Zero-Shot Everything Sketch-Based Image Retrieval, and in Explainable Style

Fengyin Lin[1*] Mingkang Li[1*] Da Li[2†] Timothy Hospedales[2,3] Yi-Zhe Song[4] Yonggang Qi[1]
[1]Beijing University of Posts and Telecommunications [2]Samsung AI Centre, Cambridge
[3]University of Edinburgh [4]SketchX, CVSSP, University of Surrey
{fylin,lmk,qiyg}@bupt.edu.cn dali.academic@gmail.com t.hospedales@ed.ac.uk y.song@surrey.ac.uk

Abstract

This paper studies the problem of zero-short sketch-based image retrieval (ZS-SBIR), however with two significant differentiators to prior art (i) we tackle all variants (inter-category, intra-category, and cross datasets) of ZS-SBIR with just one network ("everything"), and (ii) we would really like to understand how this sketch-photo matching operates ("explainable"). Our key innovation lies with the realization that such a cross-modal matching problem could be reduced to comparisons of groups of key local patches – akin to the seasoned "bag-of-words" paradigm. Just with this change, we are able to achieve both of the aforementioned goals, with the added benefit of no longer requiring external semantic knowledge. Technically, ours is a transformer-based cross-modal network, with three novel components (i) a self-attention module with a learnable tokenizer to produce visual tokens that correspond to the most informative local regions, (ii) a cross-attention module to compute local correspondences between the visual tokens across two modalities, and finally (iii) a kernel-based relation network to assemble local putative matches and produce an overall similarity metric for a sketch-photo pair. Experiments show ours indeed delivers superior performances across all ZS-SBIR settings. The all important explainable goal is elegantly achieved by visualizing cross-modal token correspondences, and for the first time, via sketch to photo synthesis by universal replacement of all matched photo patches. Code and model are available at https://github.com/buptLinfy/ZSE-SBIR.

1. Introduction

Zero-shot sketch-based image retrieval (ZS-SBIR) is a central problem to sketch understanding [8, 16, 19, 20, 28, 34, 50, 54, 55, 57, 60, 67]. The zero-shot setting is largely driven by the prevailing data scarcity problem of human sketches [19, 28, 58] – they are much harder to acquire compared with photos. As research matures on the non zero-shot [4, 25, 33, 41–43, 59, 62], and in a push to make sketch-based retrieval commercially viable, recent research efforts had mainly focused on ZS-SBIR (or the simpler few-shot setting) [16, 19, 28, 34, 50, 57, 60].

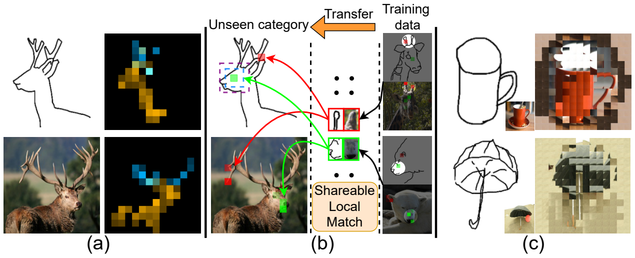

Figure 1. Attentive regions of self-/cross-attention and the learned visual correspondence for tackling *unseen* cases. (a) The proposed retrieval token [Ret] can attend to informative regions. Different colors are attention maps from different heads. (b) Cross-attention offers explainability by explicitly constructing local visual correspondence. The local matches learned from training data are shareable knowledge, which enables ZS-SBIR to work under diverse settings (inter- / intra-category and cross datasets) with just one model. (c) An input sketch can be transformed into its image by the learned correspondence, i.e., sketch patches are replaced by the closest image patches from the retrieved image.

Great strides have been made but attempts have largely aligned with the larger photo-based zero-shot literature, where the key lies in leveraging external knowledge for cross-category adaptation [19, 34]. That of conducting cross-modal matching is, however, less studied, and most prior art relies on a gold standard triplet loss with some auxiliary modules [16] to learn a joint embedding. Furthermore, as problems such as domain shift and fine-grained matching come to play, research efforts are mostly done in silo for different settings: category-level (standard) [28, 50, 60], fine-grained [2], and cross-dataset [40]. Last but definitely not least, one can not help but wonder why many of the proposed algorithm work – what is matched, and how is the transfer conducted?

This paper aims to tackle all said problems associated with the current status quo for ZS-SBIR. In particular, we advocate for (i) a single model to tackle all three settings of

[*] Equal contribution.
[†] Corresponding author.

ZS-SBIR, (ii) ditching the requirement on external knowledge to conduct category transfer, and more importantly, (iii) a way to explain why our model works (or not).

At the very core of our contribution lies with a well-explored insight that predates "deep vision", that image matching can be achieved by establishing local patch correspondences and computing a distance score based on that – yes, *loosely* similar to that of "bag of visual words" [11, 26, 51] (without building dictionaries). In our context of ZS-SBIR, we would like to conduct this matching (i) cross two diverse modalities in sketch and photo, and (ii) cross-category, granularity (fine-grained or not) and dataset (domain difference) boundaries. The biggest upside of this realization is that just as how "bag of visual words" is explainable, we can directly visualize the patch correspondences to achieve a similar level of explainability (see Figure 1).

Our solution first is a transformer-based cross-modal network, that (i) sources local patches independently in each modality, (ii) establishes patch-to-patch correspondences across two modalities, and (iii) computes matching scores based on putative correspondences. We put forward a novel design for each of the three components. We approach (i) by proposing a novel CNN-based learnable tokenizer, that is specifically tailored to sketch data. This is because the vanilla non-overlapping patch-wise tokenization proposed in ViT [18] is not friendly to the sparse nature of sketches (as most patches would belong to the uninformative blank). Our tokenizer on the other hand attends to a larger receptive field [37] hence more keen to sketch data. With this tokenizer, visual cues from nearby regions are aggregated when constructing visual tokens, so that structural information is preserved. In the same spirit of class token developed in ViT for image recognition, we introduce a learnable retrieval token to prioritize tokens for cross-modal matching.

To establish patch-to-patch correspondences, a novel cross-attention module is proposed that operates across sketch-photo modalities. Specifically, we propose cross-modal multi-head attention, in which the query embeddings are exchanged between sketch and photo branches to reason patch-level correspondences with only category-level supervision. With the putative matches in place, inspired by relation networks [53], we propose a kernel-based relation network to aggregate the correspondences and calculate a similarity score between each sketch-photo pair.

We achieve state-of-the-art performance across all said ZS-SBIR settings. Explainability is offered (i) as per tradition in terms of visualizing patch correspondences, where interesting local matches can be observed, such as the `antlers` of `deer` in Figure 1(b), regardless of a sketch being very abstract, and (ii) by replacing all patches in a sketch with their photo correspondences, to perform sketch to photo synthesis as shown in Figure 1(c).

2. Related Works

Zero-shot SBIR. Most previous works [16, 19, 28, 34, 50, 57, 60, 67] treat zero-shot SBIR (ZS-SBIR) as a category-level retrieval problem. Zero-shot learning (ZSL) algorithms [3, 7, 29, 66] typically play a central role in tackling it. The fundamental idea is to map sketch and photo into a shared semantic feature space to help alleviate the cross-domain gap. To assist knowledge transfer to match sketch and photo in unseen categories, side knowledge is normally required, such as text-based class descriptions. Essentially, these methods learn to associate sketch and photo to some class-specific feature representations for pairing. However, this matching is limited to a coarse-grained level. The visual similarities in fine details, i.e., visual local correspondences between sketch and photo, are largely ignored during training and inference. More recently, CC-DG [40] formulates ZS-SBIR as a cross-category generalization problem. In particular, CC-DG provides fine-grained SBIR by comparing sketch and photo conditioned on a visual trait that is dynamically selected from a trait bank learned from seen sketches. Although sharing the same goal of fine-grained retrieval, our approach differs from CC-DG that we explicitly learn local visual correspondence between sketch and photo, thereby offering distinct explainability.

Transformer-based cross attention. Although originating from natural language processing, the transformer [15, 56] has emerged as an effective base network for solving many vision tasks due to its powerful feature representation. Apart from applying self attention, reasoning based on cross attention has been shown effective on image classification, few-shot learning and sketch segment matching. CrossTransformer [17] finds pixel-level correspondence between a query image and a set of support images for few-shot image classification. The animation transformer (AnT) [6] learns segment-level correspondence between human-drawn animations for AI-assisted colorization. Built on vision transformer (ViT) [18], CrossViT [9] offers a dual-branch ViT which extracts multi-scale (small and large image patches) tokens and shares knowledge between two branches by exchanging the class (`CLS`) tokens, resulting an enhanced `CLS` token for classification. Despite sharing the same spirit, our model differs from CrossViT significantly in two ways: CrossViT performs within-branch attention over visual tokens. In contrast, ours carries out cross-branch interaction on visual tokens. Additionally, the output of CrossViT is an augmented `CLS` token, while ours leverages the visual tokens after cross-attention for local matching by the following token-level relation network.

Visual correspondence learning. Learning dense visual correspondence is an essential component and it has been actively explored in many vision tasks, such as structure-from-motion (SfM) [48], visual localization [38],

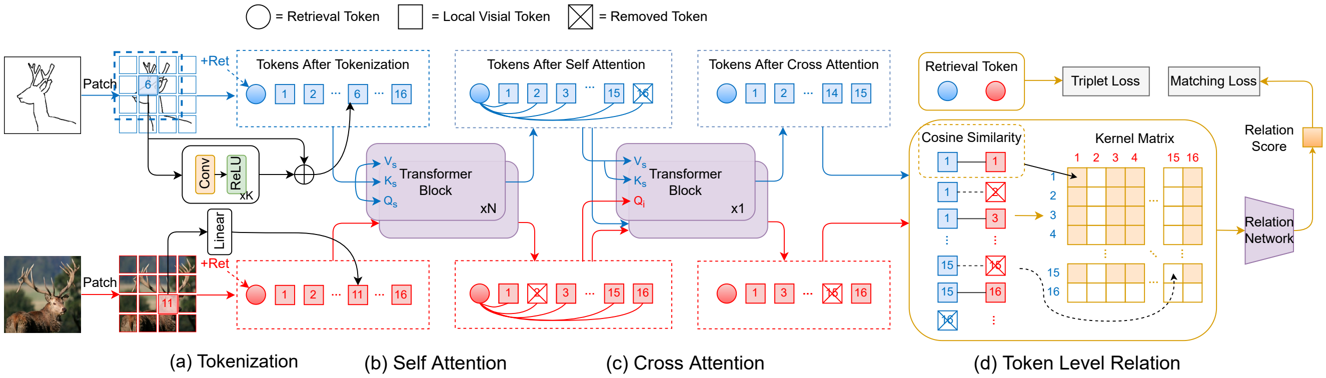

Figure 2. Network overview. (a) Learnable tokenization generates structure preserved tokens, preventing the generation of uninformative tokens. (b) Self-attention finds the most informative regions ready for local matching. (c) Cross-attention learns visual correspondence from visual tokens. A retrieval token [Ret] is added as a supervision signal during training. (d) Token-level relation network enables to *explicitly* measure the correspondences of cross-modal token pairs. Pairs of removed tokens as per token selection will not be counted.

simultaneous localization and mapping (SLAM) [12] and exemplar-based image-to-image translation [65]. Early approaches [22,36,39,61] focus on matching hand-crafted features for predicting correspondence between images. Coupled with deep features, plenty of improved results have been achieved. However, most works handle within-domain correspondence between natural images [1,10,23,24,27,30,35], or focus on dense pixel-wise correspondence learning across domains from strictly aligned cross-modal data (e.g., the edge map or semantic layout is paired with a sourced photo), and the training objectives are often highly complex (6 or more loss terms) that often requires dedicated engineering [63,65]. We instead tackle sketch to photo semantic correspondence learning at patch-level, to generalize image matching through local visual evidences.

3. Methodology

The overall scheme of our proposed framework is shown in Figure 2. Each key module is detailed in the following.

3.1. Learnable Tokenization

Given a query sketch $S \in \mathbb{R}^{h\times w\times c}$ and a gallery image $I \in \mathbb{R}^{h\times w\times c}$ to be matched, we can tokenize them into a sequence of visual tokens by using the same approach proposed in ViT [18] where images are evenly partitioned into non-overlapping patches, followed by a projection head mapping them into $S' \in \mathbb{R}^{n\times d}$ and $I' \in \mathbb{R}^{n\times d}$. However, we found that this tokenization is not friendly to sketches, which are typically composed of sparse strokes. To address this issue, we propose a learnable tokenizer, which transforms a given sketch $S \in \mathbb{R}^{h\times w\times c}$ into a sequence of visual embeddings $X \in \mathbb{R}^{n\times d}$. Specifically, the tokenizer is made up of a stack ($K = 4$) of convolution layers (Conv) with various kernel size, each followed by a non-linear activation (σ= ReLU): $X = [\sigma(\texttt{Conv}(\texttt{S}))]_{\times 4}$. Essentially, it can enlarge the receptive field when constructing visual tokens through hierarchical convolution, thereby better preserving structural cues from nearby regions. Additionally, a residual connection is introduced to rectify the vanilla tokens, then the final token embedding is: $X = X + S'$.

3.2. Self-attention with a Retrieval Token

Different from the vanilla ViT [18], we replace the vanilla class token with a retrieval token $[\texttt{Ret}]$ to facilitate our retrieval task by capturing a global representation of an image. Specifically, the retrieval token is initialized as a trainable d-dimensional token embedding $[\texttt{Ret}] \in \mathbb{R}^d$. During the model inference, all visual tokens including the retrieval token $[\texttt{Ret}]$ interact with each other through the multi-head self attention (MSA) modules, followed by MLP blocks, as per [18]. In formal, the overall forward is as

$$X_0 = [\texttt{Ret}, X^1, \dots, X^n], \tag{1}$$
$$X_l = \text{MSA}(\text{LN}(X_{l-1})) + X_{l-1}, \quad l = 1\dots L, \tag{2}$$
$$X_l = \text{MLP}(\text{LN}(X_l)) + X_l, \quad l = 1\dots L, \tag{3}$$

where residual connection is introduced in both Eq. 2 and 3. LN is layer norm and L is the number of layers. The same inference architecture is applied onto photo inputs. Specifically, the MSA module has three different projection heads $[W_q, W_k, W_v]$, which map the same token embeddings into Queries, Keys and Values. Formally, it is formulated as

$$Q = X^i \cdot W_q, \quad K = X^i \cdot W_k, \quad V = X^i \cdot W_v. \tag{4}$$

Then the scaled dot-product attention is given by

$$\texttt{s-attn}(Q, K, V) = \texttt{softmax}(\frac{QK^T}{\sqrt{d}})V. \tag{5}$$

3.3. Cross-modal Attention

The self-attention module learns an informative token-based representation of each image. To estimate local visual correspondences between sketch and photo tokens, we resort to cross attention. The idea is to find pair-wise connections between visual tokens from different modalities, i.e., sketch and photo. This can be achieved by swapping the sketch query Q_S and image query Q_I, resulting in the new Query, Key and Value tuples, i.e., (Q_S, K_I, V_I) and (Q_I, K_S, V_S). The cross modal attention is obtained by

$$\texttt{c-attn}(Q_I, K_S, V_S) = \texttt{softmax}(\frac{Q_I K_S^T}{\sqrt{d}})V_S. \tag{6}$$

In this way, sketch token embeddings are updated by the information from photo tokens. Photo token embeddings can be obtained in the same way.

3.4. Token Selection

Local visual tokens may represent background or meaningless regions which are unimportant to the retrieval. Thus, a token selection is applied to narrow the scope of attentive tokens, while this reduces computational complexity. Inspired by [31], attention scores between the retrieval token `[Ret]` and all visual tokens are leveraged as a token importance indicator. Formally, attention scores are computed as follows:

$$a = \texttt{softmax}(\frac{Q_{\texttt{[Ret]}} K^T}{\sqrt{d}}), \tag{7}$$

where $a \in \mathbb{R}^n$ and the i-th entry a^i denotes how much information the i-th token contributes to the retrieval token. Consequently, only the top-k visual tokens will be preserved according to the attention scores a, and the rest ones are discarded. In practice, token selection is performed at the 4-th, 7-th and 10-th layer in the self-attention, with keep rates r_S^{SA} and r_I^{SA} set for sketch and image respectively at the selected layers. Apart from self-attention, token selection can also be carried out during cross-attention, i.e., using the sketch retrieval token to select image visual tokens with keep rate r^{CA}, to prioritize tokens useful for retrieval.

3.5. A Kernel based Relation Network

Cosine kernel matrices generation. We further introduce a cosine kernel function after the cross attention module to explicitly measure the similarity between each pair of visual tokens. Specifically, given any pair of tokens across two modalities, i.e., X_S^i and X_I^j, the kernel matrices $M \in \mathbb{R}^{n \times n}$ is defined as

$$M_{i,j}^{S,I} = \frac{X_S^i \cdot {X_I^j}^T}{\|X_S^i\| \|X_I^j\|}. \tag{8}$$

This matrix M summarizes the cosine similarity between all pairs of sketch and photo tokens. Importantly, the formed kernel matrix M enables the explicit reasoning on token correspondences by a relation net which is described next.

Relation network. Inspired by Relation Network proposed in [53], we incorporate a relation network in our framework to estimate the matching score of a particular sketch-photo pair (S,I), based on their associated local correspondence kernel matrix $M^{S,I}$. Specifically, our relation network $R_\psi(\cdot)$ is a stack of two FC-ReLU-Dropout layers that can produce a relation score in the range of $(0, 1)$:

$$r(S, I) = \texttt{sigmoid}(R_\psi(M^{S,I})). \tag{9}$$

Unlike concatenating global image features in [53], our relation network conducts reasoning on local token similarities, thereby has the opportunity to learn which (set of) token correspondences (embedded in $M^{S,I}$) to prioritize during matching. In the end, retrieval can be performed by ranking gallery images according to their relation scores.

3.6. Losses

We exploit two losses to train our framework: A triplet loss applied on the `[Ret]` token, and a regression loss applied on our similarity score r. Given a triplet $< S_i, I_i^+, I_i^- >$, where S_i is an anchor sketch, I_i^+ is a photo with the same label to S_i while I_i^- from a different class, the triplet loss is minimized to align the positive pair $< S_i, I_i^+ >$, and push the anchor S_i away from the negative instance I_i^-. In our case, the retrieval token `[Ret]` is used as the global feature of sketches and photos, thus the triplet loss is defined as

$$\begin{aligned}\mathcal{L}_{tri} = & \frac{1}{T}\sum_{i=1}^{T} \max\{||\texttt{Ret}(S_i) - \texttt{Ret}(I_i^+)|| \\ & - ||\texttt{Ret}(S_i) - \texttt{Ret}(I_i^-)|| + m, \quad 0\}.\end{aligned} \tag{10}$$

T is the total number of triplets, and m is the margin.

In addition, a relation loss is used to measure whether a sketch-photo pair belongs to the same class/instance or not through our kernel based relation network as described in Section 3.5. Specifically, we regress the predicted relation score r to the ground-truth, i.e., $r = 1$ when matched, and $r = 0$ otherwise. Formally, the matching loss is defined as mean square error (MSE):

$$\mathcal{L}_{re} = \sum_{i=1}^{N}\sum_{j=1}^{H}(r_{i,j} - \mathbf{1}(y_i == y_j))^2, \tag{11}$$

N and H are the total numbers of query sketches and candidate photos, respectively. y is the class label. To this end, the overall loss is summed as

$$\mathcal{L} = \mathcal{L}_{tri} + \mathcal{L}_{re}. \tag{12}$$

3.7. Implementation Details

Sketch or image is scaled to 224 × 224. As stated in Section 3.1, there are four convolution layers in the tokenizer.

The kernel size of the first conv-layer is 7×7 and 3×3 for the rest. Stride is 2 for all conv-layers. Consequently, there are 196 visual tokens produced, each is represented by a $d = 768$ dimensional vector. The self-attention module is designed as per ViT [18] with 12 block and pre-trained on ImageNet-1K [14]. The cross-attention module is much lighter than the self-attention module and only contains one layer with 8 heads. AdamW is used with lr 10^{-5}.

4. Experiments

To verify the efficacy of our model, we first conduct experiments on category-level ZS-SBIR, followed by an ablation study of key components and explainability analysis to reveal why and how our approach works. Then, we conduct fine-grained SBIR experiments to verify the instance-level retrieval of our proposed model. Finally, cross-dataset ZS-SBIR experiments are conducted to verify the generalization of the learned visual correspondence across different datasets in a zero-shot setting.

4.1. Category-level ZS-SBIR

Datasets. There are three large-scale datasets that are commonly used for category-level ZS-SBIR: TU-Berlin Ext [64], Sketchy Ext [33] and QuickDraw Ext [16]. For TU-Berlin Ext, sketches are collected from TU-Berlin dataset [21] which contains 250 categories with 80 sketches per class, and 204,489 photos are sourced from ImageNet [14] and web images to pair with the sketches. Sketchy Ext is an extended version of Sketchy dataset [47], which consists of 125 categories (100 photos per class and 5-8 corresponding sketches per photo). In particular, an additional 60,502 photos are included, resulting in an enlarged photo gallery with 73,002 photos. QuickDraw Ext is the largest SBIR dataset which is composed of 330,000 sketches and 204,000 photos over 110 categories. Following [34], we split TU-Berlin Ext into 220/30 categories for training/test, and a partition of 100/25 training/test categories for Sketchy Ext. In addition, a split of 104/21 training/test classes for Sketchy Ext proposed in [28] is also used for evaluation when the testing classes are not presented in the ImageNet-1K classes. The default data split in [16] is applied to QuickDraw Ext, i.e., 80 classes for training and 30 for testing.

Competitors. We compare our method with several baselines, including ZSIH [50], CC-DG [40], DOODLE [16], SEM-PCYC [19], SAKE [34], SketchGCN [67], StyleGuide [20], PDFD [13], DSN [57], BDA-SketRet [8], SBTKNet [55], Sketch3T [44], TVT [54] and ViT-Ret/ViT-Vis [18] adapted by us. ViT-Ret means replacing the class token in ViT with a retrieval token used for matching; while ViT-Vis uses the visual tokens for matching. It should be noted that all the baselines, except CC-DG [40], StyleGuide [20] and ViT variants [18], employ external semantic information, whereas our method only relies on the learned visual correspondences between sketch-photo pairs. We also compare two variants of our model, i.e., Ours-Ret and Ours-RN, which retrieve images based on the retrieval token and the relation network separately.

Evaluation protocol. Mean average precision (mAP), precision on top 100 (Prec@100) and top 200 (Prec@200) are reported following the standard evaluation protocol.

Results. From Table 1, we can see that our proposed method achieves better or comparable results over other competitors, which is especially noteworthy since our method does not benefit from extra semantic information, i.e., text or class label. Meanwhile, we can see Ours-Ret achieves better results on TU-Berlin Ext and Sketchy Ext datasets, and works slightly worse than Ours-RN on Sketchy Ext [28] Split and QuickDraw Ext. Some qualitative results (including some failure cases) are illustrated in Figure 3. We can observe that, compared with SAKE and DSN, most of the top images retrieved using our approach more faithfully resemble to the query sketches in terms of the overall object pose and shape characteristics. And the false positives given by our method are somehow reasonable, as they are superficially similar to the query. We additionally compare ours with baselines for generalized ZS-SBIR. Results in Table 2 show that ours clearly outperforms others, suggesting a strong generalizability of our model.

Ablation study. An ablative study is conducted to examine the importance of each key component in our model. In particular, based on Ours-RN (Ours-full), we remove every individual ***component*** at a time with other parts remaining. Specifically, **w/o CA**: The cross-attention layers are removed, and the resulting visual tokens from the self-attention module are fed directly into the relation network. **w/o SA**: Self-attention layers are replaced by ResNet-50 to generate tokens. The sketch/photo feature map given by ResNet-50 is transformed into a sequence of feature vectors, serving as input tokens to the cross-attention module. **w/o cosine kernel (Cos-K)**: Instead of calculating the cosine kernel between the visual tokens, we concatenate each pair of them with a learned distance metric, which is inspired by [47], for matching. **w/o RN loss**: Relation loss in Eq. (11) is disabled during model training. **w/o `[Ret]`**: `[Ret]` and triplet loss are discarded, while only the visual tokens are used throughout the model training and testing. **w/o learnable tokenizer (L-Tok)**: The learned tokenzier is reverted back to the vanilla one, i.e., images are evenly divided into 16x16 patches. We can see from Table 3: (i) Without self-attention (SA) or cross-attention (CA), the performance drops dramatically, indicating the importance of each component. (ii) Using visual tokens only leads to some performance degradation on TU-Berlin Ext and Sketchy Ext, confirming the effect of using `[Ret]` token. (iii) Our proposed learnable tokenizer is clearly helpful on both datasets. (iv) Removing the cosine kernel is also harm-

Table 1. Category-level ZS-SBIR comparison results. "ESI" : External Semantic Information. "-" : not reported. The best and second best scores are color-coded in red and blue.

Method	ESI	$\mathbb{R}^D$	TU-Berlin Ext		Sketchy Ext		Sketchy Ext [28] Split		QuickDraw Ext	
			mAP	Prec@100	mAP	Prec@100	mAP@200	Prec@200	mAP	Prec@200
ZSIH [50]	✓	64	0.220	0.291	0.254	0.340	-	-	-	-
CC-DG [40]	✗	256	0.247	0.392	0.311	0.468	-	-	-	-
DOODLE [16]	✓	256	0.109	-	0.369	-	-	-	0.075	0.068
SEM-PCYC [19]	✓	64	0.297	0.426	0.349	0.463	-	-	-	-
SAKE [34]	✓	512	0.475	0.599	0.547	0.692	0.497	0.598	0.130	0.179
SketchGCN [67]	✓	300	0.324	0.505	0.382	0.538	-	-	-	-
StyleGuide [20]	✗	200	0.254	0.355	0.376	0.484	0.358	0.400	-	-
PDFD [13]	✓	512	0.483	0.600	0.661	0.781	-	-	-	-
ViT-Vis [18]	✗	512	0.360	0.503	0.410	0.569	0.403	0.512	0.101	0.113
ViT-Ret [18]	✗	512	0.438	0.578	0.483	0.637	0.416	0.522	0.115	0.127
DSN [57]	✓	512	0.484	0.591	0.583	0.704	-	-	-	-
BDA-SketRet [8]	✓	128	0.375	0.504	0.437	0.514	0.556	0.458	0.154	0.355
SBTKNet [55]	✓	512	0.480	0.608	0.553	0.698	0.502	0.596	-	-
Sketch3T [44]	✓	512	0.507	-	0.575	-	-	-	-	-
TVT [54]	✓	384	0.484	0.662	0.648	0.796	0.531	0.618	0.149	0.293
Ours-RN	✗	512	0.542	0.657	0.698	0.797	0.525	0.624	0.145	0.216
Ours-Ret	✗	512	0.569	0.637	0.736	0.808	0.504	0.602	0.142	0.202

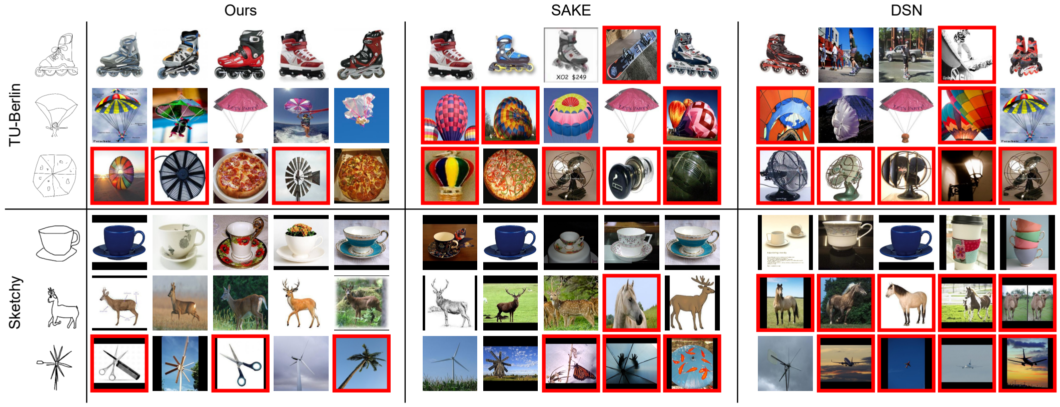

Figure 3. Exemplar comparison retrieval results for the given query sketches and the top 5 retrieved images. Red box denotes false positive.

Table 2. Generalized ZS-SBIR results.

Method	TU-Berlin Ext		Sketchy Ext	
	mAP	Prec@100	mAP	Prec@100
SEM-PCYC [19]	0.192	0.298	0.307	0.364
StyleGuide [20]	0.149	0.226	0.331	0.381
BDA-SketRet [8]	0.251	0.357	0.338	0.413
SBTKNet [55]	0.334	0.494	0.515	0.572
Ours-RN	0.432	0.460	0.634	0.651
Ours-Ret	0.464	0.485	0.656	0.670

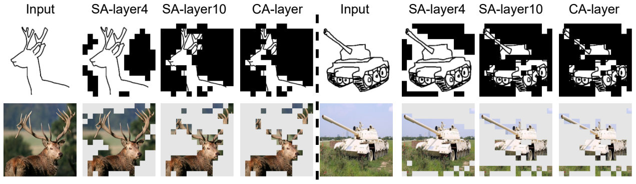

Figure 4. Visualization of token selection at different layers by setting keep rate for SA layers to 0.7 and the CA layer to 0.9.

ful, confirming its importance. ***Token selection***: Moreover, we employ a recent work [32] to study how the keep rate of tokens in our self-/cross-attention influences the final results. We can see that a lower keep rate will lead to a slight performance drop yet a significant speedup, i.e., speedier runtime per pair matching (RPM). Setting the token keep rate for both sketch and image to 0.7 is the best considering the speed gain versus the performance loss (Table 3 row in brown). Figure 4 shows visualization of token selection.

Self-attention map. To investigate what our network has learned from self-attention, we can get self-attention maps by using the retrieval token as query to measure its correlation to each visual token through vector dot-product, similar to [5]. As shown in Figure 5, we can observe that different heads can attend to different locations of an image, such as object foreground and different semantic parts, which are useful for fine-grained matching.

Cross-modal visual correspondence. To show our model's capacity of reasoning on salient local regions, we demonstrate some examples of cross-modal local visual correspondence in Figure 6. Specifically, to find estimated correspondences, we can simply measure the vector distance of each visual token pair according to Eq. 8. We can clearly see that visual correspondences in the retrieved images can be roughly located given a probe query sketch, despite the objects with different poses and backgrounds, thanks to good robustness of the cross-attention module.

Table 3. Ablation study results on manifesting importance of each key *component*, and using different *token selection rates*.

	Model	Keep Rate		TU-Berlin Ext		Sketchy Ext		RPM
		r_S^{SA}/r_I^{SA}	r^{CA}	mAP	Prec@100	mAP	Prec@100	(ms)
Components	w/o CA	-	-	0.294	0.352	0.295	0.346	-
	w/o SA	-	-	0.256	0.388	0.286	0.381	-
	w/o Cos-K	-	-	0.342	0.419	0.390	0.481	-
	w/o RN loss	-	-	0.497	0.610	0.656	0.744	-
	w/o [Ret]	-	-	0.519	0.623	0.681	0.767	-
	w/o L-Tok	-	-	0.514	0.621	0.672	0.767	-
	Ours-full	-/-	-	0.542	0.657	0.698	0.797	0.148
Token Selection	Ours-full	0.9/0.9	1.0	0.523	0.634	0.682	0.786	0.108
	Ours-full	0.7/0.7	1.0	0.509	0.619	0.671	0.778	0.056
	Ours-full	0.5/0.5	1.0	0.432	0.571	0.596	0.743	0.028
	Ours-full	0.7/0.9	1.0	0.519	0.628	0.678	0.782	0.082
	Ours-full	0.9/0.7	1.0	0.512	0.622	0.673	0.779	0.082
	Ours-full	0.7/0.7	0.9	0.510	0.618	0.668	0.774	0.055
	Ours-full	0.7/0.7	0.7	0.497	0.604	0.653	0.762	0.052

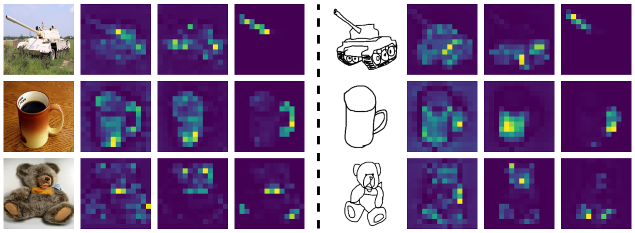

Figure 5. Attention maps of self-attention module on unseen categories. Given the tensors (heads) of the last layer of the self-attention module, we display the attention maps by using the retrieval token [Ret] as query. Original inputs are in the first column, followed by attention maps from multiple heads.

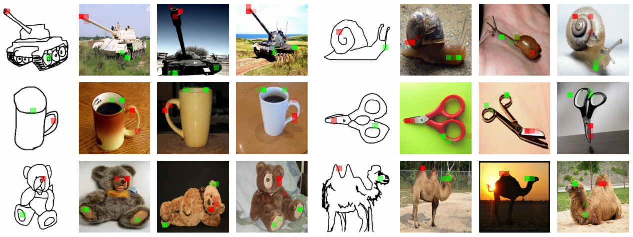

Figure 6. Visual correspondence across two modalities. Given a query sketch with two manually selected key regions (color-coded in red and green), we show the retrieved images with the corresponding matched regions (Top 3) in the same color.

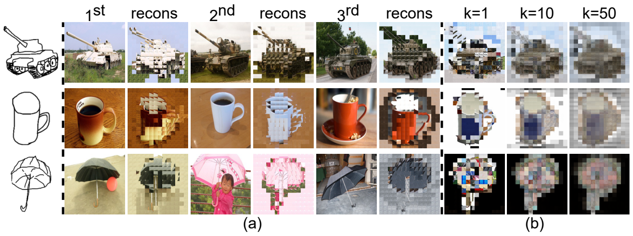

Figure 7. Cross-modal patch replacement. Given a sketch, (a) "recons" images are obtained by replacing sketch patches with the closest image patches of the top-3 retrieved images. (b) Reconstructed images using the k-nearest patches of the whole gallery.

Sketch-to-photo synthesis by patch replacement. To further inspect if reliable cross-modal correspondence has been discovered, we conduct 'sketch-to-photo synthesis' by replacing sketch patches with the closest image patches. The image patches are either found from (i) the retrieved closest image or (ii) the gallery contains all testing images of all categories in Sketchy Ext, i.e., 17,101 images with 3,351,796 patches in our case. Figure 7 (a) shows that the synthesized images can resemble the query sketch by stacking semantic closest patches from the retrieved natural images. Interestingly, irrelevant objects in the retrieved images are unselected when doing this sketch-to-photo synthesis, e.g., the tray under the red coffee cup, the balloon and the child under the umbrella. Moreover, besides using the patches from the retrieved images, sketch patches can be replaced by the patch averaged from the top k-nearest image patches from the gallery to investigate the learned correspondence more broadly. As shown in Figure 7 (b), the sketch patches can be replaced by quite reasonably similar image patches. Due to a much larger search space, the reconstructed images look scattered when $k = 1$. However, meaningful patch-level correspondence still holds, such as the searched umbrella handle and tank tracks patches. While the synthesized images get smoother and more similar content to the query sketches when k is increased. All these results demonstrate the capacity of our model for exploring the local visual correspondences between sketches and images.

How transfer happens? To gain some insights about how our model tackles unseen cases after learning local visual correspondence in training data, we take a sketch in test set as query to find the most similar image in training set, then further examine if there are common patterns of local visual matches in both training and testing sketch-photo pairs. As shown in Figure 8 (a), some shareable local matches do exist, such as the barrel and wheel of cannon and tank, suggesting the learned priors of local visual correspondence could be transferred to match regions of novel objects.

Most influential token pair? A key issue in explainability of AI systems is to be able to pinpoint the key features of the input that led to a particular decision [46]. Various off-the-shelf methods [49] exist for this in recognition systems, but it is trickier for retrieval systems as decisions operate on pairs of inputs. Our method provides the ability to answer such questions by identifying the most important feature pairing responsible for a match. Specifically, we remove one pair at a time, and return the pair that leads to maximum reduction of relation score. Some examples are demonstrated in Figure 8 (b). We can see that, for example, the most influential token is the antler of deer, which led a maximum reduction about 17% of matching score.

Computational cost analysis. We compare the GFLOPs, model size and runtime per pair of sketch-image matching (RPM) between ours and two SOTA methods, i.e., SAKE [34] and SEM-PCYC [57]. From Table 4, we can see our model has fewer parameters than SEM-PCYC. In terms of GFLOPs and RPM, it is dominated by the self-

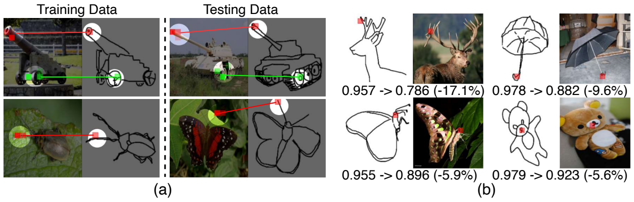

Figure 8. (a) Example of shareable local matches. The observed visual correspondences in training data show up again in testing data. (b) Example of most important token pair (red) which led maximum reduction of the matching score. Zoom in for best view.

Table 4. Comparison of computational cost.

	SAKE [34]	SEM-PCYC [19]	Ours-RN (SA+CA)	Ours⋆
# Params (M)	27.6	137.9	102.2(87.8+14.4)	102.2
GFLOPs	3.90	15.5	19.5 (17.8+1.7)	12.6 (12.0+0.6)
RPM (ms)	0.138	0.070	0.148 (0.118+0.030)	0.056 (0.048+0.008)

Table 5. Zero-shot FG-SBIR results (%). Note that all competitors are *not* zero-shot models, they are trained on Chair-V2.

Method	TripLet-SAN [62]	DSA [52]	TripLet-RL [2]
acc.@1	47.65	53.41	56.54
acc.@10	84.24	87.56	89.61
Method	StyleMeUp [45]	CC-DG [40]	Ours-RN/Ours-Ret
acc.@1	62.86	54.21	63.34/**64.31**
acc.@10	91.14	88.23	**94.53**/92.60

attention module, i.e. the core module of the used ViT backbone. However, the employed token selection can reduce the GFLOPs and RPM significantly while still deliver the SOTA performance, i.e., Ours⋆ in Table 4 which is the same model variant color-coded in brown in Table 3. More importantly, it is worth noting that our proposed CA module causes only a modest cost (8.7% of GFLOPs and 14.1% of parameters) to achieve a significant performance gain, C.F. w/o CA vs Ours-full shown in Table 3.

4.2. Zero-shot Fine-grained SBIR

Datasets. QMUL-Shoe-V2 and QMUL-Chair-V2 [62] are two commonly used benchmarks for FG-SBIR. QMUL-Shoe-V2 is composed of 6,730 shoe sketches with 2,000 associated photos, and QMUL-Chair-V2 contains 2,000 chair sketches with the corresponding 400 photos. We train our model on QMUL-Shoe-V2 then test on QMUL-Chair-V2 to conduct ZS FG-SBIR experiments.

Evaluation protocol. We follow the evaluation protocol in [45], i.e., metrics of top 1 and top 10 retrieval accuracy. I.e., credit will be given if the true positive (photo) to the query (sketch) is ranked within the top 1/10 slots.

Competitors. Five strong FG-SBIR baselines are compared, including TripLet-SAN [62], DSA [52], TripLet-RL [2], CC-DG [40], and StyleMeUp [45].

Results. From Table 5, we can see that our approach surprisingly surpasses all baselines even under an *unfair* comparison, i.e., ours is tested in a zero-shot setting, whereas all competitors are trained on the target category.

Table 6. Cross-dataset ZS-SBIR results. "S", "T" and "Q" denote Sketchy Ext, TU-Berlin Ext, and QuickDraw Ext, respectively. "(·)" denotes the number of test categories which are unseen to ensure the zero-shot setting. E.g., S→T(21) denotes that, we train on the training split of Sketchy Ext, then test on a subset (21 unseen classes) of the testing split of TU-Berlin Ext. Rows with a grey background indicate using ViT backbone for fair comparisons.

Method	**S→ T (21)**		**S→ Q (11)**		**T→ S (8)**		**T→ Q (10)**	
	mAP	Prec@100	mAP	Prec@100	mAP	Prec@100	mAP	Prec@100
CC-DG [40]	0.252	0.403	0.148	0.212	0.570	0.660	0.214	0.278
	0.308	0.434	0.156	0.227	0.624	0.693	0.231	0.296
DSN [57]	0.384	0.480	0.152	0.171	0.646	0.673	0.229	0.251
	0.356	0.469	0.149	0.178	0.613	0.654	0.218	0.246
SAKE [34]	0.421	0.549	0.183	0.250	0.657	0.722	0.248	0.340
	0.389	0.506	0.174	0.242	0.626	0.701	0.235	0.318
Ours-RN	**0.476**	**0.590**	**0.228**	**0.338**	**0.746**	**0.816**	**0.273**	**0.376**

4.3. Cross-Dataset category-level ZS-SBIR

Settings. We finally verify the ability of our method to generalize across completely different datasets in zero-shot scenario, i.e., the model is trained on dataset A then tested on dataset B, where the test classes are all unseen during training. This setting goes beyond the standard within-dataset ZS-SBIR benchmarks to evaluate transfer across entirely different benchmarks. Such cross-dataset ZS-SBIR is an even more challenging and realistic setting, since sketches from different datasets are typically drawn in diverse styles, as well as containing disjoint classes.

Results. As shown in Table 6, our method works much better than other competitors, demonstrating the preferable generalization ability of our learned visual correspondences over both the dataset and task shifts.

5. Conclusion and Future Work

We tackled ZS-SBIR, with a hope to also make it explainable. We are inspired by "old vision", and put forward a patch matching framework, that is not only explainable but also able to tackle all ZS-SBIR settings at the same time. The technical solution is a transformer-based cross-modal network, with three specific designs to tailor to the problem: (i) a self-attention module to learn the tokens, (ii) a cross-attention module to establish putative matches, and (iii) a kernel-based relation network to aggregate local matches into an overall similarity score. Last but not least, visualizations on patch-level correspondences, and sketch-to-photo synthesis through cross-modal patch replacement, provide means of explanation.

We can see the false matchings in Figure 3 are caused by the high similarity either between local tokens, e.g., windmill's paddles v.s. scissors' blades or global shapes, e.g., pizza v.s. fan. Therefore, how to take the best of global and local correspondence is worth exploring. Moreover, our patch replacement-based photo synthesis is rather coarse, thus improving synthesis quality will be a future endeavour.

Acknowledgement This work was supported by NSFC under No.61601042 and the Program for Youth Innovative Research Team of BUPT under No. 2023QNTD02.

References

[1] Kfir Aberman, Jing Liao, Mingyi Shi, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Neural best-buddies: Sparse cross-domain correspondence. *ACM TOG*, 2018. 3

[2] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *CVPR*, 2020. 1, 8

[3] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *ECCV*, 2016. 2

[4] Yang Cao, Changhu Wang, Liqing Zhang, and Lei Zhang. Edgel index for large-scale sketch-based image search. In *CVPR*, 2011. 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 6

[6] Evan Casey, Victor Perez, and Zhuoru Li. The animation transformer: Visual correspondence via segment matching. In *ICCV*, 2021. 2

[7] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. 2

[8] Ushasi Chaudhuri, Ruchika Chavan, Biplab Banerjee, Anjan Dutta, and Zeynep Akata. Bda-sketret: Bi-level domain adaptation for zero-shot sbir. *arXiv preprint arXiv:2201.06570*, 2022. 1, 5, 6

[9] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *ICCV*, 2021. 2

[10] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *NeurIPS*, 2016. 3

[11] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, 2004. 2

[12] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE TPAMI*, 2007. 3

[13] Cheng Deng, Xinxun Xu, Hao Wang, Muli Yang, and Dacheng Tao. Progressive cross-modal semantic network for zero-shot sketch-based image retrieval. *IEEE TIP*, 2020. 5, 6

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[16] 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. 1, 2, 5, 6

[17] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. In *NeurIPS*, 2020. 2

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3, 5, 6

[19] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019. 1, 2, 5, 6, 8

[20] Titir Dutta, Anurag Singh, and Soma Biswas. Styleguide: zero-shot sketch-based image retrieval using style-guided image generation. *IEEE TMM*, 2020. 1, 5, 6

[21] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM TOG*, 2012. 5

[22] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981. 3

[23] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE TPAMI*, 2017. 3

[24] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *ICCV*, 2017. 3

[25] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *CVIU*, 2013. 1

[26] Hiroharu Kato and Tatsuya Harada. Image reconstruction from bag-of-visual-words. In *CVPR*, 2014. 2

[27] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *CVPR*, 2017. 3

[28] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 1, 2, 5, 6

[29] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, 2017. 2

[30] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *CVPR*, 2019. 3

[31] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *ICLR*, 2022. 4

[32] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022. 6

[33] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017. 1, 5

[34] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV*, 2019. 1, 2, 5, 6, 7, 8

[35] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? *NeurIPS*, 2014. 3

[36] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 3

[37] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NeurIPS*, 2016. 2

[38] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 2015. 2

[39] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. In *NeurIPS*, 2018. 3

[40] 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. 1, 2, 5, 6, 8

[41] Sarthak Parui and Anurag Mittal. Similarity-invariant sketch-based image retrieval in large databases. In *ECCV*, 2014. 1

[42] Jose M Saavedra. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In *ICIP*, 2014. 1

[43] Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based image retrieval using learned keyshapes (lks). In *BMVC*, 2015. 1

[44] 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. 5, 6

[45] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*, 2021. 8

[46] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019. 7

[47] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 2016. 5

[48] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2

[49] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 7

[50] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018. 1, 2, 5, 6

[51] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 2

[52] 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. 8

[53] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018. 2, 4

[54] 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. 1, 5, 6

[55] Osman Tursun, Simon Denman, Sridha Sridharan, Ethan Goan, and Clinton Fookes. An efficient framework for zero-shot sketch-based image retrieval. *Pattern Recognition*, 2022. 1, 5, 6

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[57] Zhipeng Wang, Hao Wang, Jiexi Yan, Aming Wu, and Cheng Deng. Domain-smoothing network for zero-shot sketch-based image retrieval. *IJCAI*, 2021. 1, 2, 5, 6, 7, 8

[58] Peng Xu, Timothy M Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep learning for free-hand sketch: A survey. *IEEE TPAMI*, 2022. 1

[59] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, Zhanyu Ma, and Jun Guo. Sketchmate: Deep hashing for million-scale human sketch retrieval. In *CVPR*, 2018. 1

[60] Yang Yang, Yadan Luo, Weilun Chen, Fumin Shen, Jie Shao, and Heng Tao Shen. Zero-shot hashing via transferring supervised knowledge. In *ACM MM*, 2016. 1, 2

[61] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *ECCV*, 2016. 3

[62] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016. 1, 8

[63] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, and Changgong Zhang. Marginal contrastive correspondence for guided image generation. In *CVPR*, 2022. 3

[64] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *CVPR*, 2016. 5

[65] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*, 2020. 3

[66] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015. 2

[67] Zhaolong Zhang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. Zero-shot sketch-based image retrieval via graph convolution network. In *AAAI*, 2020. 1, 2, 5, 6