT (ViT-B/32 for categorical and ViT-B/16 for fine-grained) as backbone. For FG-ZS-SBIR, $\mathbf{z}_{\text{img}} \in \mathbb{R}^{768}$ is extracted from the

$\mathbf{E}_{\text{img}}$ without the last projection layer. The text embedding is projected to the same size by a projection layer. Also, we strategically freeze the text encoder during training to enhance efficiency, supported by our findings that this maintains performance unaffected for FG-ZS-SBIR. Modality classifier is implemented with a FC layer.

We use AdamW optimizer with early stopping. The learning rate for the image/text encoder is set deliberately lower than the rest, to prevent catastrophic forgetting: (5e-8, 5e-3) for ZS-SBIR, and (5e-5, 1e-2) for FG-ZS-SBIR. We use batch size of 256 on Sketch-Ext and TU-Berlin-Ext for ZS-SBIR, while 64 on QuickDraw-Ext for ZS-SBIR and on Sketchy for FG-ZS-SBIR. We grid search the weight coefficients $\lambda_{\{\text{sem, mc, rec, ortho}\}}$ and set to (0.1, 0.1, 0.02, 0) for Sketchy-Ext, (0.1, 0.1, 0, 0) for TU-Berlin-Ext, and (0.2, 0.2, 0.2, 0.2) for QuickDraw-Ext. For FG-ZS-SBIR, we use (0.5, 0.1, 0.02, 0.1, 0.02) for $\lambda_{\{\text{ps, sem, mc, rec, ortho}\}}$. We report mean performance of three independent experiments. We implement our model with PyTorch [23] and experiment on a NVIDIA RTX A6000 GPU.

4. Experimental Settings

Datasets. We evaluate our method on well-known ZS-SBIR datasets: Sketchy Extended, TU-Berlin Extended and QuickDraw Extended. Sketchy Extended [18] consists of 73,002 images on 125 categories, on average 604 sketches and 584 images per class, extending Sketchy [29]. For zero-shot experiments, we set aside 21 classes that are not present in the 1,000 classes of ImageNet for testing, leaving the rest 104 classes for training, following [41].

TU-Berlin Extended [18] extends the TU-Berlin [8], originally designed for sketch classification and composed of 20,000 sketches on 250 object categories in a balanced manner, by incorporating 204,489 natural images from [43]. After this extension, each category has 787 images on average, but highly imbalanced. We follow the partition protocol in [31], where 30 classes are randomly selected for testing, leaving the rest 220 classes for training. Due to the significant imbalances in the numbers of real images in each class, [31] also made sure that each test category has at least 400 images when choosing the test set.

QuickDraw Extended [5] is a large-scale dataset designed for ZS-SBIR. Using Google Quick, Draw! data, it contains 110 categories with 330,000 sketches, including 3,000 amateur sketches per category. It also has 204,000 images taken from Flickr tagged with the corresponding label. For split, we follow [41, 5] to ensure the 30 test classes free from ImageNet, using the rest 80 classes for training.

Baselines. We compare our method with state-of-the-art methods for ZS-SBIR and FG-ZS-SBIR. For ZS-SBIR, we compare with CNN-based models (SEM-PCYC [7], SAKE [19], OCEAN [45], BDA [3], and Sketch-3T [28]) and ViT-based models (TVT [35], PSKD [38], SaA [25], ZSE [15], and CLIP-AT[26]). For generalized ZS-SBIR, we additional compare with STL[9].

Table 1. Comparison for Categorical ZS-SBIR

	Model	Sketchy-Ext		TU-Berlin-Ext		QuickDraw-Ext	
		m@200	P@200	m@all	P@100	m@all	P@200
CNN	SEM-PCYC [7]	-	-	29.7	42.6	-	-
	SAKE [19]	49.7	59.8	47.5	59.9	-	-
	OCEAN [45]	-	-	33.3	46.7	-	-
	BDA [3]	45.8	55.6	37.4	50.4	15.4	35.5
	Sketch-3T [28]	-	62.4	50.7	-	-	-
ViT	TVT [35]	53.1	61.8	48.4	66.2	14.9	29.3
	PSKD [38]	56.0	64.5	50.2	66.2	15.0	29.8
	SaA [25]	53.5	63.0	49.0	60.8	14.8	-
	ZSE[Ret] [15]	50.4	60.2	56.9	63.7	14.2	20.2
	ZSE[RN] [15]	52.5	62.4	54.2	65.7	14.5	21.6
	CLIP-AT* [26]	63.6	71.0	65.9	76.7	29.3	36.4
	Ours [clip]	68.5	74.9	70.5	77.6	32.2	41.9
	Ours [original]	68.5	74.9	**70.7**	77.6	31.7	41.6
	Ours [converted]	**69.1**	**75.5**	70.5	**77.7**	**32.7**	**42.5**

*Indicates our reproduction, using codes provided by [35].

Our Model Variants. We present three variations of our proposed approach. Initially, we train our model solely with $\mathcal{L}_{\text{clip}}$ to assess the impact of indirect alignment technique, labeled with **[clip]**. On our full model, we may use the original embeddings $\mathbf{z}_{\{\text{img, txt}\}}$, marked with **[original]**, or the converted ones $\mathbf{z}'_{\{\text{img, txt}\}}$, labeled with **[converted]**.

Evaluations Metrics. We use two standard retrieval metrics: mean Average Precision (mAP@k) and Precision (Prec@k). mAP calculates average precision at different recall levels, while precision measures relevance from the top k retrieved items. k is set following the standard in literature: $k = 200$ on Sketchy-Ext [18], $k = 100$ on TU-Berlin-Ext [18] only for Precision, and $k = 200$ on QuickDraw-Ext [5] only for Precision. For FG-SBIR, we report accuracy@$\{1, 10\}$, the ratio of sketches correctly matching with top-k retrieved photographs, following previous studies.

5. Results and Discussion

5.1. Performance Analysis

Categorical ZS-SBIR. Table 1 compares our approaches with SOTA methods in ZS-SBIR across diverse datasets. The results indicate that the methods with CLIP (CLIP-AT [26] and ours) demonstrate superior performance, proving the effect of leveraging rich semantic information for enhancing zero-shot retrieval. Moreover, ours[clip] outperform all other methods, highlighting the efficacy of our proposed indirect alignment. Notably, we observe a consistent improvement in mAP across all methods by a minimum of 8% for Sketchy, 7% for TU-Berlin, and 12% for QuickDraw, proving our approach is more suitable than earlier triplet methods that directly correlate sketches and photos.

We also observe that our [original] and [converted] surpass [clip] version, indicating efficacy of the additional loss

Table 2. Comparison for Generalized ZS-SBIR

Model	Sketchy-Ext		TU-Berlin-Ext	
	mAP@200	Prec@200	mAP@all	Prec@100
SEM-PCYC [7]	-	-	19.2	29.8
OCEAN [45]	-	-	31.2	34.1
BDA [3]	22.6	33.7	25.1	35.7
SaA [25]	-	-	29.0	38.1
ZSE[Ret] [15]	-	-	46.4	48.5
ZSE[RN] [15]	-	-	43.2	46.0
STL [9]	**63.4**	53.8	40.2	49.8
CLIP-AT*	55.6	62.7	60.9	63.8
Ours [converted]	62.3	**68.5**	**62.6**	**67.8**

Table 3. Comparison on the FG-ZS-SBIR task

Model	Acc@1	Acc@5
CrossGrad [30]	13.40	34.90
CC-DG [21]	22.60	49.00
SketchPVT [27]	**30.24**	51.65
CLIP-AT [26]	28.68	**62.34**
Ours [original]	29.96	58.53
Ours [converted]	29.80	57.94

terms. Furthermore, [converted] surpassing [original] implies our methods effectively address the modality gap.

Generalized ZS-SBIR. In Table 2, we observe that our method consistently outperforms baselines on the Generalized ZS-SBIR. Results highlight again the validity of employing CLIP, evident when comparing the CLIP-AT [26] and ours with all other methods (excluding mAP@200 on Sketchy-Ext, where STL [9] performs the best). Our method outperforms the previous state-of-the-art models by 12% and 3% in mAP, and by 9% and 6% in precision on Sketchy and TU, respectively, showing our method's suitability for both zero-shot learning and generalized settings.

Fine-grained ZS-SBIR. In fine grained setting, we expect our method would not perform well, as our model is better suited for coarse-grained retrieval by design. Surprisingly, however, Table 3 shows that ours comparably performs with state-of-the-art models specially designed for fine-grained setting, with minimal changes described in Sec. 3.3.

Interestingly, we observe that our [original] embeddings outperform our [converted] in this setting, unlike the ZS-SBIR task. We speculate that the conversion to the exact position in the target modality space is more challenging in the fine-grained setting, leading to a slight decline in performance with converted embeddings.

Visualization. Fig. 3 visualizes our learned embedding vectors corresponding to various modalities and classes. We randomly select 100 vectors from *songbird* (pink), *sword* (blue), and *wheelchair* (green) and plot with t-SNE projection. Photos are marked with •, sketches are with ▲, and the converted photo embeddings from sketches are with +, for each class. We observe clear modality gap between photos and sketches, and also see that converted embeddings are clearly closer to the target, distinguished from the origin.

5.2. Ablation Study

Effect of Loss Terms. To evaluate the impact of individual loss terms, we conduct an ablation study to add loss terms

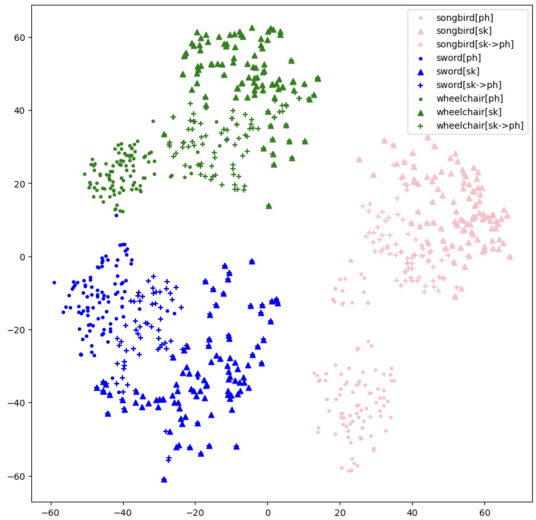

Figure 3. T-SNE visualization on Sketchy-Ext [18]

Table 4. Ablation on Loss Terms for ZS-SBIR

Loss					Sketchy-Ext		TU-Berlin-Ext		QuickDraw-Ext	
$\mathcal{L}_{\text{clip}}$	$\mathcal{L}_{\text{sem}}$	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{mc}}$	$\mathcal{L}_{\text{ortho}}$	m@200	P@200	m@all	P@100	m@all	P@200
✓					68.50	74.90	**70.53**	77.62	32.17	41.85
✓	✓				68.86	75.21	70.52	77.67	32.56	42.44
✓	✓	✓			**69.05**	75.44	70.49	77.68	32.20	42.29
✓	✓	✓	✓		**69.05**	**75.46**	70.45	77.64	32.16	42.25
✓	✓	✓	✓	✓	69.02	75.44	70.40	**77.73**	**32.74**	**42.49**

Table 5. ZS-SBIR on Sketchy-ext with unpaired datasets

Method	mAP@200	Prec@200
CLIP-AT* / 0.8	62.59	70.26
Ours / 0.8	67.97	74.51
Ours / 0.8 + photo / 0.2	68.25	74.67
Ours / 0.8 + sketch / 0.2	**68.56**	**75.11**

one by one, reported in Table 4. (Note that $\mathcal{L}_{\text{clip}}$ is always retained as an essential term for contrasting vectors.)

In general, a better result is achieved when more loss terms are used. Taking a closer look, however, we observe slightly different patterns across datasets. On Sketchy-Ext [18], most proposed losses, except for the orthogonal loss, demonstrate effectiveness when added. We speculate that the more stringent constraints imposed by the orthogonal loss slightly impede achieving superior outcomes. On QuickDraw-Ext [5], the inclusion of $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{mc}}$ leads to a slight decline in performance, while the orthogonal loss significantly improves the performance, unlike the case on Sketchy-Ext. This suggests the potential interactions among these loss terms. Experiments on TU-Berlin-Ext [18] show distinct trends. Introducing more terms brings about a decline in mAP, while the opposite is observed in precision.

Effect of training on additional unpaired dataset. Our indirect alignment not only leads to improved performance but also facilitates semi-supervised approach, taking advantage of additional datasets containing only photographs or sketches. To assess the impact of additional unpaired data, we conduct an experiment on Sketchy-Ext [18] as follows. First, a model is trained on 80% of the seen categories (83

Table 6. Ablation on Model Types for ZS-SBIR

Loss	Sketchy-Ext[18]		TU-Berlin-Ext[18]		QuickDraw-Ext[5]	
	m@200	P@200	m@all	P@100	m@all	P@200
Ours full	**69.05**	**75.46**	**70.49**	**77.68**	**32.74**	**42.49**
Ours (g frozen)	67.25	74.23	68.72	75.12	30.83	37.43

Table 7. Ablation on Textual Intervention

$\lambda_{\text{clip-img}}$	1.0	0.8	0.5
Acc@1	29.96	27.31	24.71
Acc@5	58.53	54.79	70.71

out of 104). Then, instances of either photos or sketches are added from the remaining seen classes, corresponding to the 20% of the entire seen categories (21 out of 104).

As shown in Table 5, seeing more samples even from unrelated classes improves the results. This is aligned with observations in previous studies [1, 26, 27]. We emphasize that other methods assuming paired sketches and photos are not eligible to take advantage of these unimodal datasets.

Taking a deeper look, adding more sketches yields greater advantage than adding more photos. Augmenting sketches is more efficient but expensive. We leave further investigation as a potential future work.

Effect of textual supervision. In categorical SBIR, our proposed model aligns sketches and photos solely via texts. We hypothesize that the text encoder would play an important role, since the image distribution should, to some extent, follow the fixed text distribution if the text encoder is frozen and the text distribution is fixed.

Table 6 evaluates the effect of fine-tuning the text encoder. As expected, the performance slightly drops when the text encoder remains frozen. We interpret that realigning the text encoder leads to a more effective utilization of latent space, perhaps achieving higher uniformity [40] and better adaptation to SBIR datasets, while simultaneously preventing catastrophic forgetting.

For FG-SBIR, we explore several different values for $\lambda_{\text{clip-img}}$ in Table 7. We observe that alignment with the class hinders the exact alignment of identical instances between sketches and photos, even when we unfreeze the text encoder. This is in contrast to [26], which suggests that such alignment facilitates accurate alignment in the latent space.

5.3. Qualitative Results

We qualitatively compare our retrieval results with the strongest baseline, CLIP-AT [26], in Fig. 4 (categorical) and Fig. 5 (fine-grained). Refer to the supplementary material for more examples. Overall, our method retrieves correct photos for most classes.

For the fine-grained task, even incorrectly marked images turn out to be often correct. For instance, for a drawing of a wind turbine (row 2 in Fig. 5), all 5 photos retrieved by our model are indeed wind turbines, although they are not labeled in the ground truth. The baseline also retrieves 3 wind turbine images, but it includes two windmill images

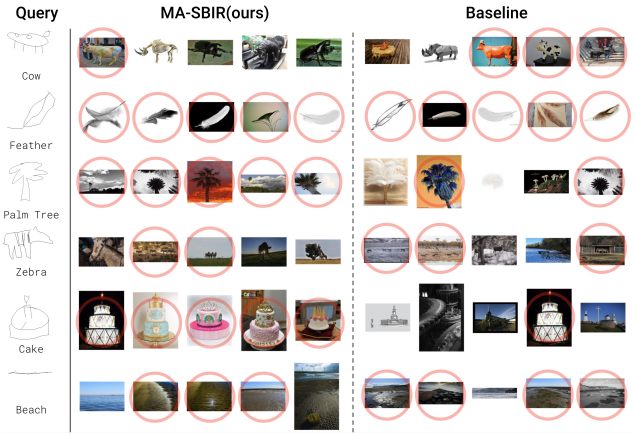

Figure 4. **Categorical-ZS-SBIR.** Top-5 Retrieved images on QuickDraw-Ext [5]. Correct samples are circled.

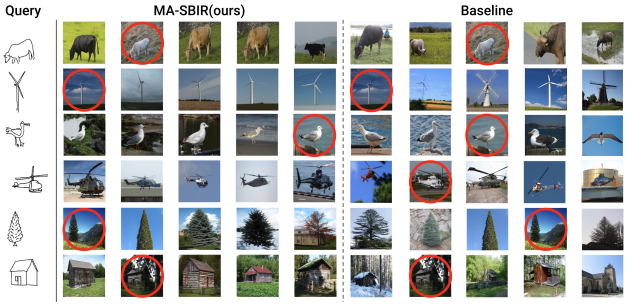

Figure 5. **FG-ZS-SBIR.** Top-5 Retrieved images on Sketchy [5]. Correct samples are circled.

as well (at the 3rd and 5th). In row 3 and 4 of Fig. 5, our method ranks the true image lower than the baseline does, but the retrieved photos actually exhibits greater visual similarity with the queried sketches.

6. Summary and Limitations

We present a novel approach that aligns the joint embedding space by disentangling modality-specific nuances from semantic content. Our method excels in multiple zero-shot scenarios of sketch-based image retrieval, setting a new performance benchmark.

Despite the success on categorical setting, however, our model turns out not to align well at instance level. This is somewhat expected, as our model is designed to be trainable with unpaired sketches and photos, but this is not applicable on fine-grained setting. Improving fine-grained alignment without paired examples will be an interesting future work.

Acknowledgement. This work was supported by the New Faculty Startup Fund from Seoul National University and by National Research Foundation (NRF) grant (No. 2021H1D3A2A03038607/50%, 2022R1C1C1010627/20%, RS-2023-00222663/10%) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No. 2022-0-00264/20%) funded by the government of Korea.

References

[1] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *CVPR*, 2021.

[2] Ushasi Chaudhuri, Biplab Banerjee, Avik Bhattacharya, and Mihai Datcu. CrossATNet - a novel cross-attention based framework for sketch-based image retrieval. *Image and Vision Computing*, 104:104003, 2020.

[3] Ushasi Chaudhuri, Ruchika Chavan, Biplab Banerjee, Anjan Dutta, and Zeynep Akata. BDA-SketRet: Bi-level domain adaptation for zero-shot sbir. *Neurocomput.*, 514(C):245–255, dec 2022.

[4] J. Collomosse, T. Bui, and H. Jin. LiveSketch: Query perturbations for guided sketch-based visual search. In *CVPR*, 2019.

[5] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, 2019.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[7] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019.

[8] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):44:1–44:10, 2012.

[9] Ce Ge, Jingyu Wang, Qi Qi, Haifeng Sun, Tong Xu, and Jianxin Liao. Semi-transductive learning for generalized zero-shot sketch-based image retrieval. In *AAAI*, volume 37, 2023.

[10] Longteng Guo, Jing Liu, Yuhang Wang, Zhonghua Luo, Wei Wen, and Hanqing Lu. Sketch-based image retrieval using generative adversarial networks. In *ACM MM*, 2017.

[11] Fei Huang, Cheng Jin, Yuejie Zhang, Kangnian Weng, Tao Zhang, and Weiguo Fan. Sketch-based image retrieval with deep visual semantic descriptor. *Pattern Recognition*, 76:537–548, 2018.

[12] Taotao Jing, Haifeng Xia, Jihun Hamm, and Zhengming Ding. Augmented multimodality fusion for generalized zero-shot sketch-based visual retrieval. *IEEE Transactions on Image Processing*, 31:3657–3668, 2022.

[13] Jianjun Lei, Yuxin Song, Bo Peng, Zhanyu Ma, Ling Shao, and Yi-Zhe Song. Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):3226–3237, 2020.

[14] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NIPS*, 2022.

[15] Fengyin Lin, Mingkang Li, Da Li, Timothy Hospedales, Yi-Zhe Song, and Yonggang Qi. Zero-shot everything sketch-based image retrieval, and in explainable style. In *CVPR*, 2023.

[16] Hangyu Lin, Yanwei Fu, Peng Lu, Shaogang Gong, Xiangyang Xue, and Yu-Gang Jiang. TC-Net for iSBIR: Triplet classification network for instance-level sketch based image retrieval. In *ACM MM*, 2019.

[17] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017.

[18] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017.

[19] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L. Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV*, 2019.

[20] Peng Lu, Gao Huang, Hangyu Lin, Wenming Yang, Guodong Guo, and Yanwei Fu. Domain-aware se network for sketch-based image retrieval with multiplicative euclidean margin softmax. In *ACM MM*, 2021.

[21] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Generalising fine-grained sketch-based image retrieval. In *CVPR*, 2019.

[22] Kaiyue Pang, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *CVPR*, 2020.

[23] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPSW*, 2017.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[25] Leo Sampaio Ferraz Ribeiro and Moacir Antonelli Ponti. Sketch-an-anchor: Sub-epoch fast model adaptation for zero-shot sketch-based image retrieval. *arxiv:2303.16769*, 2023.

[26] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *CVPR*, 2023.

[27] Aneeshan Sain, Ayan Kumar Bhunia, Subhadeep Koley, Pinaki Nath Chowdhury, Soumitri Chattopadhyay, Tao Xiang, and Yi-Zhe Song. Exploiting unlabelled photos for stronger fine-grained sbir. In *CVPR*, 2023.

[28] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch3T: Test-time training for zero-shot sbir. In *CVPR*, 2022.

[29] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Trans. Graph.*, 35(4), jul 2016.

[30] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi.

Generalizing across domains via cross-gradient training. In *ICLR*, 2018.

[31] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018.

[32] Peiyang Shi, Michael C. Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in CLIP. In *ICLRW*, 2023.

[33] 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.

[34] Haifeng Sun, Jiaqing Xu, Jingyu Wang, Qi Qi, Ce Ge, and Jianxin Liao. DLI-Net: Dual local interaction network for fine-grained sketch-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7177–7189, 2022.

[35] Jialin Tian, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. TVT: Three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval. In *AAAI*, 2022.

[36] Jialin Tian, Xing Xu, Zheng Wang, Fumin Shen, and Xin Liu. Relationship-preserving knowledge distillation for zero-shot sketch based image retrieval. In *ACM MM*, 2021.

[37] Hao Wang, Cheng Deng, Tongliang Liu, and Dacheng Tao. Transferable coupled network for zero-shot sketch-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9181–9194, 2022.

[38] Kai Wang, Yifan Wang, Xing Xu, Xin Liu, Weihua Ou, and Huimin Lu. Prototype-based selective knowledge distillation for zero-shot sketch based image retrieval. In *ACM MM*, 2022.

[39] Luo Wang, Xueming Qian, Xingjun Zhang, and Xingsong Hou. Sketch-based image retrieval with multi-clustering re-ranking. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4929–4943, 2020.

[40] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

[41] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018.

[42] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *CVPR*, 2016.

[43] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. SketchNet: Sketch classification with web images. In *CVPR*, 2016.

[44] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018.

[45] Jiawen Zhu, Xing Xu, Fumin Shen, Roy Ka-Wei Lee, Zheng Wang, and Heng Tao Shen. Ocean: A dual learning approach for generalized zero-shot sketch-based image retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2020.