# Augmented Multimodality Fusion for Generalized Zero-Shot Sketch-Based Visual Retrieval

Taotao Jing, Haifeng Xia, Jihun Hamm, and Zhengming Ding, *Member, IEEE*

***Abstract*— Zero-shot sketch-based image retrieval (ZS-SBIR) has attracted great attention recently, due to the potential application of sketch-based retrieval under zero-shot scenarios, where the categories of query sketches and gallery photos are not observed in the training stage. However, it is still under insufficient exploration for the general and practical scenario when the query sketches and gallery photos contain both seen and unseen categories. Such a problem is defined as generalized zero-shot sketch-based image retrieval (GZS-SBIR), which is the focus of this work. To this end, we propose a novel Augmented Multi-modality Fusion (AMF) framework to generalize seen concepts to unobserved ones efficiently. Specifically, a novel knowledge discovery module named cross-domain augmentation is designed in both visual and semantic space to mimic novel knowledge unseen from the training stage, which is the key to handling the GZS-SBIR challenge. Moreover, a triplet domain alignment module is proposed to couple the cross-domain distribution between photo and sketch in visual space. To enhance the robustness of our model, we explore embedding propagation to refine both visual and semantic features by removing undesired noise. Eventually, visual-semantic fusion representations are concatenated for further domain discrimination and task-specific recognition, which tend to trigger the cross-domain alignment in both visual and semantic feature space. Experimental evaluations are conducted on popular ZS-SBIR benchmarks as well as a new evaluation protocol designed for GZS-SBIR from DomainNet dataset with more diverse sub-domains, and the promising results demonstrate the superiority of the proposed solution over other baselines. The source code is available at https://github.com/scottjingtt/AMF_GZS_SBIR.git.**

***Index Terms*— Sketch-based image retrieval, generalized zero-shot SBIR, multi-modality fusion.**

## I. Introduction

SKETCH-BASED image retrieval (SBIR) enables users to obtain desired photo samples through searching a large-scale database given a probe sketch [1]–[9]. Compared to querying textual descriptions [10]–[14], free-hand sketches are easier to create on the pervasive touchscreen devices in real-life scenarios, which can express the expectation and target candidates more effectively in a visual and concise way, such as shape and pose. Such benefits and expectations drive the increasing attention and progress about SBIR lately.

The bottleneck of conventional SBIR is the presumption of identical label space between the train and test data, that is to say, all categories evaluated in the test stage must have already been observed in the training data, as shown in Fig. 1(a), which the realistic scenarios difficultly satisfy. Zero-shot sketch-based image retrieval (ZS-SBIR) [1]–[3], [15] addresses a special situation when the given probe sketches and retrieval gallery consist of data from categories unseen during the training phase, as Fig. 1(b) shows, which is proposed as a combination of zero-shot learning (ZSL) [16]–[23] and sketch-based image retrieval. The challenges of ZS-SBIR mainly result from domain discrepancy across sketch and photo images, semantic difference between training and test data, as well as insufficient knowledge of unseen categories. To solve the ZS-SBIR problem, massive efforts involve the category-wise semantic information of the training data to explore a visual-semantic relationship from sketch/photo towards the semantic space, then operate retrieval in the low-dimensional semantic feature space [24]–[31].

Conventional ZS-SBIR methods [15], [25], [26] only attempt to identify the instances of unseen classes, while ignoring the retrieval on seen categories. However, practical applications fail to acquire prior knowledge of whether the given probe sketch comes from unseen categories or not. The disadvantage of ZS-SBIR further motivates the development of generalized zero-shot sketch-based image retrieval (GZS-SBIR) [28], [32], [33] where the test data consists of both *seen* and *unseen* classes. In terms of the GZS-SBIR scenario, as shown in Fig. 1 (c), a well-trained retrieval model can not only overcome the challenges of cross-domain discrepancy and insufficient knowledge of unseen classes as in ZS-SBIR, but also achieve promising performance when given probe sketches and gallery photos from either seen or unseen categories.

Unfortunately, the existing literature made limited efforts to define the evaluation standard and solution specifically for the GZS-SBIR challenge. For example, Dutta and Akata [25] directly apply their framework designed for the ZS-SBIR task to a generalized ZS-SBIR setting, which expands the searching space of the gallery photos covering both seen and unseen categories. Later on, Dutta *et al.* [28] propose an unseen class sample detection approach to recognize photos from unseen categories to compare with the probe sketch. Differently, Zhu *et al.* [33] design another generalized ZS-SBIR evaluation protocol by picking out a subset of seen classes together with unseen classes making up the retrieval gallery. However, all these works presume that the probe sketches are from unseen categories, while only the retrieval gallery contains some photos from seen classes.

Considering the aforementioned challenges of GZS-SBIR and the limitations of prior works, we propose a novel Augmented Multi-modality Fusion (AMF) framework by

Manuscript received June 30, 2021; revised December 14, 2021 and February 12, 2022; accepted May 2, 2022. Date of publication May 16, 2022; date of current version May 26, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nikos Deligiannis. *(Corresponding author: Taotao Jing.)*

The authors are with the Department of Computer Science, Tulane University, New Orleans, LA 70118 USA (e-mail: tjing@tulane.edu; hxia@tulane.edu; jhamm3@tulane.edu; zding1@tulane.edu).

Digital Object Identifier 10.1109/TIP.2022.3173815

1941-0042 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.

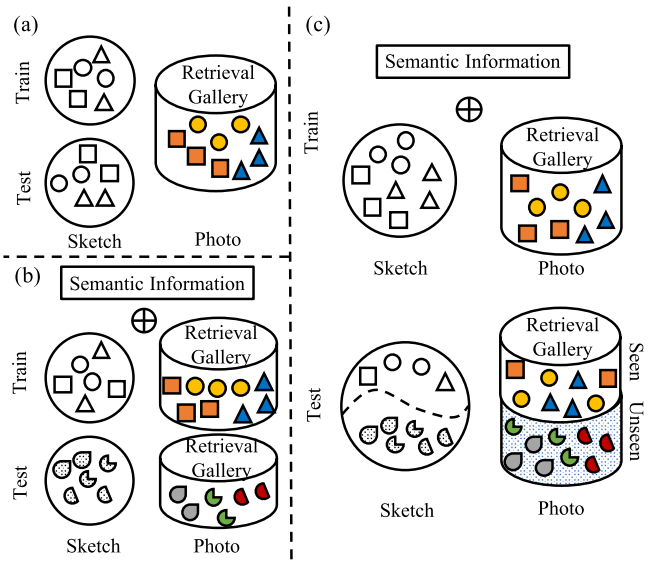



Fig. 1. Illustration of three scenarios of sketch-based image retrieval problem: (a) Conventional sketch-based image retrieval (SBIR), where the probe sketches and gallery photos are from the same classes; (b) Zero-shot sketch-based image retrieval (ZS-SBIR), where the test data are from different class space compared to the training data; (c) Generalized zero-shot sketch-based image retrieval (GZS-SBIR), where the sketches and gallery photos in the test phase come from both *seen* and *unseen* categories compared to the training stage.

leveraging photo-sketch and visual-semantic mismatch simultaneously. Specifically, both sketches and photos are mapped into a domain-invariant visual feature space, then projected to the semantic embedding space. Meanwhile, embedding propagation is explored to refine all the feature representations via removing undesired outliers and noises, and cross-domain alignment alleviates the distribution gap between the sketch and photo domains in two levels. The retrieval process is achieved in a visual-semantic fusion feature space to take advantage of multi-modality benefits. Moreover, cross-domain augmentation in both visual and semantic space expands the data space and promotes the generalizability of the framework to novel knowledge. To evaluate the proposed method for the GZS-SBIR problem, we apply our model on several existing benchmarks following previous works, and construct another new GZS-SBIR evaluation protocol, testing the overall ability on both seen and unseen categories. Our main contributions of this work can be summarized as follows:

- First of all, a novel augmented multi-modality fusion framework is proposed to address generalized zero-shot sketch-based image retrieval (GZS-SBIR) via alleviating the distribution difference across domains and leveraging visual and semantic knowledge simultaneously.
- Secondly, the cross-domain augmentation in both visual and semantic space via mix-up strategy expands the searching range in the feature space and enriches the feature/semantic patterns, which promotes the generalizability of the model for novel knowledge.
- Finally, promising experimental results of the proposed model on existing GZS-SBIR benchmarks and a new evaluation protocol emphasize its effectiveness and superiority over prior GZS-SBIR methods.

## II. Related Work

### A. Zero-Shot Sketch-Based Image Retrieval

Given a free-hand drawn sketch, searching the most corresponding samples from a retrieval gallery consisting of plenty of natural images is defined as the problem sketch-based image retrieval (SBIR). Different from SBIR tasks focusing on the same class space for both sketch and images in training and test stages, users sometimes may look for some objects not present in the training data, but the new categories data do exist in the test data retrieval gallery. Zero-shot sketch-based image retrieval (ZS-SBIR) considers such special scenario, when the test data, both provided sketches and images in the retrieval gallery, is new to the model compared to the training data [15], [25]–[27], [34].

To name a few, CVAE proposes a method based on deep conditional generative model taking the sketch as input and fill the missing information stochastically [34]. ZSIH builds an end-to-end three-network architecture to mitigate the sketch-image heterogeneity and enhance the semantic relations among data [15]. SEM-PCYC proposes a semantically aligned paired cycle-consistent generative adversarial network including two branches translating sketch and image bidirectionally, as well as combining textual and hierarchical side information via an auto-encoder scheme to select discriminating knowledge [25]. Doodle2Search mines the mutual information among domains to alleviate the domain gap and jointly model sketches and photos into a common embedding space [26]. SAKE fine-tunes the model pre-trained on ImageNet dataset in an economical way and leverages semantic information to preserve inter-class relationship knowledge [35]. StyleGuid uses style-guided fake-image generation strategy to address the knowledge gap across classes and domain gap between query and search sets [28]. SketchGCN utilizes the graph convolution network to simultaneously consider both visual and semantic information, and the semantic information can be generated from the visual information using a conditional variational auto-encoder [24]. SAN proposes a generative method based on stacked adversarial network to generate high-quality samples, while takes advantage of Siamese Network to learn a better distance metric compared to nearest neighbor search [27]. PDFD decomposes visual features into domain features and semantic ones, and then the semantic features are projected into common space as retrieval features, and the progressive projection strategy maintains strong semantic supervision [29].

### B. Generalized Zero-Shot Sketch-Based Image Retrieval

Although some ZS-SBIR works consider the scenario where the test data comes from categories unseen during training stage, they only focus on the unseen classes photos in the retrieval gallery, but ignores the overall performance when the probe sketches are from both *seen* and *unseen* categories. Generalized zero-shot sketch-based image retrieval (GZS-SBIR) takes both seen and unseen data into account for applications [25], [28], [32], [33]. Unfortunately, such GZS-SBIR is under insufficient exploration and the evaluation protocols are ambiguous in prior works. Some ZS-SBIR efforts evaluate

Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.

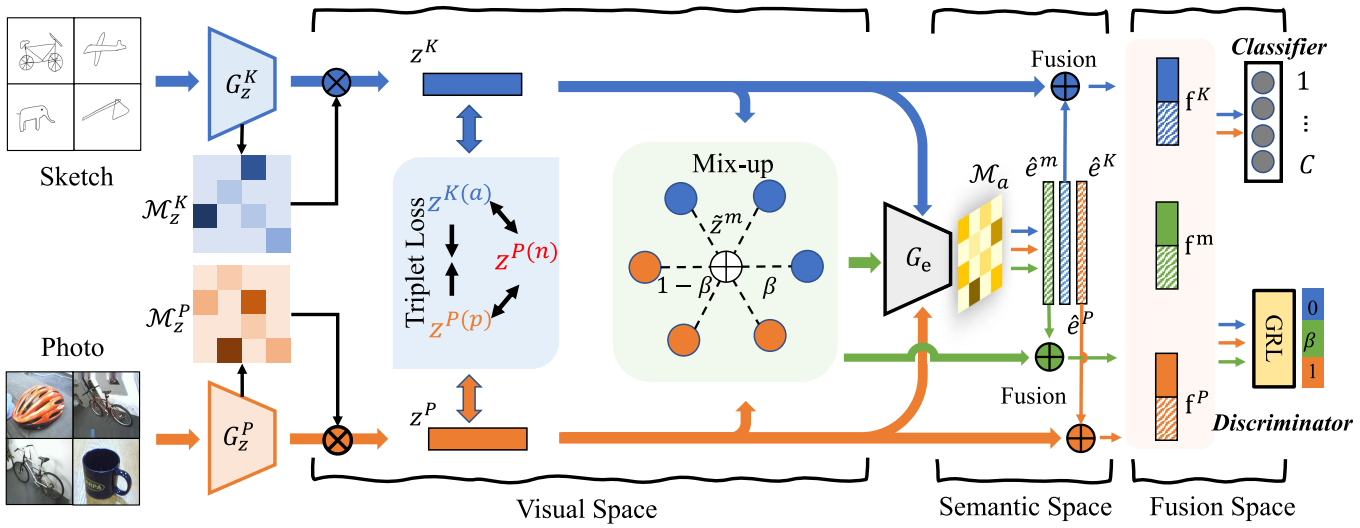



Fig. 2. Illustration of the proposed framework. The model consists of two visual feature generators $G_z^{K/P}(\cdot)$ mapping the sketch and photo samples into a shared domain invariant visual space, then based on the visual features, corresponding semantic information is predicted by the projector $G_e(\cdot)$. Triplet loss and cross-domain mix-up strategies are leveraged to align the distribution across sketch and photo domains, and the embedding propagation based on refines the features through removing undesired noise. The visual-semantic fusion representations are recognized by the classifier $C(\cdot)$ and domain discriminator $D(\cdot)$, which is also used to operate the retrieval task.

their model on the GZS-SBIR setting involving some training data from seen classes into the retrieval gallery [15], [25], [27], [28]. OCEAN [33] differently takes a subset of seen classes into account together with unseen classes to make up the searching class space for GZS-SBIR. Typically, [28] proposed a specific technique to detect if the images from retrieval gallery belong to seen or unseen categories, which presuming that the probe sketches is from unseen categories, while such presumption reduces the generalizability and applicability of GZS-SBIR to real-world scenarios.

To address issues described before, we propose a novel framework seeking to address the GZS-SBIR problem. Moreover, a new evaluation protocol based on DomainNet [36], a large-scale dataset consisting of different visual domains, is proposed to evaluate the overall performance of the model under the GZS-SBIR settting.

## III. The Proposed Algorithm

### A. Preliminaries and Motivation

Given a training set $\mathcal{D}_{tr} = \{\mathcal{K}_{tr}^s, \mathcal{P}_{tr}^s\}$ consisting of data from two domains *sketch* ($\mathcal{K}$) and *photo* ($\mathcal{P}$) drawn from *seen* categories. Both sketch and photo images as well as corresponding class labels are available for training. Specifically, $\mathcal{K}_{tr}^s = \{\mathbf{X}_{tr}^K, \mathbf{Y}_{tr}^K\} = \{(\mathbf{x}_{tr}^{K(i)}, y_{tr}^{K(i)})\}_{i=1}^{n_{tr}^K}$ and $\mathcal{P}_{tr}^s = \{\mathbf{X}_{tr}^P, \mathbf{Y}_{tr}^P\} = \{(\mathbf{x}_{tr}^{P(i)}, y_{tr}^{P(i)})\}_{i=1}^{n_{tr}^P}$ denote the datasets of *sketch* and *photo*, respectively. $y_{tr}^{K/P(i)} \in \mathcal{Y}^s$, where $\mathcal{Y}^s$ means the *seen* categories in the training set. Different from SBIR, GZS-SBIR is more practical and challenging as the probe sketches and retrieval gallery contain both *seen* and *unseen* categories in the test stage, i.e., $\mathcal{Y}_{te} = \mathcal{Y}^s \bigcup \mathcal{Y}^u$, where $\mathcal{Y}^s$ and $\mathcal{Y}^u$ denote *seen* and *unseen* categories, respectively. Mathematically, the test data consists of probe sketches and photo gallery, i.e., $\mathcal{D}_{te} = \{\mathcal{K}_{te}, \mathcal{P}_{te}\}$, where $\mathcal{P}_{te} = \mathcal{P}_{te}^s \bigcup \mathcal{P}_{te}^u$ and $\mathcal{K}_{te} = \mathcal{K}_{te}^s \bigcup \mathcal{K}_{te}^u$, and $\mathcal{P}_{te}^s/\mathcal{K}_{te}^s$ are drawn from the *seen* categories $\mathcal{Y}^s$, while $\mathcal{P}_{te}^u/\mathcal{K}_{te}^u$ are from the *unseen* categories $\mathcal{Y}^u$. Moreover, following previous ZS-SBIR works, semantic information obtained by a word embedding extractor with corresponding class name as input is also accepted for training, denoted as $\mathcal{E}_{tr} = \{\mathbf{e}^c, c \in \mathcal{Y}^s\}$ [25], [26]. For sketches and photos in $\mathcal{K}_{tr}$ and $\mathcal{P}_{tr}$, corresponding semantic information are denoted as $\mathbf{E}_{tr}^K = \{\mathbf{e}_{tr}^{K(i)}\}_{i=1}^{n_{tr}^K}$ and $\mathbf{E}_{tr}^P = \{\mathbf{e}_{tr}^{P(i)}\}_{i=1}^{n_{tr}^P}$, respectively. It is noteworthy that the semantic information for each sample is only decided by its class name, which means semantic representations are shared across sketch and photo domains for the same category. During the test phase, neither class labels nor the semantic knowledge are available, which is similar to the real-world applications.

However, existing ZS-SBIR works have three main shortcomings when they are applied to GZS-SBIR tasks. First, none of previous works are designed specifically to benefit the generalizability of the framework on both seen and unseen categories, which is still under insufficient exploration. Second, although some of previous ZS-SBIR works also evaluate their models on "generalized" tasks, the evaluation protocol is inconsistent and ambiguous because they didn't explicitly explore the performance on the novel classes. We observe extremely low performance on the unseen categories compared to seen classes in generalized tasks, which is not obviously recognized with the overall average results previous works adopted, as the number of seen classes is much larger than the number of unseen classes. Thirdly, we notice that previous works use a subset of the training photos from the seen categories together with the unseen classes samples constructing the test phase generalized retrieval gallery, which is not reasonable due to the reuse of training data for evaluation. Moreover, most existing SBIR tasks only focus on the sketch and photo domains, and we realize the necessity and expectation to retrieve images from different domains or styles (e.g., quick draw, art) besides natural photo. All these motivate us to

Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.

propose a novel framework to solve the GZS-SBIR problem and evaluate it on a new experimental protocol, which involves various visual domains besides sketch and photo and addresses the aforementioned drawbacks of previous works. Details about the new evaluation protocol and explored datasets will be illustrated in IV-A.

### B. Framework Overview

. We solve the GZS-SBIR problem from a domain adaptation perspective with the proposed framework illustrated as Fig. 2 including three modules. ***(i) visual-semantic supervised projection***, mapping the sketches and photos to a shared domain-invariant visual space through two visual feature extractors $G_z^K(\cdot)$ and $G_z^P(\cdot)$, then the visual features and corresponding semantic representations are bridged via the visual-semantic projector $G_e(\cdot)$, supervised by task-specific recognition over classifier $C(\cdot)$. ***(ii) cross-domain distribution alignment***, alleviating the domain distribution gap across sketch and photo domains leveraging both visual and semantic knowledge with multi-modality features fusion. ***(iii) novel knowledge exploration***, promoting the generalization ability of the model with the help of cross-domain augmentation in both visual and semantic space.

### C. Visual-Semantic Supervised Projection

In order to address the retrieval task in a domain-invariant feature space, the sketch and photo samples are projected into a shared visual space by $G_z^{K/P}(\cdot)$, i.e., $\mathbf{Z}_{tr}^{K/P} = G_z^{K/P}(\mathbf{X}_{tr}^{K/P})$. Then the corresponding semantic information is predicted by $G_e(\cdot)$, which is denoted as $\hat{\mathbf{E}}_{tr}^{K/P} = G_e(\mathbf{Z}_{tr}^{K/P})$.

*1) Feature Refinement Through Embedding Propagation:* Moreover, to remove the influence and distraction caused by undesired noise and outlier data, we apply the embedding propagation technique to smooth the manifold and eliminate the noise of both visual and semantic embeddings. The refined representations can be denoted as:

$$\begin{cases} \mathbf{Z}_{tr}^{K/P} = \mathbf{M}_z^{K/P} G_z^{K/P}(\mathbf{X}_{tr}^{K/P}) \\ \hat{\mathbf{E}}_{tr}^{K/P} = \mathbf{M}_e^{K/P} G_e(\mathbf{Z}_{tr}^{K/P}) \end{cases}, \tag{1}$$

where $\mathbf{M}_{z/e}^{K/P}$ is the visual/semantic embedding propagator calculated based on the features $\mathbf{Z}_{tr}^{K/P}$ and $\hat{\mathbf{E}}_{tr}^{K/P}$. For simplification, we explain the procedures to build the embedding propagator matrix $\mathbf{M}_z^K$ for the visual features of one sketch training batch as an example, and the visual features of sketch samples in the batch are denoted as $\{\mathbf{z}^i, i \in [1, \text{batch size}]\}$. Specifically, the adjacency matrix $A$ per batch is calculated as $A^{ij} = \exp(\frac{-(d^{ij})^2}{\sigma^2})$, where $A^{ii} = 0, \forall i$. $d^{ij} = \|\mathbf{z}^i - \mathbf{z}^j\|_2$ is the distance between two samples visual features $\mathbf{z}^i$ and $\mathbf{z}^j$, and $\sigma^2$ is a scaling factor set as $\text{Var}((d^{ij})^2)$ [37]. Then following [37], the embedding propagator $\mathbf{M}_z^K = (I - \alpha L)^{-1}$, in which $I$ is the identity matrix, $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, $D^{ii} = \sum_j A^{ij}$, and $\alpha \in \mathbb{R}$ is a scaling factor fixed as suggested by [37]. Following the same procedure as [37], the embedding propagator for sketch and photo visual features $\mathbf{M}_z^{K/P}$ can be calculated with $\mathbf{Z}_{tr}^{K/P}$ in one training batch, and for the semantic features the embedding propagator $\mathbf{M}_e^{K/P}$ can be calculated based on $\hat{\mathbf{E}}_{tr}^{K/P}$ for each batch. After the embedding propagation, the visual and semantic features of each sample are refined as a weighted combination of all its neighbors, removing undesired noise and smoothing the manifold [37]. In this work, we accept the refined visual features $\mathbf{Z}_{tr}^{K/P}$ and predicted semantic embeddings $\hat{\mathbf{E}}_{tr}^{K/P}$ for all training objectives if not specified.

*2) Mapping From Visual to Semantic Space:* The predicted semantic embeddings based on the sketch and photo visual features are denoted as $\hat{\mathbf{e}}_{tr}^K \in \hat{\mathbf{E}}_{tr}^K$ and $\hat{\mathbf{e}}_{tr}^P \in \hat{\mathbf{E}}_{tr}^P$ for each sample. The visual to semantic projector $G_e(\cdot)$ seeks to map the sketch and photo visual features to the ground-truth semantic vector $\mathbf{e}_{tr}^K$ and $\mathbf{e}_{tr}^P$, which are extracted by the pretrained word-to-vector network $W(\cdot)$ with the corresponding class name as input, and the learning objective is minimizing:

$$\mathcal{L}_{sem} = \frac{1}{n_{tr}^K + n_{tr}^P} \sum_{\mathbf{e}_{tr}^i \in \mathbf{E}_{tr}^K \bigcup \mathbf{E}_{tr}^P} \mathcal{H}(\hat{\mathbf{e}}_{tr}^i, \mathbf{e}_{tr}^i), \tag{2}$$

where $\mathcal{H}(\hat{\mathbf{e}}_{tr}^i, \mathbf{e}_{tr}^i)$ measures the difference between the predicted semantic embeddings and the ground-truth semantic vectors, and $\mathcal{H}(\mathbf{a}, \mathbf{b}) = \frac{1}{2}\big(1 - \frac{\mathbf{a}\mathbf{b}^\top}{\|\mathbf{a}\|\|\mathbf{b}\|}\big)$ for two vectors $\mathbf{a}$ and $\mathbf{b}$. Such supervised semantic embedding prediction loss maps the visual features of samples from the same class into the same semantic embedding, regardless of which domain they are drawn from.

*3) Discriminative Supervision via Multi-Modality Fusion:* Considering the involvement of the semantic knowledge obtained from a pretrained word2vector network, we argue the different characteristics and knowledge between visual and semantic representations. To take advantage of the multi-modality knowledge, visual-semantic fused representations are constructed as: $\mathbf{f}^i = \mathbf{z}^i \oplus \hat{\mathbf{e}}^i$, where $\oplus$ is concatenating operation, $\mathbf{z}^i \in \mathbf{Z}_{tr}^{K/P}$ and $\hat{\mathbf{e}}^i \in \hat{\mathbf{E}_{tr}^{K/P}}$ are visual features and predicted semantic embeddings, respectively. The fused multi-modality representation $\mathbf{f}^i$ is input to the object recognition classifier $C(\cdot)$ to optimize the model via discriminative supervision, and the training objective is to minimize:

$$\mathcal{L}_{cls} = -\frac{1}{n_{tr}^K + n_{tr}^P} \sum_{\mathbf{f}^i \in \mathcal{D}_{tr}} y^i \log C(\mathbf{f}^i). \tag{3}$$

### D. Cross-Domain Distribution Alignment

*1) Domain-Shift Elimination in Visual-Space:* We seek to conduct the retrieval task in a domain-invariant feature space, so we first map both sketches and photos into a shared visual feature space where the two domains features have same or similar discriminative distribution. Then cross-domain alignment is conducted in the visual space to eliminate distribution gap between the sketch and photo domains. Specifically, for each training batch, several triplet sets are constructed as $\mathcal{T} = \{(\mathbf{x}_a^K, y_a^K), (\mathbf{x}_p^P, y_p^P), (\mathbf{x}_n^P, y_n^P)\}$ drawn from $\mathcal{K}_{tr}$ and $\mathcal{P}_{tr}$, such that $y_a^K = y_p^P$ but $y_a^K \neq y_n^P$. The semantic embeddings corresponds to the class labels, $y_a^K = y_p^P$ and $y_n^P$, are denoted as $\mathbf{e}_a^K \in \mathbf{E}_{tr}$ and $\mathbf{e}_{p/n}^P \in \mathbf{E}_{tr}$, respectively. We construct one triplet set for each sketch sample in the training set,

Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.

i.e., $\mathbf{x}_a^K \in \mathcal{K}_{tr}$, together with randomly selected $\mathbf{x}_{p/n}^P \in \mathcal{P}_{tr}^P$ satisfying the requirements. Thus for all $n_{tr}^K$ training sketches, there will be $n_{tr}^K$ triplet sets constructed.

To eliminate the distribution difference between the sketch and photo domains in the visual feature space, for samples in one triplet set $\mathcal{T}$, the visual feature extractor $G_z^{K/P}$ are optimized to reduce the distance between two samples from the same class, while increase the distances between either two samples if they are from different categories. Denoting the refined visual features of the samples in one triplet set produced by the visual feature extractor as $\mathbf{z}_a^K = G_z^K(\mathbf{x}_a^K)$ and $\mathbf{z}_{p/n}^P = G_z^P(\mathbf{x}_{p/n}^P)$ from Eq. (1), the objective function is defined to minimize the triplet loss:

$$\mathcal{L}_{vis} = \frac{1}{n_{tr}^K} \sum_{i=1}^{n_{tr}^K} \mathbf{max}\{0, \mu + \delta_+^i - \delta_-^i\}, \tag{4}$$

where $\delta_+^i = \|\mathbf{z}_a^{K(i)} - \mathbf{z}_p^{P(i)}\|_2$, $\delta_-^i = \|\mathbf{z}_a^{K(i)} - \mathbf{z}_n^{P(i)}\|_2$, and $\mu > 0$ is a marginal parameter fixed as default. The triplet loss is applied to all triplet sets constructed on $n_{tr}^K$ sketches.

*2) Adversarial Adaptation With Fused Features:* Furthermore, the visual-semantic fused representations $\mathbf{f}^i = \mathbf{z}^i \oplus \hat{\mathbf{e}}^i$ are passed to the domain discriminator $D(\cdot)$ to recognize if it is from sketch or photo domain, optimizing the framework in an adversarial manner by minimizing:

$$\mathcal{L}_{dis} = -\frac{1}{n_{tr}^K + n_{tr}^P} \sum_{\mathbf{f}^i \in \mathcal{D}_{tr}} \big(l_d^i \log D(\mathbf{f}^i) + (1 - l_d^i) \log D(\mathbf{f}^i)\big), \tag{5}$$

where $l_d^i \in \{0, 1\}$ is the domain label for the sample $\mathbf{f}^i$, defined as 0 for sketches, while 1 for photos. Moreover, gradient reverse layer (GRL) is inserted before the domain discriminator, which can reverse the gradient direction in back-propagation process, making the model can be optimized through standard back-propagation training strategy [38].

### *E. Novel Knowledge Discovery via Cross-Domain Mix-up*

Lack of knowledge of the unseen categories is one of the most crucial challenges to solve GZS-SBIR. Inspired by zero-shot learning, the semantic attributes or parts of different categories combined together will result in novel species or objects never seen before [18], [39]–[43]. We apply the mix-up strategy to both visual and semantic features across domains and categories, which can create abundant intermediate status synthetic samples between the sketch and photo domains, as well as between various pairs of two categories. Such interaction of features from various domains and categories are not restricted within the seen categories, thus they can expand the searching range in the feature space and enrich the feature patterns, which can be considered as exploring novel knowledge never presented in the training data [44]–[46]. Specifically, the visual and semantic synthetic features through mix-up are constructed as:

$$\begin{cases} \tilde{\mathbf{z}}^{m(i)} = (1 - \beta^i)\mathbf{z}^{K(j)} + \beta^i \mathbf{z}^{P(k)} \\ \tilde{\mathbf{e}}^{m(i)} = (1 - \beta^i)\mathbf{e}^{K(j)} + \beta^i \mathbf{e}^{P(k)} \end{cases} \tag{6}$$

where $\beta^i \in [0, 1], \beta^i \sim \mathbf{Beta}(\gamma, \gamma)$ is a random cross domain mix-up ratio, $\mathbf{z}^{K(j)}$ and $\mathbf{z}^{P(k)}$ are randomly constructed sketch and photo samples pair from sketch and photo domains, respectively, and $\mathbf{e}^{K(j)}$ and $\mathbf{e}^{P(k)}$ are corresponding ground-truth semantic embeddings. For each training batch, the same number as the batch size of synthesized samples are created, i.e., $\mathcal{D}_m = \{\tilde{\mathbf{z}}^{m(i)}, \tilde{\mathbf{e}}^{m(i)}, \beta^i\}_{i=1}^{n_{tr}^K}$, and $\gamma = 2$ for all experiments following [39].

To improve the generalization ability of the visual to semantic projector $G_e(\cdot)$ and learn a continuous visual-semantic relationship for novel knowledge never present in the training data, the synthetic visual feature $\tilde{\mathbf{z}}^{m(i)}$ is mapped to predict the synthetic semantic feature $\tilde{\mathbf{e}}^{m(i)}$. Specifically, the synthetic visual feature $\tilde{\mathbf{z}}^{m(i)}$ is input to the visual-to-semantic projector $G_e(\cdot)$ which outputs the predicted semantic embedding $\hat{\mathbf{e}}^{m(i)} = \mathcal{M}_e^m G_e(\tilde{\mathbf{z}}^{m(i)})$. As mentioned before, the visual and semantic features are the refined via the embedding propagation described in Eq. (1), and the refinement propagator $\mathcal{M}_e^m$ for the synthetic features is calculated following the same rule. Then the semantic-embedding prediction loss for synthetic data is similarly defined as:

$$\mathcal{L}_{sem}^{mix} = \frac{1}{n_{tr}^K} \sum_{i=1}^{n_{tr}^K} \mathcal{H}(\hat{\mathbf{e}}^{m(i)}, \tilde{\mathbf{e}}^{m(i)}). \tag{7}$$

Furthermore, as $\beta^i \in [0, 1]$ is the mix-up ratio in Eq. (6), and the domain labels of the sketch and photo are defined as 0 and 1, respectively. The synthetic features lie in somewhere in between the sketch and photo domains, thus $\beta^i = (1 - \beta^i) \times 0 + \beta^i \times 1$ is more proper to be assigned as a soft domain label for the synthesized sample $\tilde{\mathbf{z}}^{m(i)}$. Combining the synthetic visual features $\tilde{\mathbf{z}}^{m(i)}$ and the predicted semantic embeddings $\hat{\mathbf{e}}^{m(i)}$ to make up the visual-semantic fusion representations for synthetic samples as $\tilde{\mathbf{f}}^{m(i)} = \tilde{\mathbf{z}}^{m(i)} \oplus \hat{\mathbf{e}}^{m(i)}$. Then the domain discrimination loss for the synthetic samples with soft domain labels is defined as:

$$\mathcal{L}_{dis}^{mix} = \frac{1}{n_{tr}^K} \sum_{i=1}^{n_{tr}^K} \big(\beta^i \log D(\tilde{\mathbf{f}}^{m(i)}) + (1 - \beta^i) \log D(\tilde{\mathbf{f}}^{m(i)})\big). \tag{8}$$

By integrating real and augmented data together, we can refine the semantic embedding prediction loss and soft label domain discrimination loss as:

$$\begin{aligned} \mathcal{L}'_{sem} &= \mathcal{L}_{sem} + \mathcal{L}_{sem}^{mix}, \\ \mathcal{L}'_{dis} &= \mathcal{L}_{dis} + \mathcal{L}_{dis}^{mix}. \end{aligned} \tag{9}$$

### *F. Overall Objective*

Combining the task-specific recognition loss, visual space domain-shift elimination loss, together with semantic embedding prediction loss and adversarial domain discrimination loss for both real and synthetic samples, we obtain our overall learning objective as:

$$\min_{G_z^K, G_z^P, G_e, C, D} \mathcal{L}_{cls} + \mathcal{L}_{vis} + \mathcal{L}'_{sem} + \mathcal{L}'_{dis}. \tag{10}$$

Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.

Thanks to the use of a gradient reverse layer (GRL) inserted before the discriminator $D(\cdot)$, all the trainable networks can be updated through standard back-propagation optimization rule, without bother of alternative adversarial training [47], [48].

## IV. Experiments

### A. Datasets and Experimental Settings

*1) Sketchy-Extended Dataset [49]:* it consists of fine-grained sketches and corresponding photos drawn from 125 categories. Originally there are 75,471 hand-drawn sketches and 12,500 photos, then Liu *et al.* extended the photo set yielding a total of 73,002 photos available [4]. Some of previous zero-shot sketch-based image retrieval (ZS-SBIR) works randomly split the 125 categories into 100 as *seen* classes and the rest 25 classes as *unseen* [25]. However, considering to the reliance on the backbone networks pretrained on ImageNet dataset [50], e.g., VGG-16 [51] or ResNet-50 [52], Yelamarthi *et al.* proposes another partition taking 21 categories never present in the 1,000 ImageNet dataset as *unseen* categories for test, while the rest 104 classes data is used for training [34]. All experiments in this work on Sketchy-Extended dataset follow the ImageNet-Exclusive split. To evaluate the performance of our proposed method for the GZS-SBIR tasks, following the settings of [25], [26], 20% of the *seen* classes training photos are used to extend the unseen categories photos retrieval gallery resulting in the searching space covering both *seen* and *unseen* categories.

*2) TUBerlin-Extended Dataset [53]:* it contains 20,000 sketches uniformly distributed in 250 categories, and 204,489 real images from the same categories are provided by [4] and used for retrieval tasks. Similarly, for ZS-SBIR task, we carefully picked out 30 categories never present in the ImageNet 1,000 categories as the *unseen* classes for test, while the rest 220 classes of sketches and real images are *seen* categories for training [35]. To construct GZS-SBIR task, 20% of the *seen* categories real images are added to the *unseen* images to make generalized retrieval gallery containing both *seen* and *unseen* classes.

*3) A New GZS-SBIR Evaluation Protocol:* All previous works tried to explore GZS-SBIR problem only extended the retrieval gallery to cover both seen and unseen classes, ignoring the universal performance of the model when the query sketches may also come from both seen and unseen categories. Besides, both Sketchy-Extended and TUBerlin-Extended datasets only consider sketch-photo domains, ignoring the demand to retrieve images from different domains in addition to photos. These two issues motivate us to propose a new evaluation protocol specified for GZS-SBIR.

In reality, we would retrieve from a more enriched database with various styles or domains, such as photos or painting, given probe sketches. Thus, we construct several retrieval tasks based on the DomainNet with 6 "sketch-like" and "photo-like" domains [36]. Specifically, DomainNet contains about 0.6 million samples from 345 categories distributed in 6 domains: *Sketch* (Sk), *Real* (Re), *Quickdraw* (Qu), *Painting* (Pa), *Infograph* (In), *Clipart* (Cl). The *Sketch* and *Quickdraw* are two "sketch-like" domains and we use the samples from one of them as the probe sketches. Samples in the *Sketch* domain are more professional and accurate to the corresponding photo objects, while domain *Quickdraw* consists of rough conceptual abstractions of objects produced in an amateur drawing style. The differences of *Quickdraw* domain compared to *Sketch* domain introduce more realistic challenges, e.g., large domain gap between the amateur style sketches and photos, large intra-class variability caused by the high abstraction level of different drawers. The rest four domains are treated as "photo-like" domains, which perform as the retrieval gallery in each task. By selecting one of the "sketch-like" domain as the probe set, and one of the "photo-like" domains making up the retrieval gallery, we can construct 8 GZS-SBIR tasks.

TABLE I
Statistical Characteristics of Experimental Datasets

| Dataset | domain | # samples | seen/unseen split |
|---|---|---|---|
| Sketchy | Sketch | 75,471 | 104 / 21 |
| | *Photo* | 73,002 | |
| TU-Berline | Sketch | 20,000 | 220 / 30 |
| | *Photo* | 204,489 | |
| DomainNet | Sketch | 69,128 | 300 / 45 |
| | QuickDraw | 172,500 | |
| | *Real* | 172,947 | |
| | *Infograph* | 51,605 | |
| | *Clipart* | 48,129 | |
| | *Painting* | 72,266 | |

To evaluate the proposed model on GZS-SBIR problem and test its universal performance over both seen and unseen categories, we propose a realistic experimental split for the tasks on DomainNet. Specifically, we carefully choose 45 categories never present in the ImageNet dataset as the *unseen* categories, which will be used only in the test phase, and the rest 300 categories work as *seen* categories. For the "sketch-like" domains, i.e., *Sketch* and *Quickdraw*, we follow the rule of [36] to split the *seen* categories data into *train* and *test* sets. During training stage, only *train* set from *seen* categories are available, while for the test stage, the *test* set data from *seen* categories and all *unseen* categories sketches are used to evaluate. For the retrieval gallery, i.e., "photo-like" domains, all *seen* categories data build up the retrieval gallery for the training, while in the test phase, all *unseen* categories data together with the same number of randomly drawn *seen* classes samples make up the retrieval gallery for evaluation.

### B. Implementation and Evaluation Metrics

*1) Implementation Details:* We accept ImageNet [50] pretrained VGG-16 [51] as the backbone for experiments on Sketchy-Extended and TUBerlin-Extended datasets, and ResNet-50 [52] for the new evaluation protocol on DomainNet. Specifically, the last fully-connected layer of the backbone is replaced by two trainable fully-connected layers with the output dimension of the hidden layer is 1,024 and the last layer output dimension is 512, resulting in two different visual feature generator $G_z^{K/P}(\cdot)$ for sketches and photos, respectively. The convolutional layers are initialized and fixed as the pretrained parameters. Besides, the visual-to-semantic projector $G_e(\cdot)$ is a two-layer fully-connected layer network with hidden layer and last layer output dimension as 300,

Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.

TABLE II
COMPARISON OF THE GENERALIZED ZERO-SHOT SKETCH-BASED IMAGE RETRIEVAL PERFORMANCE ON DOMAINNET. (RESNET-50)

| | | **Doodle2Search** [26] | | | | **SEM-PCYC** [25] | | | | **Ours** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task | TestSet | **mAP@all** | Prec@100 | mAP@200 | Prec@200 | **mAP@all** | Prec@100 | mAP@200 | Prec@200 | **mAP@all** | Prec@100 | mAP@200 | Prec@200 |
| Sk → Re | S | 0.1868 | 0.1563 | 0.2119 | 0.108 | 0.2699 | 0.2046 | 0.2905 | 0.1144 | **0.3318** | **0.2539** | **0.3425** | **0.1621** |
| | U | 0.0965 | 0.1272 | 0.1489 | 0.1261 | 0.0570 | 0.0716 | 0.0865 | 0.0732 | **0.1365** | **0.1633** | **0.1866** | **0.1633** |
| | H | 0.1273 | 0.1403 | 0.1749 | 0.1164 | 0.0941 | 0.1061 | 0.1333 | 0.0893 | **0.1934** | **0.1988** | **0.2416** | **0.1627** |
| Sk → Cl | S | 0.1741 | 0.0721 | 0.2056 | 0.0440 | 0.2714 | 0.0718 | 0.2815 | 0.0366 | **0.3776** | **0.1225** | **0.3799** | **0.0686** |
| | U | 0.0861 | 0.1298 | 0.1572 | 0.1103 | 0.0411 | 0.0527 | 0.0740 | 0.0459 | **0.0941** | **0.1341** | **0.1618** | **0.1157** |
| | H | 0.1152 | 0.0927 | 0.1782 | 0.0629 | 0.0714 | 0.0608 | 0.1172 | 0.0407 | **0.1507** | **0.1280** | **0.2356** | **0.0861** |
| Sk → In | S | 0.0866 | 0.0388 | 0.1324 | 0.0242 | 0.1325 | 0.0031 | 0.0173 | 0.0017 | **0.3317** | **0.0991** | **0.3438** | **0.0573** |
| | U | 0.0467 | **0.0769** | **0.1058** | **0.0639** | 0.0134 | 0.0055 | 0.01 | 0.0081 | **0.0486** | 0.0691 | 0.0954 | 0.0606 |
| | H | 0.0607 | 0.0516 | 0.1176 | 0.0351 | 0.0243 | 0.0040 | 0.0127 | 0.0028 | **0.0848** | **0.0814** | **0.1494** | **0.0589** |
| Sk → Pa | S | 0.1311 | 0.0898 | 0.1738 | 0.0598 | 0.2697 | 0.1307 | 0.2762 | 0.0734 | **0.3693** | **0.1806** | **0.3763** | **0.1100** |
| | U | 0.0664 | 0.1021 | 0.1295 | 0.0972 | 0.0319 | 0.0381 | 0.0534 | 0.0375 | **0.0901** | **0.1283** | **0.1530** | **0.1241** |
| | H | 0.0882 | 0.0956 | 0.1484 | 0.0740 | 0.0308 | 0.0590 | 0.0895 | 0.0496 | **0.1449** | **0.1500** | **0.2175** | **0.1166** |
| Qu → Re | S | 0.1855 | 0.1507 | 0.2189 | 0.1048 | 0.2195 | 0.1653 | 0.2367 | 0.1001 | **0.4252** | **0.3149** | **0.4287** | **0.2002** |
| | U | 0.0663 | 0.0836 | 0.1002 | 0.0844 | 0.0323 | 0.0339 | 0.0472 | 0.0355 | **0.0749** | **0.0779** | **0.1001** | **0.0867** |
| | H | 0.0977 | 0.1075 | 0.1375 | 0.0935 | 0.0563 | 0.0563 | 0.0787 | 0.0524 | **0.1274** | **0.1249** | **0.1623** | **0.1210** |
| Qu → Cl | S | 0.1925 | 0.0775 | 0.2332 | 0.0466 | 0.2060 | 0.0581 | 0.2324 | 0.0302 | **0.4407** | **0.1409** | **0.4408** | **0.0784** |
| | U | 0.0523 | 0.0687 | 0.0894 | 0.0625 | 0.0341 | 0.0412 | 0.062 | 0.0356 | **0.0609** | **0.0836** | **0.1116** | **0.0736** |
| | H | 0.0823 | 0.0728 | 0.1293 | 0.0534 | 0.0341 | 0.0412 | 0.0620 | 0.0356 | **0.1070** | **0.1049** | **0.1781** | **0.0759** |
| Qu → In | S | 0.0581 | 0.0269 | 0.0988 | 0.0183 | 0.1413 | 0.0459 | 0.1637 | 0.0256 | **0.4547** | **0.1214** | **0.4633** | **0.0692** |
| | U | 0.0272 | **0.0395** | **0.0599** | 0.0356 | 0.0229 | 0.0212 | 0.036 | 0.0208 | **0.0298** | 0.0392 | 0.0525 | **0.0370** |
| | H | 0.0371 | 0.0320 | 0.0746 | 0.0242 | 0.0394 | 0.0290 | 0.0590 | 0.0230 | **0.0559** | **0.0593** | **0.0943** | **0.0482** |
| Qu → Pa | S | 0.1381 | 0.0705 | 0.1777 | 0.0471 | 0.2393 | 0.1084 | 0.2528 | 0.0619 | **0.4541** | **0.2044** | **0.4554** | **0.1220** |
| | U | 0.0384 | 0.0527 | 0.0720 | 0.0504 | 0.0248 | 0.0145 | 0.0280 | 0.0163 | **0.0458** | **0.0540** | **0.0780** | **0.0548** |
| | H | 0.0601 | 0.0603 | 0.1025 | 0.0487 | 0.0449 | 0.0256 | 0.0504 | 0.0258 | **0.0832** | **0.0854** | **0.1332** | **0.0756** |

which is the same as the output dimension of the word-to-vector model pretrained on Google News dataset ($\sim$ 100 billion words) [55], when given the class name as input. The classifier $C(\cdot)$ and discriminator $D(\cdot)$ are both two-layer fully-connected layer network with hidden layer output dimension as 512. It is noteworthy that the output dimension of $C(\cdot)$ is the number of *seen* categories as the *unseen* data is not available for training. All hidden layers use ReLU$(\cdot)$ as the activation function, and the whole network parameters are optimized by Adam optimizer with learning rate as $10^{-3}$. Early stop strategy is applied with 10% of the training data as validation set, and the maximal training epochs are set as 500. For parameters, we follow previous works [25], [26], [37], [39] to fix $\mu$ as 1 and $\alpha$ as 0.2. The cross-domain mix-up ratio $\beta^i \in [0, 1]$ is randomly drawn from Beta$(\gamma, \gamma)$ where $\gamma = 2$ [39].

*2) Evaluation Metrics:* Following prior works [25], mean average precision of all retrieved images (**mAP@all**) and top 200 results (**mAP@200**), the precision of top 200 (**Prec@200**) and top 100 results (**Prec@100**) are accepted as the evaluation metrics. Moreover, for the experimental tasks on DomainNet, the performance on the *seen* categories probe sketches are denoted as "S", and for the *unseen* categories are reported as "U". Besides, the harmonic mean of the *seen* and *unseen* categories are computed as $\mathrm{H} = \frac{2\times \mathrm{S}\times \mathrm{U}}{\mathrm{S}+\mathrm{U}}$.

### C. Comparison Results

In this section, we report the results of the proposed model applied to the new evaluation protocol in Table II and Table III, as well as two popular benchmarks (Sketchy-Extended and TUBerlin-Extended) with ImageNet-Exclusive Split in Table V following prior evaluation metrics [25], [26], [28], [29], [34], [54].

First, for the new evaluation protocol constructed on DomainNet dataset, we compare with four existing ZS-SBIR works, i.e., Doodle2Search [26], SEM-PCYC [25], PDFD [29], and TCN [54], which are implemented based on the official codes released by the authors. From the results in Table II and Table III, we observe that our proposed model outperforms all compared baselines for all tasks under different evaluation metrics. In most tasks, our model not just achieves the best performance on the seen classes, but also obtains promising results for the unseen categories, resulting in the best overall performance. For example, in Table II, our model achieves the best mAP@all results for task Sk→Re (S:0.3318, U: 0.1365, H: 0.1934), outperforming both compared baselines significantly. Moreover, we apply the iterative quantization (ITQ) to obtain binary codes for the representations of sketches and photos, and report the results in Table III [56]. It is noteworthy that in Table III, PDFD-binary achieves the best performance on some tasks for unseen categories, but fails on the seen classes, leading to lower overall results than ours.

Moreover, Table V reports the results on the Sketchy-Extended and TUBerlin-Extended benchmarks, as well as compared baselines following the same settings and splits. The results are quoted from the original paper or produced by the official codes applied to the ImageNet-Exclusive split. From the results, we notice that our proposed model achieves the best performance for both ZS-SBIR and GZS-SBIR settings on both Sketchy-Extended and TUBerlin-Extended datasets. For example, compared to the baseline on Sketchy-Extended dataset, our model improves the mAP@all over 0.06 and 0.07 for ZS-SBIR and "Generalized ZS-SBIR", respectively. It is noteworthy that the "Generalized ZS-SBIR" tasks in Table V considered in prior works on

Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.

TABLE III
COMPARISON OF THE GENERALIZED ZERO-SHOT SKETCH-BASED IMAGE RETRIEVAL PERFORMANCE ON DOMAINNET. (RESNET-50)

| | | **PDFD** [29] | | **TCN** [54] | | **Ours** | | **PDFD-binary** [29] | | **TCN-binary** [54] | | **Ours-binary** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task | TestSet | **mAP@all** | Prec@100 | **mAP@all** | Prec@100 | **mAP@all** | Prec@100 | **mAP@all** | Prec@100 | **mAP@all** | Prec@100 | **mAP@all** | Prec@100 |
| Sk → Re | S | 0.0979 | 0.0969 | 0.0587 | 0.0636 | **0.3318** | **0.2539** | 0.1510 | 0.1440 | 0.0328 | 0.0441 | 0.3256 | 0.2479 |
| | U | 0.1464 | 0.2294 | 0.1189 | 0.1899 | 0.1365 | 0.1633 | **0.1554** | **0.2377** | 0.0837 | 0.1464 | 0.1301 | 0.1558 |
| | H | 0.1173 | 0.1362 | 0.0786 | 0.0953 | **0.1934** | **0.1988** | 0.1532 | 0.1793 | 0.0471 | 0.0678 | 0.1859 | 0.1913 |
| Sk → Cl | S | 0.1101 | 0.0568 | 0.0378 | 0.0260 | **0.3776** | 0.1225 | 0.1507 | 0.0759 | 0.0220 | 0.0195 | 0.3731 | **0.1269** |
| | U | 0.0378 | 0.0260 | 0.0938 | 0.1492 | 0.0941 | 0.1341 | **0.1221** | **0.1860** | 0.0645 | 0.1135 | 0.0883 | 0.1254 |
| | H | 0.0563 | 0.0357 | 0.0539 | 0.0443 | 0.**1507** | **0.1280** | 0.1349 | 0.1078 | 0.0328 | 0.0333 | 0.1428 | 0.1261 |
| Sk → In | S | 0.0789 | 0.0411 | 0.0130 | 0.0108 | **0.3317** | **0.0991** | 0.0953 | 0.0485 | 0.0080 | 0.0087 | 0.3196 | 0.0815 |
| | U | 0.0406 | 0.0693 | 0.0322 | 0.0555 | 0.0486 | 0.0691 | **0.0493** | **0.0849** | 0.0270 | 0.0484 | 0.0339 | 0.0527 |
| | H | 0.0536 | 0.0516 | 0.0185 | 0.0181 | **0.0848** | **0.0814** | 0.065 | 0.0617 | 0.0123 | 0.0147 | 0.0613 | 0.0640 |
| Sk → Pa | S | 0.1014 | 0.0728 | 0.0328 | 0.0310 | **0.3693** | **0.1806** | 0.1342 | 0.0945 | 0.0182 | 0.0219 | 0.3601 | 0.1725 |
| | U | 0.0956 | 0.1642 | 0.0840 | 01435 | 0.0901 | 0.1283 | **0.1036** | **0.1745** | 0.0603 | 0.1117 | 0.0828 | 0.1179 |
| | H | 0.0984 | 0.1009 | 0.0472 | 0.0620 | **0.1449** | **0.1500** | 0.1169 | 0.1226 | 0.028 | 0.0366 | 0.1346 | 0.1401 |
| Qu → Re | S | 0.0174 | 0.0207 | 0.0064 | 0.0071 | **0.4252** | **0.3149** | 0.0461 | 0.0550 | 0.0049 | 0.0063 | 0.4136 | 0.3007 |
| | U | 0.0691 | **0.1021** | 0.0404 | 0.0623 | **0.0749** | 0.0779 | 0.0610 | 0.1006 | 0.0313 | 0.0531 | 0.0664 | 0.0627 |
| | H | 0.0278 | 0.0344 | 0.011 | 0.0127 | **0.1274** | **0.1249** | 0.0525 | 0.0711 | 0.0085 | 0.0113 | 0.1144 | 0.1038 |
| Qu → Cl | S | 0.0492 | 0.0346 | 0.0078 | 0.0072 | **0.4407** | **0.1409** | 0.0722 | 0.0497 | 0.0057 | 0.0060 | 0.4332 | 0.1267 |
| | U | 0.0545 | **0.0881** | 0.0347 | 0.0549 | **0.0609** | 0.0836 | 0.0536 | 0.0831 | 0.0284 | 0.0450 | 0.0525 | 0.0738 |
| | H | 0.0517 | 0.0497 | 0.0127 | 0.0127 | **0.1070** | **0.1049** | 0.0615 | 0.0622 | 0.0095 | 0.0106 | 0.0937 | 0.0933 |
| Qu → In | S | 0.0190 | 0.0164 | 0.0036 | 0.0030 | **0.4547** | **0.1214** | 0.0311 | 0.0241 | 0.0031 | 0.0029 | 0.4384 | 0.1131 |
| | U | 0.0195 | 0.0289 | 0.0161 | 0.0203 | **0.0298** | **0.0392** | 0.0238 | **0.0392** | 0.0158 | 0.0205 | 0.0257 | 0.0316 |
| | H | 0.0192 | 0.0209 | 0.0059 | 0.0052 | **0.0559** | **0.0593** | 0.027 | 0.0298 | 0.0052 | 0.0051 | 0.0486 | 0.0494 |
| Qu → Pa | S | 0.0252 | 0.0235 | 0.0049 | 0.0051 | **0.4541** | **0.2044** | 0.0417 | 0.0378 | 0.0039 | 0.0046 | 0.4423 | 0.1825 |
| | U | 0.0294 | 0.0479 | 0.0234 | 0.0330 | **0.0458** | **0.0540** | 0.0291 | 0.0464 | 0.0198 | 0.0283 | 0.0327 | 0.0442 |
| | H | 0.0271 | 0.0315 | 0.0081 | 0.0088 | **0.0832** | **0.0854** | 0.0343 | 0.0417 | 0.0065 | 0.0079 | 0.0609 | 0.0712 |

the Sketchy-Extended and TUBerlin-Extended datasets is a sub-task of the GZS-SBIR problem on DomainNet addressed in this paper, because they only consider the probe sketches from the *unseen* categories.

### *D. Empirical Analysis*

*1) Time and Space Cost Analysis:* In Table IV, we compare the average retrieval time cost of our model and two compared baselines [29], [54] on selected tasks from DomainNet dataset. Following previous works calculating the retrieval time cost [25], during evaluation, all images in the retrieval gallery are projected into the latent embedding space and stored in the memory, then given one query sketch sample, the time cost of calculating similarities between the query sketch and all images in the gallery, as well as sorting the similarities is recorded, and the average time cost of one query sketch over all query samples is reported in Table IV. Denoting the number of samples in the retrieval gallery as $N$, dimension of latent representation as $D$, the time complexity and space complexity are decided by $N$ and $D$, without much differences observed between seen and unseen categories. From the experimental results, we observe that retrieving by binary coding is much faster than by the original features.

*2) Feature Retrieval Comparison:* To explore the difference among the visual features, semantic features, and visual-semantic fusion representations, we show the t-SNE of three kind features extracted by our model for the sketch data in task Sk→Pa on DomainNet as Fig. 3. The seen and unseen categories are displayed in two separate sub-figures for better observation. We notice that the visual-semantic features are more discriminative distributed compared to the visual and semantic features. Moreover, from the experimental results showed in Fig. 3, we observe that the visual-semantic fusion representations achieve the best retrieval performance compared to the others. Such observation demonstrates the effectiveness and contribution of the proposed visual-semantic fusion scheme benefiting the retrieval task.

TABLE IV
COMPARISON OF THE RETRIEVAL TIME COST

| Method | Feature Size | Inference Time Cost (s) Sk → Cl | Sk → Re | Qu → Cl | Qu → Re |
|---|---|---|---|---|---|
| PDFD [29] | 512 | $3.5 \times 10^{-3}$ | $2.1 \times 10^{-2}$ | $6.5 \times 10^{-3}$ | $2.0 \times 10^{-2}$ |
| PDFD-binary [29] | 512 | $2.1 \times 10^{-3}$ | $1.3 \times 10^{-2}$ | $4.1 \times 10^{-3}$ | $1.3 \times 10^{-2}$ |
| TCN [54] | 512 | $3.5 \times 10^{-3}$ | $1.8 \times 10^{-2}$ | $3.3 \times 10^{-3}$ | $2.0 \times 10^{-2}$ |
| TCN-binary [54] | 512 | $2.1 \times 10^{-3}$ | $9.6 \times 10^{-3}$ | $2.1 \times 10^{-3}$ | $1.2 \times 10^{-2}$ |
| Ours | 512 | $3.3 \times 10^{-3}$ | $2.0 \times 10^{-2}$ | $3.1 \times 10^{-3}$ | $1.2 \times 10^{-2}$ |
| Ours-binary | 512 | $1.8 \times 10^{-3}$ | $1.2 \times 10^{-2}$ | $1.8 \times 10^{-3}$ | $1.1 \times 10^{-2}$ |

*3) Ablation Study:* In addition, Fig. 4 shows the results on DomainNet Sk→Pa task obtained by different variants of our model by removing one specific term to explore the contribution of each component. We focus on the contributions of three modules of our model designed for the generalized capability of the model, and we observe their crucial contributions for the performance on unseen classes. Specifically, "w/o Triplet" denotes variant without the visual-space domain-shift elimination via the triplet loss, "w/o Mix-up" means the variant without cross-domain mix-up for novel knowledge exploration, and "w/o EP" represents the variant by removing the features refinement through embedding propagation strategy. From the results, we can observe that our

Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.

TABLE V

COMPARISON OF (GENERALIZED) ZS-SBIR PERFORMANCE ON SKETCHY-EXTENDED AND TUBERLIN-EXTENDED (IMAGENET-EXCLUSIVE SPLIT)

| | Method | Sketchy-Extended [49] | | | | TUBerlin-Extended [53] | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | mAP@all | Prec@100 | mAP@200 | Prec@200 | mAP@all | Prec@100 | mAP@200 | Prec@200 |
| ZS-SBIR | SEM-PCYC [25] | 0.2554 | 0.3688 | 0.4047 | 0.3474 | 0.1842 | 0.2478 | 0.2713 | 0.2364 |
| | Doodle2Search [26] | 0.3691 | - | 0.4606 | 0.3704 | 0.0923 | 0.0891 | 0.1432 | 0.1018 |
| | Ours | **0.4280** | **0.5046** | **0.5287** | **0.4766** | **0.1963** | **0.2863** | **0.3074** | **0.2690** |
| Generalized ZS-SBIR | SEM-PCYC [25] | 0.2355 | 0.3405 | 0.3749 | 0.3229 | 0.1415 | 0.2192 | 0.2411 | 0.2098 |
| | Doodle2Search [26] | 0.3104 | 0.3747 | 0.4214 | 0.3348 | 0.0439 | 0.0491 | 0.0685 | 0.0448 |
| | Ours | **0.3800** | **0.4383** | **0.4545** | **0.4249** | **0.1622** | **0.2376** | **0.2587** | **0.2255** |

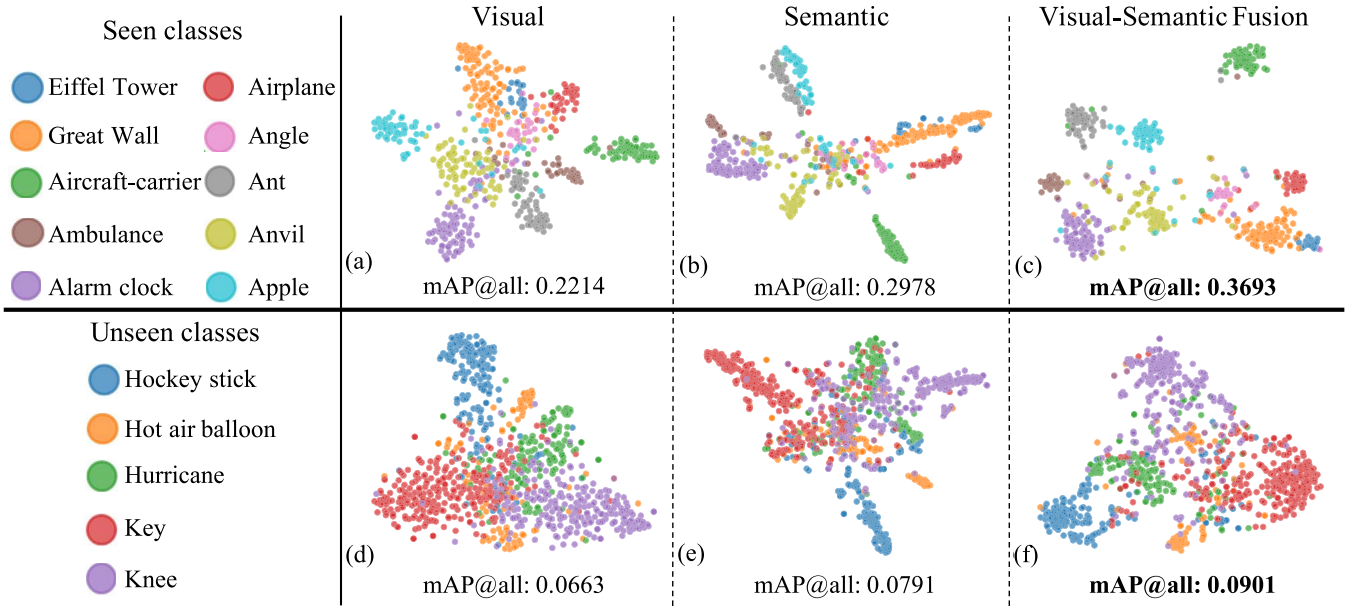



Fig. 3. Visualization of different features of the probe sketches for task Sk→Pa on DomainNet dataset. (a-c) show the "visual features", "semantic features", and "visual-semantic fusion" of selected 10 classes from *seen* categories, and (d-f) show 5 selected classes from *unseen* categories. All data are from the test sets and each color represents a particular category.

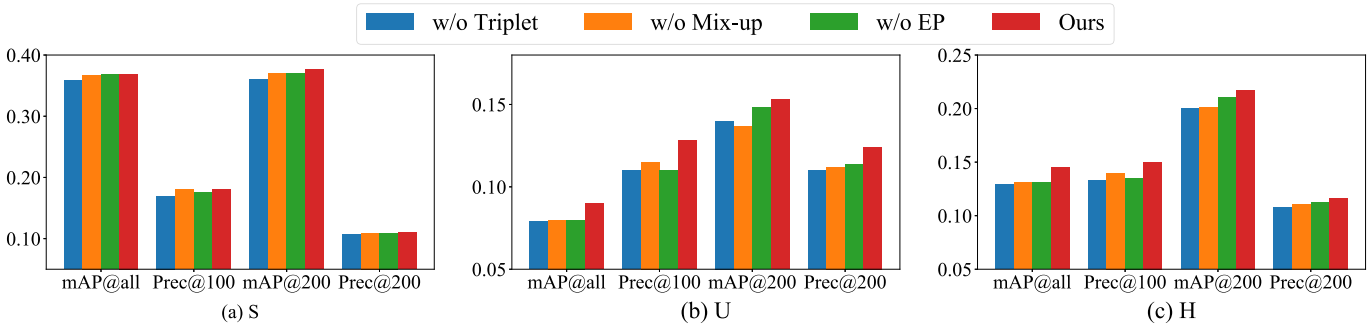



Fig. 4. Ablation analysis about each loss term contribution for task Sk→Pa on DomainNet dataset. S and U denote the performance on seen, unseen categories, respectively, and H is the harmonic mean of S and U.

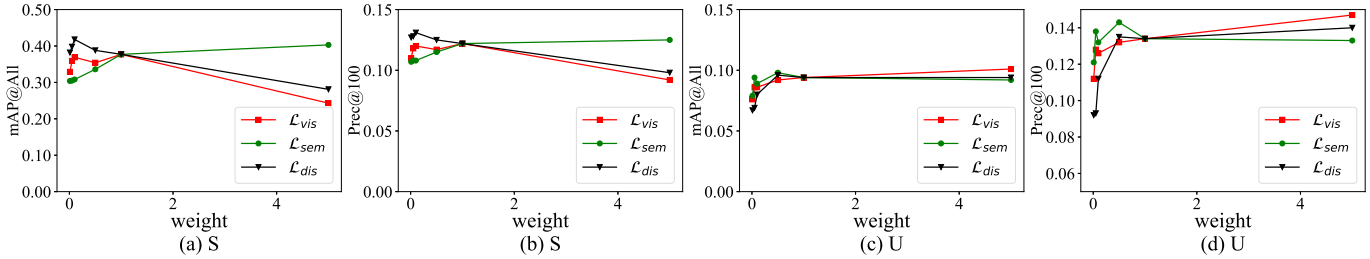



Fig. 5. Parameter analysis about the weights of loss terms of task Sk to Cl on DomainNet. (a)(b) show mAP@All and Prec@100 results of seen classes (S), and (c)(d) show mAP@All and Prec@100 results of unseen categories (U).

Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.

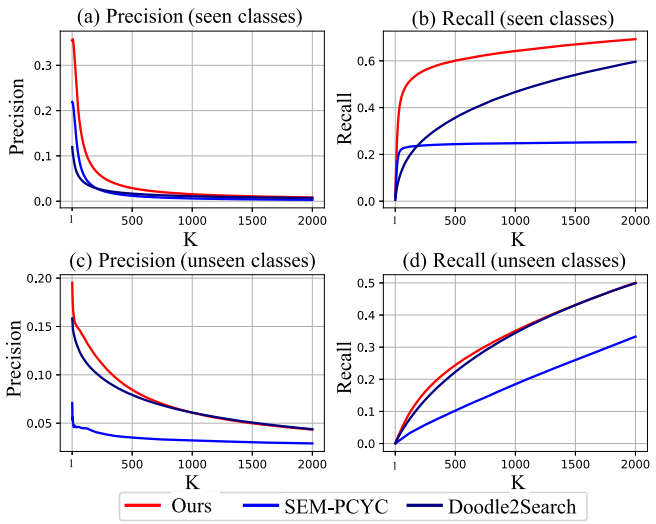



Fig. 6. Precision and Recall curves for task Sk→Cl on DomainNet.

complete model achieves the best performance for most tasks compared to all variants. Specifically, the visual-space domain alignment ("Triplet") performs crucial to all tasks, while the features refinement ("EP") and the cross-domain augmentation ("Mix-up") play important roles especially for the *unseen* categories. The observations of the significant contribution per each proposed module over novel categories demonstrate the contribution and effectiveness of the proposed model.

*4) Parameter Sensitivity Analysis:* In order to futher analyze the parameter sensitivity and the role of each loss term, we vary the weight for each loss term from 0.01 to 5, and report the results of task Sk → Cli on DomainNet dataset in Fig. 5. From the results, we observe that the best results for seen classes occur at around 0.1 for all three loss terms, and for the unseen classes, our model reach the best performance around 0.1. Specifically, for the unseen categories Prec@100, the performance is turning better, when we keep increasing the weights for $L_{vis}$ and $L_{dis}$.

*5) Precision-Recall Curve:* Moreover, to comprehensively evaluate the proposed model, we show the precision and recall curve of the results produced by our model and two compared baselines, SEM-PCYC and Doodle2Search, on the task Sk→Cl as Fig. 6. Specifically, figure (a-b) shows the precision and recall with different top-K retrieval samples for *seen* classes, respectively, and (c-d) are for *unseen* categories. From the results, we notice that the curves produced by our model is always higher than other methods, proving the effectiveness of this work to the compared baselines, which demonstrates the superiority of the model.

*6) Qualitative Results:* To demonstrate the effectiveness of the proposed model more intuitively, we show some retrieval results for different tasks on DomainNet in Fig. 7. Specifically, given one probe sample of Sketch/Quickdraw domain from seen/unseen categories, the top-7 retrieved images from 4 different retrieval galleries consisting of data constructed on different domains are displayed. All retrieved images are ordered based on the rank from left to right. From the results, we observe that our model achieves promising results for all 4 different retrieval galleries from different domains, no matter

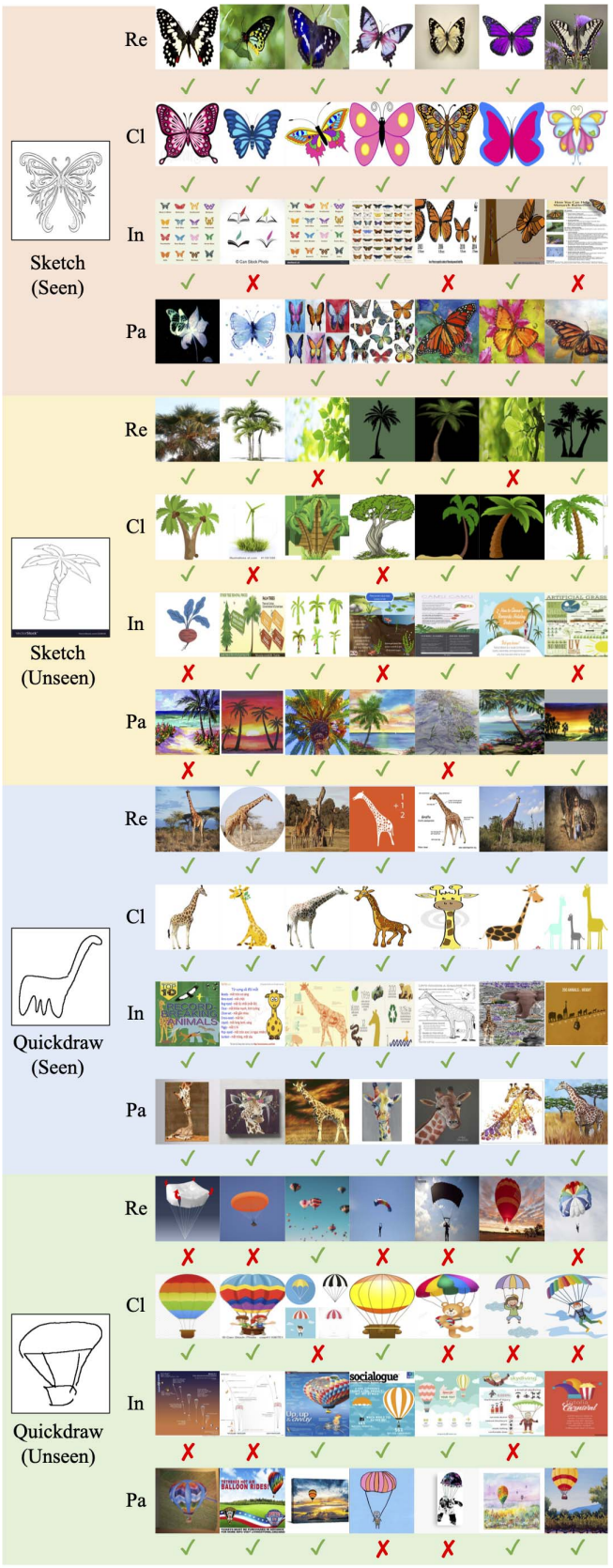



Fig. 7. Top-7 retrieved samples from various target domains given the same Sketch/Quickdraw sample from seen/unseen categories.

the query sample is from Sketch or Quickdraw domain for the seen categories. Moreover, the tasks on unseen classes are much harder, for most tasks, the majority of retrieved

Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.

samples by our model are correct. The most challenging tasks are Sk → In and Qu→ Re for *unseen* categories. The reason is that the Infograph domain has large domain distribution difference compared to the Sketch domain as images contain text and graphs irrelevant to the query object. In addition to the cross-domain distribution difference and the abstraction of *Quickdraw* domain samples, the large number of photos in the *Real* domain which are more complex in background and may contain more objects make the retrieval much more challenging. For these two tasks, the precision of top-7 results of our model are around 46.67% and 33.33%, respectively. From the retrieved samples, we notice some results are visually similar to the probes, although are annotated with "wrong" categories compared to the probe sketch. For example, the retrieved results for the sample "hot-air-balloon" contains several "parachute" images having similar shapes and colors.

## V. Conclusion

In this paper, we present a novel Augmented Multi-modality Fusion (AMF) framework to address the generalized zero-shot sketch-based image retrieval (GZS-SBIR) problem, where the probe sketches and retrieval gallery consist of samples from both seen and unseen categories. Specifically, both probe sketches and photos in the retrieval gallery are mapped into a domain-invariant visual space and a common semantic space. The distribution difference across-domains are alleviated in the visual space, and the visual-semantic fusion representations are constructed to leverage the multi-modality knowledge benefiting domain discrimination and task-specific classification. Moreover, cross-domain augmentation in both visual and semantic space synthesizes novel feature patterns to promote the generalizability of the model and boost the performance for novel categories. In addition to existing benchmarks, a new evaluation protocol on a large-scale dataset consisting of various domains is designed for GZS-SBIR. The superior experimental results on all evaluation benchmarks demonstrate our contributions.

## References


[1] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy, "Sketch me that shoe," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 799–807.

[2] F. Wang, L. Kang, and Y. Li, "Sketch-based 3D shape retrieval using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1875–1883.

[3] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao, "SketchNet: Sketch classification with web images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1105–1113.

[4] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2862–2871.

[5] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 102–112, Jan. 2019.

[6] Z. Gao, L. Wang, and L. Zhou, "A probabilistic approach to cross-region matching-based image retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1191–1204, Mar. 2019.

[7] Z. Li, C. Deng, E. Yang, and D. Tao, "Staged sketch-to-image synthesis via semi-supervised generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 2694–2705, 2021.

[8] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 3626–3637, 2020.

[9] H. Xia, T. Jing, C. Chen, and Z. Ding, "Semi-supervised domain adaptive retrieval via discriminative hashing learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3853–3861.

[10] Z. Wang *et al.*, "CAMP: Cross-modal adaptive message passing for text-image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5764–5773.

[11] Y. Chen, S. Gong, and L. Bazzani, "Image search with text feedback by visiolinguistic attention learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3001–3011.

[12] Y. Jing, W. Wang, L. Wang, and T. Tan, "Learning aligned image-text representations using graph attentive relational network," *IEEE Trans. Image Process.*, vol. 30, pp. 1840–1852, 2021.

[13] C. Deng, E. Yang, T. Liu, W. Liu, J. Li, and D. Tao, "Unsupervised semantic-preserving adversarial hashing for image search," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4032–4044, Aug. 2019.

[14] F. Huang, X. Zhang, Z. Zhao, and Z. Li, "Bi-directional spatial-semantic attention networks for image-text matching," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2008–2020, Apr. 2019.

[15] Y. Shen, L. Liu, F. Shen, and L. Shao, "Zero-shot sketch-image hashing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3598–3607.

[16] S. Antol, C. L. Zitnick, and D. Parikh, "Zero-shot learning via visual abstraction," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 401–416.

[17] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong, "Transductive multi-view embedding for zero-shot recognition and annotation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 584–599.

[18] M. R. Vyas, H. Venkateswara, and S. Panchanathan, "Leveraging seen and unseen semantic relationships for generative zero-shot learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 70–86.

[19] Y. Liu and T. Tuytelaars, "A deep multi-modal explanation model for zero-shot learning," *IEEE Trans. Image Process.*, vol. 29, pp. 4788–4803, 2020.

[20] R. Gao *et al.*, "Zero-VAE-GAN: Generating unseen features for generalized and transductive zero-shot learning," *IEEE Trans. Image Process.*, vol. 29, pp. 3665–3680, 2020.

[21] K. Wei, C. Deng, X. Yang, and D. Tao, "Incremental zero-shot learning," *IEEE Trans. Cybern.*, early access, Sep. 30, 2021, doi: 10.1109/TCYB.2021.3110369.

[22] S. Chen *et al.*, "FREE: Feature refinement for generalized zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 122–131.

[23] Z. Chen *et al.*, "Semantics disentangling for generalized zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 8712–8720.

[24] Z. Zhang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Zero-shot sketch-based image retrieval via graph convolution network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12943–12950.

[25] A. Dutta and Z. Akata, "Semantically tied paired cycle consistency for any-shot sketch-based image retrieval," *Int. J. Comput. Vis.*, vol. 128, no. 10, pp. 2684–2703, 2020.

[26] S. Dey, P. Riba, A. Dutta, J. Llados, and Y.-Z. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2179–2188.

[27] A. Pandey, A. Mishra, V. K. Verma, A. Mittal, and H. A. Murthy, "Stacked adversarial network for zero-shot sketch based image retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2540–2549.

[28] T. Dutta, A. Singh, and S. Biswas, "StyleGuide: Zero-shot sketch-based image retrieval using style-guided image generation," *IEEE Trans. Multimedia*, vol. 23, pp. 2833–2842, 2021.

[29] X. Xu, M. Yang, Y. Yang, and H. Wang, "Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 984–990.

[30] C. Deng, X. Xu, H. Wang, M. Yang, and D. Tao, "Progressive cross-modal semantic network for zero-shot sketch-based image retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 8892–8902, 2020.

[31] K. Lin, X. Xu, L. Gao, Z. Wang, and H. T. Shen, "Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11515–11522.

[32] W. Thong, P. Mettes, and C. G. M. Snoek, "Open cross-domain visual search," *Comput. Vis. Image Understand.*, vol. 200, Nov. 2020, Art. no. 103045.


Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.

[33] J. Zhu, X. Xu, F. Shen, R. K.-W. Lee, Z. Wang, and H. T. Shen, "Ocean: A dual learning approach for generalized zero-shot sketch-based image retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2020, pp. 1–6.

[34] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal, "A zero-shot framework for sketch based image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 300–317.

[35] Q. Liu, L. Xie, H. Wang, and A. Yuille, "Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3662–3671.

[36] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1406–1415.

[37] P. Rodríguez, I. Laradji, A. Drouin, and A. Lacoste, "Embedding propagation: Smoother manifold for few-shot classification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 121–138.

[38] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[39] M. Xu *et al.*, "Adversarial domain adaptation with domain mixup," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, pp. 6502–6509.

[40] T. Jing, B. Xu, and Z. Ding, "Towards fair knowledge transfer for imbalanced domain adaptation," *IEEE Trans. Image Process.*, vol. 30, pp. 8200–8211, 2021.

[41] T. Jing, H. Liu, and Z. Ding, "Towards novel target discovery through open-set domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9322–9331.

[42] T. Jing, H. Xia, and Z. Ding, "Adaptively-accumulated knowledge transfer for partial domain adaptation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1606–1614.

[43] T. Jing and Z. Ding, "Adversarial dual distinct classifiers for unsupervised domain adaptation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 605–614.

[44] M. F. Naeem, Y. Xian, F. Tombari, and Z. Akata, "Learning graph embeddings for compositional zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 953–962.

[45] M. Mancini, M. F. Naeem, Y. Xian, and Z. Akata, "Open world compositional zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5222–5230.

[46] D. Huynh and E. Elhamifar, "Compositional zero-shot learning via fine-grained dense feature composition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–12.

[47] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 23, 2022, doi: 10.1109/TKDE.2021.3130191.

[48] Y. Li, Q. Wang, J. Zhang, L. Hu, and W. Ouyang, "The theoretical research of generative adversarial networks: An overview," *Neurocomputing*, vol. 435, pp. 26–41, May 2021.

[49] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–12, Jul. 2016.

[50] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

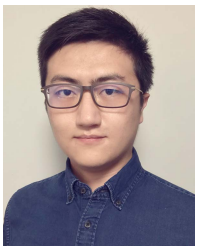
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[53] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, 2012.

[54] H. Wang, C. Deng, T. Liu, and D. Tao, "Transferable coupled network for zero-shot sketch-based image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 27, 2021, doi: 10.1109/TPAMI.2021.3123315.

[55] J. Bhatta, D. Shrestha, S. Nepal, S. Pandey, and S. Koirala, "Efficient estimation of nepali word representations in vector space," *J. Innov. Eng. Educ.*, vol. 3, no. 1, pp. 71–77, Mar. 2020.

[56] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.

**Taotao Jing** received the B.S. degree in electronic science and technology from Xi'an Jiaotong University, Xi'an, China, in 2016, and the M.S. degree in computer system engineering from Northeastern University, Boston, USA, in 2018. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Tulane University. His research interests lie in computer vision, transfer learning, and deep learning.

**Haifeng Xia** received the B.S. degree in information and computer science from Huazhong Agricultural University, Wuhan, China, in 2016, and the M.S. degree in computational mathematics from Sun Yat-sen University, Guangzhou, China, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Tulane University. His research interests mainly include computer vision and machine learning.

**Jihun Hamm** received the Ph.D. degree from the University of Pennsylvania in 2008, supervised by Dr. Daniel Lee. He has been an Associate Professor of computer science at Tulane University, since 2019. His research interest is in machine learning, from theory and to applications. He has worked on efficient algorithms for adversarial machine learning, deep learning, privacy and security, optimization, and nonlinear dimensionality reduction. He also has a background in biomedical engineering and has worked on medical data analysis, computational anatomy, and modeling human behaviors. His approach can be summarized as using machine learning to find novel solutions for challenging problems in the applied fields. His work in machine learning has been published in top venues such as ICML, NIPS, CVPR, *JMLR*, and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. His work has also been published in medical research venues, such as MICCAI, MedIA, and IEEE TRANSACTIONS ON MEDICAL IMAGING. The academic community has recognized his contributions, among other awards, he has earned the Best Paper Award from MedIA in 2010 and Google Faculty Research Award in 2015.

**Zhengming Ding** (Member, IEEE) received the B.Eng. degree in information security and the M.Eng. degree in computer software and theory from the University of Electronic Science and Technology of China (UESTC), China, in 2010 and 2013, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Northeastern University, USA, in 2018. He has been a Faculty Member affiliated with the Department of Computer Science, Tulane University, since 2021. Prior that, he was a Faculty Member affiliated with the Department of Computer, Information and Technology, Indiana University-Purdue University Indianapolis. His research interests include transfer learning, multiview learning, and deep learning. He is a member of ACM and AAAI. He received the National Institute of Justice Fellowship (2016–2018). He was a recipient of the Best Paper Award (SPIE 2016) and Best Paper Candidate (ACM MM 2017). He is currently an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *Journal of Electronic Imaging* (JEI), and *IET Image Processing*.

Authorized licensed use limited to: Anhui Normal University. Downloaded on October 26,2022 at 04:08:32 UTC from IEEE Xplore. Restrictions apply.