You'll Never Walk Alone: A Sketch and Text Duet for Fine-Grained Image Retrieval

Subhadeep Koley[1,2] Ayan Kumar Bhunia[1] Aneeshan Sain[1] Pinaki Nath Chowdhury[1]
Tao Xiang[1,2] Yi-Zhe Song[1,2]
[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{s.koley, a.bhunia, a.sain, p.chowdhury, t.xiang, y.song}@surrey.ac.uk

https://subhadeepkoley.github.io/Sketch2Word

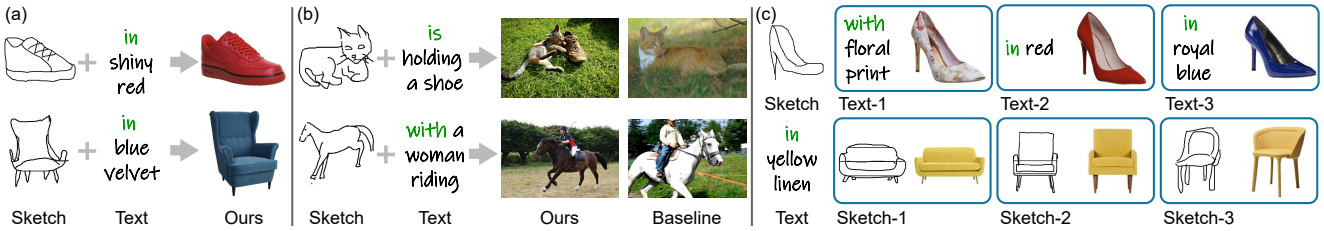

Figure 1. (a) Photos retrieved by our method, depicting precise control over *both* shape and appearance. (b) Unlike baseline sketch+text composed retrieval framework, our method seamlessly composes the *structural* and *contextual* cues of sketch and text queries respectively. (c) With a *fixed* sketch query, our method retrieves *different* images for *different* textual descriptions and vice-versa, depicting the *complementarity* of sketch and text modalities in sketch+text-based composed image retrieval. For a fixed sketch, the visual *attributes* from different textual descriptions are visibly reflected in the retrieved images while maintaining shape consistency. Similarly, fixing the attributes provided via text, shapes of retrieved images change corresponding to different sketch queries.

Abstract

Two primary input modalities prevail in image retrieval: sketch and text. While text is widely used for inter-category retrieval tasks, sketches have been established as the sole preferred modality for fine-grained image retrieval due to their ability to capture intricate visual details. In this paper, we question the reliance on sketches alone for fine-grained image retrieval by simultaneously exploring the fine-grained representation capabilities of both sketch and text, orchestrating a duet between the two. The end result enables precise retrievals previously unattainable, allowing users to pose ever-finer queries and incorporate attributes like colour and contextual cues from text. For this purpose, we introduce a novel compositionality framework, effectively combining sketches and text using pre-trained CLIP models, while eliminating the need for extensive fine-grained textual descriptions. Last but not least, our system extends to novel applications in composed image retrieval, domain attribute transfer, and fine-grained generation, providing solutions for various real-world scenarios.

1. Introduction

Sketch and text represent the two most common [11, 59] input modalities in the realm of image retrieval. The choice between these modalities depends on the nature of the retrieval problem, especially when fine-grained distinctions are required [18, 59, 60, 69]. In inter-category retrieval, text dominates as the primary modality, exemplified by widely-used platforms like Google Images. However, when the challenge transitions to fine-grained image retrieval, sketches take the spotlight [11, 59, 60]. Sketches promise to capture fine-grained visual cues that can be cumbersome or even impossible for text to express [11]. Research in this domain predominantly revolves around harnessing the unique qualities of sketches, exploring aspects such as style [54], abstraction [32], and more [4, 18, 59].

In this paper, we question this notion that "sketch is everything" and, for the first time, simultaneously delve into the fine-grained representation capabilities of both sketch and text, and in turn orchestrate a duet between these two modalities for fine-grained image retrieval. The outcome is a novel retrieval experience where a sketch and text work in harmony, enabling users to achieve precise retrievals that were previously unattainable. Now, users can locate "that" specific shoe, considering not only fine-grained pattern cues from sketches but also incorporating attributes like colour and texture from text (Fig. 1a). This synergy extends to scenarios where text offers contextual cues to a given sketch, such as a cat holding a shoe for that matter (Fig. 1b)!

While sketch and text synergy has been studied before [11, 59], it has predominantly focused on *scene-level/category* retrieval, where paired sketch and text descriptions for a given scene/category are readily available in datasets (*e.g.*, FS-COCO [10], CM-Places [8]). Our contention is that the synergy between sketches and text, while notable, is not as pronounced in category-level retrieval as it is for *fine-grained* retrieval [8, 11]. Indeed, for category-level retrieval, one might argue the necessity of sketches, as the descriptive power of text might already suffice [59].

However, when the desire extends to a horse/cat with a specific pose, an accompanying sketch becomes indispensable (as illustrated in Fig. 1b).

Our foremost challenge centres around fine-grained compositionality, specifically investigating how sketches and text can serve as complementary components in fine-grained queries. Our goal is to maintain the semantics of both modalities, ensuring, for example, that a horse in Fig. 1b corresponds precisely to the respective sketch and associated text, rather than any generic horse with a woman riding. To tackle this challenge, we harness the capabilities of CLIP [51], leveraging its implicit grammatical-composition capability. We achieve this via CLIP [51] to create a *fine-grained textual equivalent* of the input sketch, referred to as a "pseudo-word token". This token, when combined with text input, forms a fine-grained textual query that seamlessly integrates both sketch and text features, allowing them to work in synergy within the text domain.

The second challenge pertains to alleviating the requirement of collecting a dataset of fine-grained sketch and text pairs. We also aim to emulate the inference-time distribution of text input. The key innovation lies in the hypothesis that the fine-grained description embedded in a photo (P) can be approximated by that of a sketch (S) plus text (T), leading to T = P - S. This relationship illustrates how the absence of text can be approximated by the difference signal between the photo and the sketch in the latent embedding space. We incorporate this difference signal as a *proxy* for the missing textual description during training to make it inference-time equivalent through a novel *compositionality constraint*. Furthermore, we utilise short phrases generated by a lightweight GPT [6] as a *neutral-text regulariser* to ensure that the synergy works without disrupting the grammatical structure of CLIP's [51] language manifold.

Last but not least, in addition to addressing the challenges in fine-grained image retrieval, our system opens the door to a range of novel applications in the field of composed image retrieval such as object-sketch-based scene image retrieval, domain attribute transfer, and sketch+text-based fine-grained image generation.

In summary, *(i)* we address the challenge of fine-grained image retrieval by leveraging the synergy between freehand sketches and textual descriptions, extending retrieval beyond traditional category-level distinctions. *(ii)* we introduce a novel compositionality framework, effectively combining sketches and text using pre-trained CLIP models, eliminating the need for extensive fine-grained textual descriptions. *(iii)* our system unlocks novel applications like object-sketch-based scene retrieval, domain attribute transfer, and sketch+text-based fine-grained image generation.

2. Related Works

Sketch-Based Image Retrieval (SBIR). Starting at category-level, SBIR is tasked to fetch a photo of the same category as that of a given query sketch. Earlier deep-learning methods [15, 39, 68, 71] generally train Siamese-like networks [15] over a distance-metric in a cross-modal joint embedding space [14]. Moving forward to *fine-grained* SBIR (FG-SBIR), the aim is to retrieve one particular *photo-instance* from a gallery of same-category photos corresponding to a query sketch. FG-SBIR has progressed from a deep-triplet ranking-based Siamese network [70] to further enhancements involving higher-order attention [61] or auxiliary losses [37], and local feature alignment [67], to name a few. While most works focus on various applications like early-retrieval [3], cross-category generalisation [5, 49] and even zero-shot FG-SBIR [33, 55], others delved deeper to explore sketch-specific traits, like hierarchy [53], or style-diversity [54], for better retrieval. Recent extensions include retrieving a scene image based on a scene-sketch (*i.e.*, Scene-level FG-SBIR), employing cross-modal region associativity [9], enhanced further with text-query [11]. Unlike existing *fully-supervised* sketch+text-based methods [11, 59], relying on *paired* training triplets (*i.e.*, sketch, text, and ground truth photo), our approach alleviates the need for such triplets, simplifying the challenge of collecting fine-grained sketch-text-photo dataset.

Text-Based Image Retrieval (TBIR). Over the years, much emphasis has been paid to textual query-based image retrieval by learning a joint embedding space via ranking loss [19, 28, 50]. This was further augmented by cross-modal message passing [64], hard triplet mining [20], and one-to-many probabilistic mapping [12], to name a few. Thanks to internet-scale paired image-text datasets TBIR has become highly competitive and one of the most active areas of research [51], leading to expansive techniques like Oscar [36], CLIP [51], ALIGN [27], etc. The recent paradigm of *Textual Inversion* [13, 21] deals with inverting input image(s) into a pseudo-word token in the language space of pre-trained vision-language models for downstream tasks like personalised image retrieval [13], composed retrieval [2, 57], etc. Composed image retrieval (CIR) aims to retrieve images from a combined query of text and image pairs [1]. Existing CIR methods typically leverage pre-trained CLIP [1, 2, 57], or resort to image-text feature fusion [23, 25, 62, 66]. Textual inversion-based CIR methods [2, 57] either use million-scale image dataset to train inversion networks [57], or uses time-consuming two-stage optimisation-based approach [2]. Nevertheless, these *image*+text composed retrieval frameworks [2, 57] can not handle the huge domain gap of *sparse sketch*+text composed image retrieval. Our method, on the other hand, explicitly models the sketch-photo *difference* within itself for fine-grained sketch+text composed multi-modal retrieval.

Sketch+Text Joint Multi-Modal Learning. While sketch is a suitable [10] fine-grained query medium for AI systems, certain aspects fall outside its scope like qualita-

tive attributes (*e.g.*, colour, shade, etc.) [11]. Text aptly describes these qualitative attributes (*e.g.*, colour) which leads to "text as query" being extensively studied [11], albeit largely for category-level tasks (*e.g.*, TBIR). Realising this potential of sketch-text-photo association, joint multi-modal sketch+text query was heavily used in downstream vision tasks. On generation, sketch-to-image models used text-as-guidance to specify class-label [24]. Simultaneously text-to-image diffusion models used sketches as semantic-guidance for the encoder [44], or decoder [72], or external classifier guidance [63], NeRF-editing [43, 45] using sketch+text conditioning [42], and sketch-conditioned image captioning [11]. On retrieval, the importance of sketch+text was realised with FS-COCO [10] dataset collecting scene-level sketch-text-photo dataset. Recent works include supervised approaches of Sangkloy *et al*. [59] using CLIP, and Scenetrilogy [11] using conditional invertible neural networks for sketch+text-based image retrieval. Despite its benefit, collecting fine-grained textual descriptions, is quite cumbersome [10], which bottlenecks further exploration of fine-grained sketch-text-photo association for downstream tasks till date.

3. Revisiting CLIP

CLIP [51] consists of a text encoder and an image encoder, trained with a multi-class N-pair contrastive loss [51] on internet-scale ($\sim 400M$) image-text pair dataset, with an aim to learn a joint-embedding space that minimises the cosine similarity between the matching image-text pairs while maximising the same for random unpaired ones [51]. The image encoder ($\mathbf{V}$) usually employs a Vision Transformer (ViT) [17], to encode an input image $\mathcal{I}$ into a visual feature as $\mathbf{i} = \mathbf{V}(\mathcal{I}) \in \mathbb{R}^d$. The text encoder ($\mathbf{T}$) inputs a sequence of words $\mathcal{W} = \{w^0, w^1, \dots, w^k\}$ and applies a lower-cased byte pair encoding (BPE), followed by a learnable word embedding layer $\mathbf{T_w}$ ($49,152$ vocab size) [51], to convert each word w^i into a *word token embedding* $\mathbf{w}_e^i = \mathbf{T_w}(w^i)$ of size $\mathbb{R}^d$. This sequence of word token embeddings $\mathcal{W}_e = \{\mathbf{w}_e^0, \mathbf{w}_e^1, \dots, \mathbf{w}_e^k\}$ is then passed via a transformer $\mathbf{T_t}$, to provide the final textual feature as $\mathbf{w} = \mathbf{T_t}(\mathcal{W}_e) \in \mathbb{R}^d$, taken from the last hidden state of the final transformer layer.

4. Sketch-Based Composed Image Retrieval

Motivation. Combining structural cues from *sketch* with additional *textual description* results in a powerful query for image retrieval. Existing works [11, 59] on such compositionality usually extract sketch and text features via separate encoders, and add or concatenate (followed by additional learnable layers) them, to obtain the composed query feature. This has two major issues: *(i)* it needs sketch-*associated* textual description – absent from fine-grained SBIR datasets [58, 70], and *(ii)* combining the two features naively may distort the optimal sketch-text feature correlation needed, to correctly represent a composed semantic. This *compositionality* is however more explicit in the textual domain [2, 13, 57] where combining individual words/phrases form a composed semantic, *e.g.*, 'a cat' and 'brown' together infers *'a brown cat'*. Following CLIP's rise in various downstream tasks [30, 55, 73], we thus aim to leverage its text encoder's input text space to tackle sketch+text compositionality. In particular, inspired by textual inversion literature [21], we aim to represent a sketch as a *pseudo-word token* that emulates its visual concept in equivalent word-embedding space, and combine its textual description via connecting phrases like 'with', 'in', 'and', etc. (the full list in § Suppl.) to obtain "⟨pseudo-word token⟩ ⟨connecting phrase⟩ ⟨text description⟩" as a query. Passing this via CLIP's text encoder would provide a sketch+text composed representation, that *can* be compared against gallery image features pre-computed via CLIP's vision encoder. The goal here is to learn sketch+text compositionality via CLIP, unsupervised, without any expensive paired textual description.

Overall Framework. Here, we aim to design a Sketch-Based Composed Image Retrieval (SBCIR) framework harnessing the Vision-Language (V-L) embedding of pre-trained CLIP [51] using only the sketch-photo pairing readily available in sketch datasets [10, 58, 70], *without* any annotated textual description. Accordingly, we embed the sketch into a pseudo-word token by first passing it through CLIP's visual encoder $\mathbf{V}$, followed by a visual-to-word converter $\mathbf{C_{v2w}}$, which is then passed into CLIP's text transformer $\mathbf{T_t}$ along with a few learnable prompts and additional textual descriptions (during inference) to obtain the composed query embedding. Specifically, we have three salient designs – *(i)* a novel compositionality constraint imposed via sketch/photo difference-signal to imitate the missing textual description (during training) and *neutral text* (Sec. 4.2) to preserve the grammatical structure of the input text-space of CLIP's text encoder, *(ii)* generalisable continuous prompt-learning (Sec. 4.3) over handcrafted textual prompts, and *(iii)* fine-grained matching (Sec. 4.4) between composed query and paired photo embedding via region-aware triplet loss and an auxiliary generative loss.

4.1. Baseline SBCIR

Conventional CLIP-based SBIR [55] maps both sketch query and its target photo in the joint embedding space using the visual encoder ($\mathbf{V}$). On the contrary, we convert a query sketch to a *pseudo-word token* and pass it through CLIP's textual encoder ($\mathbf{T_t}$) to generate its representation. In particular, given a photo $\mathcal{P}$, we generate its latent embedding ($\mathbf{p}$) as $\mathbf{p} = \mathbf{V}(\mathcal{P}) \in \mathbb{R}^{1\times d}$. For a sketch however, we generate its equivalent word embedding as $\mathbf{s}^w = \mathbf{C_{v2w}}(\mathbf{V}(\mathcal{S})) \in \mathbb{R}^{1\times d}$. $\mathbf{s}^w$ signifies the equivalent language representation of the visual sketch query.

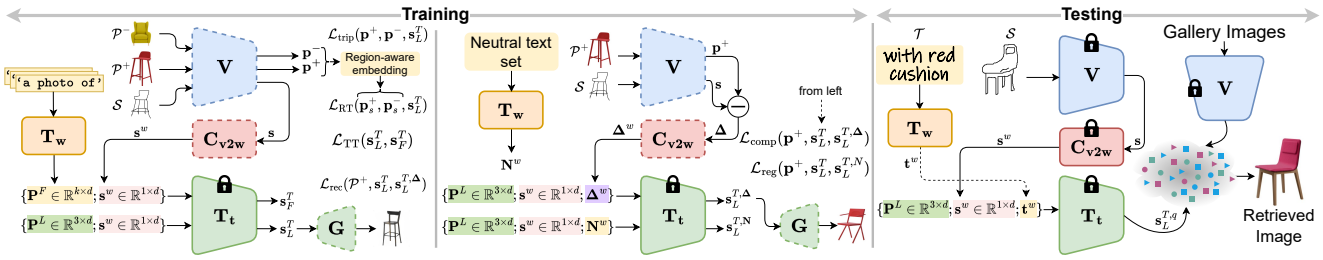

Figure 2. A query sketch ($\mathcal{S}$) is passed via CLIP's visual encoder ($\mathbf{V}$) followed by the visual-to-word converter ($\mathbf{C_{v2w}}$) to obtain pseudo-word token embedding ($\mathbf{s}^w$). It is then appended with a learnable continuous prompt $\mathbf{P}^L \in \mathbb{R}^{3\times d}$ and passed via frozen $\mathbf{T_t}$ to produce the final sketch embedding $\mathbf{s}_L^T$. *Compositionality constraint* (middle) is importantly a part of our multitask training (*not* a two-stage approach [2, 57]), where we compute $\mathbf{s}_L^{T,\Delta}$ (Sec. 4.2) by passing the *sketch-photo difference signal* $\mathbf{\Delta}$ via $\mathbf{C_{v2w}}$ and appending as $\mathbf{s}_L^{T,\Delta}$=$\{\mathbf{P}^L;\mathbf{s}^w;\mathbf{\Delta}^w\}$, using which $\mathcal{L}_{\text{comp}}$ is imposed. However, as this numeric signal $\mathbf{\Delta}^w$ does not exist in CLIP's [51] input text manifold, it may disrupt its grammatical syntax. Thus, we mine a set of "neutral text" (via GPT [6]) to impose a regularisation loss $\mathcal{L}_{\text{reg}}$. Apart from $\mathcal{L}_{\text{trip}}$, we use $\mathcal{L}_{\text{RT}}$ (region-aware triplet) with $\mathbf{s}_L^T$ and photo embeddings $\mathbf{p}^+/\mathbf{p}^-$ to enforce fine-grained matching. Additionally, a reconstruction loss $\mathcal{L}_{\text{rec}}$ trains a UNet decoder ($\mathbf{G}$) for further cross-modal alignment (Sec. 4.4). Furthermore, $\mathcal{L}_{\text{TT}}$, using a pre-defined set of standard language prompts, brings learnable prompts closer to *actual* English prompts for *unseen set* generalisation (Sec. 4.3). Specifically, we only train the `LayerNorm` of $\mathbf{V}$, $\mathbf{C_{v2w}}$, $\mathbf{P}^L$, and $\mathbf{G}$. The testing pipeline is shown on the right. *(Best view when zoomed.)*

We design the visual-to-word converter network $\mathbf{C_{v2w}}$ using 3-Layer MLP with ReLU [46]. Inspired from prompt learning [73, 74], we prepend a $\mathbb{R}^{3\times d}$ learnable continuous prompt vector $\mathbf{P}^L$ with $\mathbf{s}^w$ and pass this through $\mathbf{T_t}$ to finally obtain $\mathbf{s}_L^T = \mathbf{T_t}(\{\mathbf{P}^L;\mathbf{s}^w\}) \in \mathbb{R}^{1\times d}$. Baseline SBCIR trains over triplet loss [65] to perform cross-modal matching between $\mathbf{s}_L^T$, and positive ($\mathbf{p}^+$) and negative ($\mathbf{p}^-$) photo features. With margin $\mu_{\text{trip}} > 0$, triplet loss aims to minimise the distance $\delta(\cdot,\cdot)$ between $\mathbf{s}_L^T$ and $\mathbf{p}^+$, while increasing that from a random negative feature $\mathbf{p}^-$.

$$\mathcal{L}_{\text{trip}} = \texttt{max}(0, \mu_{\text{trip}} + \delta(\mathbf{s}_L^T, \mathbf{p}^+) - \delta(\mathbf{s}_L^T, \mathbf{p}^-)) \quad (1)$$

We only train the `LayerNorm` parameters of $\mathbf{V}$ [55], $\mathbf{C_{v2w}}$, and $\mathbf{P}^L$, while the rest including $\mathbf{T_t}$ remains frozen. Consequently, we can leverage the zero-shot compositionality of CLIP's text encoder $\mathbf{T}$ to perform SBCIR even if *no* text was provided during training. During inference, we append the word embedding of optional textual query ($\mathcal{T}$) as $\mathbf{t}^w = \mathbf{T_w}(\mathcal{T})$ to form $\{\mathbf{P}^L;\mathbf{s}^w;\mathbf{t}^w\}$, and pass it through frozen $\mathbf{T_t}$ to get the final composed query vector.

However, baseline SBCIR has a few major limitations – *(i)* most importantly, without access to paired textual descriptions during training, it *solely* relies on the frozen CLIP encoder [51] for zero-shot compositionality. We thus set out to explore how we can *imitate* the effect of adding this query text during training to learn the compositionality, *(ii)* while CLIP [51] performs fairly well for *category-level* semantic matching [55, 59], its *off-the-shelf* adaptation in the *fine-grained* setting is sub-optimal as seen in certain CLIP-based works [41, 55]. So, how can we make the cross-modal matching more *fine-grained*? *(iii)* prompt learning is prone to overfitting on the training set, delivering poor test-set zero-shot performance [7]. Furthermore, vanilla prompt learning is not robust [7] and is prone to distort the composed query embedding, particularly when combined with the optional text coming from an *in-the-wild* scenario.

4.2. Learning Compositionality Constraint

While our method does not rely on paired textual descriptions *during training*, users can provide optional textual descriptions *during inference*, to specify the desired *additional information*, which gets composed with the query sketch to retrieve the correct image. We handle this training-testing disparity, by hypothesising the *additional information* to be equivalent to the difference between the query sketch and its paired photo. To this end, we compute the sketch-photo *difference* signal embedding $\mathbf{\Delta}^w = \mathbf{C_{v2w}}(|\mathbf{p}^+ - \mathbf{s}|)$ and append it as $\{\mathbf{P}^L;\mathbf{s}^w;\mathbf{\Delta}^w\}$, which upon passing through $\mathbf{T_t}$ generates the $\mathbf{s}_L^{T,\mathbf{\Delta}}$ embedding. Here $\mathbf{\Delta}^w$ could be considered as a "single vector" pseudo-word token *imitating* the difference between sketch and photo, which ideally would be substituted with real query text during inference. Now, as per our hypothesis, we enforce compositionality constraint $\mathcal{L}_{\text{comp}}$ (with $\mu_{\text{comp}} > 0$) that ensures that the distance between $\mathbf{p}^+$ and $\mathbf{s}_L^{T,\mathbf{\Delta}}$ is less than the same between $\mathbf{p}^+$ and $\mathbf{s}_L^T$, as $\mathbf{\Delta}^w$ reinforces $\mathbf{s}_L^T$ with *additional information*.

$$\mathcal{L}_{\text{comp}} = \texttt{max}(0, \mu_{\text{comp}} + \delta(\mathbf{s}_L^{T,\mathbf{\Delta}}, \mathbf{p}^+) - \delta(\mathbf{s}_L^T, \mathbf{p}^+)) \quad (2)$$

Although $\mathbf{\Delta}$ enforces compositionality, this mere numeric signal does not exist in CLIP's [51] input text manifold and might break its grammatical syntax. Furthermore, this additional signal might train $\mathbf{C_{v2w}}$ sub-optimally, rendering its output incompatible with CLIP's input text manifold. Thus, to restrict the adverse effect of $\mathbf{\Delta}$ to a minimum level, we regularise the training via a *"neutral-text"* set containing a list of 3-5 word *generic* description of a freehand sketch (*e.g.*, "`with a line drawing`"). To this end, we replace $\mathbf{\Delta}^w$ from composed query $\mathbf{s}_L^{T,\mathbf{\Delta}}$ with any one random phrase from the neutral-text set, to generate neutral-text

enriched composed representation $\{\mathbf{P}^L; \mathbf{s}^w; \mathbf{N}^w\}$, which upon passing through $\mathbf{T_t}$ generates the $\mathbf{s}_L^{T,N}$ embedding. Here $\mathbf{N}^w$ is the word-embedding of the chosen neutral phrase. We posit that using a *generic description* for a sketch should neither enhance nor impair the composed query. Thus, we enforce the distance between $\mathbf{p}^+$ and $\mathbf{s}_L^{T,\mathbf{\Delta}}$ to be *equivalent* to the distance between $\mathbf{p}^+$ and $\mathbf{s}_L^{T,N}$.

$$\mathcal{L}_{\text{reg}} = ||\delta(\mathbf{s}_L^{T,\mathbf{\Delta}}, \mathbf{p}^+) - \delta(\mathbf{s}_L^{T,N}, \mathbf{p}^+)||_2 \quad (3)$$

We prompt a lightweight GPT [6], to generate 100 different 3-5 word phrases (more in § Suppl.) describing a "freehand sketch" to form our optimum neutral text set.

4.3. Generalised Prompt Learning

Prompt learning literature [7, 73, 74] dictates that handcrafted fixed language prompts (*e.g.*, "`a photo of`", "`an image of`") generalise better on *unseen* sets, while learnable continuous prompts depict better performance on *seen* sets used to learn it. While we are using learnable continuous prompts $\mathbf{P}^L$ over handcrafted ones, we impose a *text-to-text* generalisation loss that enforces the learned prompts to be similar to a set of handcrafted language prompts in the text embedding space, so that it generalises beyond the seen training set. In particular, at every instance, we randomly pick one handcrafted fixed language prompt d_i from a set of handcrafted [74] fixed language prompts (more in § Suppl.) $\mathcal{D}$ and prepend its word-embedding representation $\mathbf{P}^F = \mathbf{T_w}(d_i)$ to $\mathbf{s}^w$ as $\{\mathbf{P}^F, \mathbf{s}^w\}$ and pass it via $\mathbf{T_t}$ to generate the fixed representation $\mathbf{s}_F^T$. Now we employ $\mathcal{L}_{\text{TT}}$ to enforce the sketch query representation with the learned prompts $\mathbf{s}_L^T$ to be similar with that of the fixed language prompt $\mathbf{s}_F^T$ as:

$$\mathcal{L}_{\text{TT}} = ||\mathbf{s}_F^T - \mathbf{s}_L^T||_2 \quad (4)$$

The utility of $\mathcal{L}_{\text{TT}}$ is multi-fold – *(i)* it alleviates seen set overfitting, *(ii)* typically, learned prompts reside in the sparse regions of the CLIP manifold [2], limiting its intractability with actual query texts during inference. Regularising $\mathbf{P}^L$ with actual language supervision suppresses this issue, and *(iii)* the diverse list of fixed prompts, acts as an additional augmentation in the language domain [7], which reinforces the robustness of the learned prompts.

4.4. Fine-Grained Matching

Region-Aware Triplet Loss. To further improve the cross-modal fine-grained matching, we consider the CLIP [51] vision encoder $\mathbf{V}$ (employed through a vision transformer) that breaks the input image ($\mathcal{P}$) into T patches and passes them via transformer layer to generate patch-wise feature $\mathbf{p}_r = \mathbf{V}(\mathcal{P}) \in \mathbb{R}^{T \times d}$, where T is the number of patches. To enforce region-wise associativity, we use the patch-wise embedding $\mathbf{p}_r$ from all T patches and calculate the patch-level correlation (a) between global sketch query feature $\mathbf{s}_L^T \in \mathbb{R}^{1 \times d}$ and $\mathbf{p}_r$ as: $a = (\mathbf{p}_r \cdot \mathbf{s}_L^T) \in \mathbb{R}^{T \times 1}$, which is `SoftMax` normalised across the patch dimension. Every value a_i denotes the associativity between the global sketch query and patch-wise photo features. Now we take a weighted sum across all patch embeddings to get a region-aware photo feature as: $\mathbf{p}_s = \sum_{i=1}^{T}(a_i \times \mathbf{p}_r^i)$. We utilise this *region-aware embeddings* $\mathbf{p}_s^+$ and $\mathbf{p}_s^-$ from the positive and the negative photo respectively to impose one additional region-aware triplet loss $\mathcal{L}_{\text{RT}}$ with margin $\mu_{\text{RT}} > 0$ as:

$$\mathcal{L}_{\text{RT}} = \texttt{max}(0, \mu_{\text{RT}} + \delta(\mathbf{s}_L^T, \mathbf{p}_s^+) - \delta(\mathbf{s}_L^T, \mathbf{p}_s^-)) \quad (5)$$

It is noteworthy that although the $\mathcal{L}_{\text{RT}}$ acts as a proxy to better align the joint embedding space for fine-grained matching, we perform inference on the *global feature*.

Auxiliary Generator Guidance. With sketches, triplet loss typically performs fine-grained shape matching [3, 70], ignoring the *fine-grained* appearance features (*e.g.*, colour, texture). Being a *composed* retrieval framework, we aim to encompass appearance traits along with structural ones in the visual domain. Considering the proven efficacy of cross-modal translation [48] in fine-grained matching, we impose a sketch-to-photo reconstruction objective, where given the sketch query representation, we train a simple UNet [52] decoder ($\mathbf{G}$) to reconstruct the ground truth photo using pixel-level l_2 reconstruction loss. Please note, that the aim of $\mathcal{L}_{\text{rec}}$ is not to enforce photorealistic image generation, but to impose an appearance guidance.

$$\mathcal{L}_{\text{rec}} = ||\mathcal{P}^+ - \mathbf{G}(\mathbf{s}_L^T)||_2 + ||\mathcal{P}^+ - \mathbf{G}(\mathbf{s}_L^{T,\mathbf{\Delta}})||_2 \quad (6)$$

To sum up, our overall training objective becomes: $\mathcal{L}_{\text{total}}$ = $\lambda_1\mathcal{L}_{\text{trip}}$ + $\lambda_2\mathcal{L}_{\text{comp}}$ + $\lambda_3\mathcal{L}_{\text{reg}}$ + $\lambda_4\mathcal{L}_{\text{TT}}$ + $\lambda_5\mathcal{L}_{\text{RT}}$ + $\lambda_6\mathcal{L}_{\text{rec}}$. Our model employs multitask learning with multiple losses updating the `LayerNorm` of $\mathbf{V}$, $\mathbf{C_{v2w}}$, $\mathbf{P}^L$, and $\mathbf{G}$. During inference, we discard the UNet decoder and use frozen $\mathbf{V}$, $\mathbf{T_t}$, and $\mathbf{C_{v2w}}$ to first generate a pseudo-word token from query sketch as: $\mathbf{s}^w = \mathbf{C_{v2w}}(\mathbf{V}(\mathcal{S}))$. We prepend $\mathbf{P}^L$ to $\mathbf{s}^w$ followed by appending the tokenised representation of the additional user-given textual description $\mathbf{t}^w = \mathbf{T_w}(\mathcal{T})$ to form the final composed query token $\{\mathbf{P}^L, \mathbf{s}^w, \mathbf{t}^w\}$. This composed query token, upon passing through $\mathbf{T_t}$, generates the final composed query feature $\mathbf{s}_L^{T,q}$. With pre-computed visual features of the gallery images (retrieval candidates), we perform retrieval by comparing the distance between the gallery features and $\mathbf{s}_L^{T,q}$ (both l_2 normalised).

5. Experiments

Datasets. We evaluate on the following datasets. QMUL-ShoeV2 [70] and QMUL-ChairV2 [70] contain $2000/6730$, and $400/1800$ sketches/photos respectively, with fine-grained association. The Sketchy [58] dataset comprises $12,500$ photos across 125 classes with at least 5 sketches per photo. For scene-level retrieval, we use FS-COCO [10] and SketchyCOCO [22] containing $10,000$ and $14,081$ paired sketch-text-photo triplets respectively, where images and textual captions are sourced from MS-COCO [38]. We

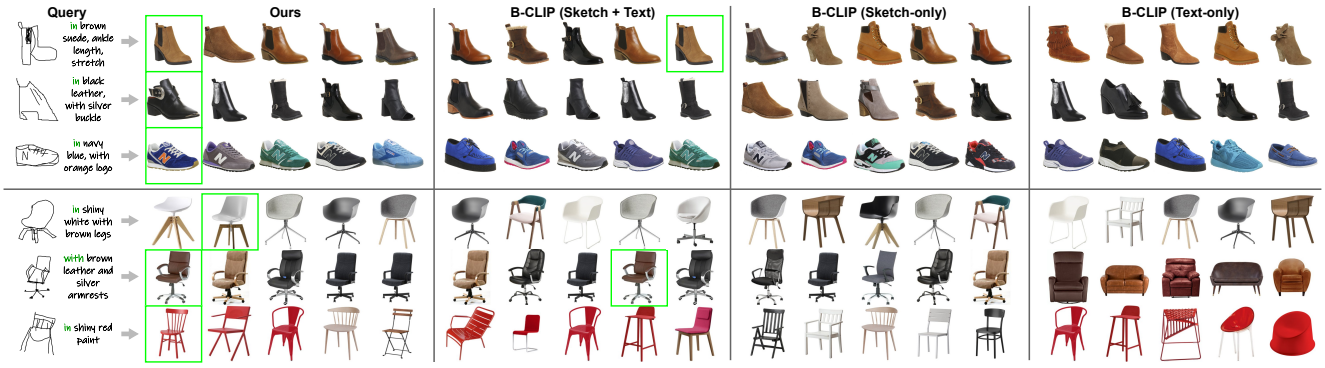

Figure 3. Top-5 *fine-grained retrieval* result comparison on ShoeV2/ChairV2. GT photos are green-bordered. *(Zoom-in for best-view)*

also use the ImageNet-R(endition) [26] image-only dataset, which consists $30,000$ images across 200 ImageNet [16] classes and 16 domains with domain annotations.

Implementation Details. We use a pre-trained ViT-L/14 CLIP vision encoder [51] in all experiments with an embedding dimension $d = 768$. The prompt vectors are trained with a learning rate of 10^{-5}, keeping the encoder frozen (except `LayerNorm` layers). The UNet decoder and the visual-to-word converter are trained with a learning rate of 10^{-4} and 10^{-3} respectively. We train the model for 100 epochs using AdamW [40] optimiser with 0.09 weight decay, and a batch size of 128. Values of $\lambda_{1,2,3,4,5,6}$ are set to 1, 0.5, 0.1, 0.1, 1, and 1, empirically.

Competitors. We evaluate from three perspectives – *(i) Sketch or Text-only **B**aselines:* Here we validate the additional accuracy gain achieved by sketch+text composition over well-studied individual sketch/text-only retrieval paradigms. For sketch-only baselines, we use triplet loss-based frameworks [55] of **B-DINOv2 (S)** and **B-CLIP (S)**, using DINOv2 [47] and CLIP [51] ViT-L/14 as backbones respectively. Given the de facto usage of pre-trained CLIP [51] in TBIR, for *text-only* baselines we introduce **B-CLIP (T)**, which employs frozen CLIP [51] vision and language encoder to retrieve by comparing query text feature against gallery image features. *(ii) Sketch+Text Composed SoTA:* To judge CLIP's off-the-shelf potential in sketch+text composition, we employ **B-CLIP (S+T)** which uses the mean of sketch and text features from CLIP's frozen vision and language encoders for retrieval. Now, as most supervised sketch+text SoTA models [1, 11, 59] train using *paired textual captions*, we employ SoTA image captioner BLIP [35] to generate textual captions for training images which however are often noisy, generic and non-discriminative [35]. Among supervised SoTAs, **Combiner** [1] trains a network to fuse paired image-text features of *frozen* CLIP encoders [51] to generate multi-modal query features, while **TASK-former** [59] merges paired sketch and text features via element-wise addition and *trains* CLIP's vision and language encoders [51] end-to-end. **SceneTrilogy** [11] models sketch-text-photo joint-embedding by training an invertible neural network. *(iii) Unsupervised Sketch+Text Composition:* Leveraging *unlabelled photos*, **Pic2Word** [57] learns a textual-inversion network to map input visual query into a pseudo-word token in CLIP's textual embedding space for retrieval. Unlike Pic2Word, **SEARLE** [2] generates a set of pseudo-word tokens from *unlabelled photos* using optimisation-based textual-inversion, and then uses those image-token pairs to learn a textual-inversion network. While such methods either train a textual-inversion network on a massive $3M$ image dataset [57] or use time-consuming optimisation-based textual-inversion for image-token pair generation [2], we exploit pre-trained CLIP model to address sketch+text compositionality in an *unsupervised* manner without any associated textual descriptions. We adapt Combiner [1], Pic2Word [57], and SEARLE [2], by replacing the input *image* with *sketch*. For fairness, we keep the same training/testing paradigm for all competing methods.

Evaluation Setup. In the fine-grained setup, we aim to retrieve the target image using the composed query formed by an input sketch and text. We use a train:test split of 90:10 for Sketchy [58] (following [56]), 7000:3000 for FS-COCO [10] and 1015:210 for SketchyCOCO [22]. For ShoeV2/ChairV2 we use $1800/300$ $(6051/1275)$ photos (sketches) for training and the rest for testing. Notably, our method *does not* use captions during training. For *evaluation*, we manually collect fine-grained captions for each of the test-set images of ShoeV2, ChairV2, and Sketchy. We compose the query as $\{\mathbf{P}^L; \mathbf{s}^w; [\texttt{text}]\}$, where $[\texttt{text}]$ denotes the word embedding of textual query with suitable prepositions (*e.g.*, 'with', 'in'). We use Acc.@q to denote the percentage of sketches with true-matched photos in the top-q retrieved images.

5.1. Performance Analysis

Tab. 1 delineates the quantitative results while the qualitative ones are shown in Fig. 3. In the fine-grained composed retrieval setup (Tab. 1), our method outperforms baselines and SoTAs significantly on all datasets, indicating its efficiency in seamlessly combining fine-grained sketch with textual description. This gain is likely due to the regularisa-

Table 1. Results for *fine-grained* object-level and scene-level composed retrieval.

Methods	Object-level						Scene-level			
	ShoeV2		ChairV2		Sketchy		FS-COCO		SketchyCOCO	
	Acc.@5	Acc.@10	Acc.@5	Acc.@10	Acc.@5	Acc.@10	Acc.@5	Acc.@10	Acc.@5	Acc.@10
B-CLIP (S)	9.8	17.5	16.7	18.4	6.8	11.3	5.9	9.7	6.8	10.2
B-DINOv2 (S)	10.2	19.4	17.9	20.2	8.5	15.1	7.6	11.4	9.4	12.2
B-CLIP (T)	9.1	16.6	15.4	17.8	5.9	10.2	5.5	10.1	6.7	10.6
B-CLIP (S+T)	19.1	30.8	30.2	32.3	10.1	20.2	10.2	15.4	11.2	20.2
Combiner [1]	24.7	40.2	35.7	39.9	15.7	33.7	11.6	22.1	15.9	32.2
TASK-former [59]	27.7	44.1	40.7	45.2	17.8	35.2	12.7	24.2	19.4	34.7
SceneTrilogy [11]	29.1	46.2	43.4	46.8	19.7	37.2	14.5	28.3	20.4	40.2
Pic2Word [57]	34.7	58.4	55.7	62.1	22.5	48.7	16.7	32.6	24.4	46.0
SEARLE [2]	38.4	64.8	60.8	66.4	25.3	54.2	17.7	35.9	26.0	50.4
Proposed	**47.3**	**79.1**	**73.5**	**81.4**	**30.6**	**64.2**	**22.7**	**43.5**	**33.4**	**61.1**
Avg. Improvement	*+24.7*	*+45.5*	*+38.3*	*+42.6*	*+15.9*	*+34.6*	*+11.3*	*+22.4*	*+17.8*	*+32.5*

tion provided by our region-aware contrastive loss and generator guidance. Our closest competitors (*i.e.*, Pic2Word [57], and SEARLE [2]) attempt to handle composed retrieval by either training data-hungry textual-inversion network [57] or use optimisation-based textual-inversion for image-token pair generation followed by training an inversion network [2]. Surprisingly, our method achieves the highest Acc.@5 of 47.3 (73.5) in ShoeV2 (ChairV2) without the $3M$ data-requirement of Pic2Word [57], or the complicated two-stage approach of SEARLE [2].

Being more challenging than object-level [9], baseline methods perform quite poorly (Tab. 1) for scene-image retrieval. However, thanks to the increased interaction capability (learned via compositionality constraint) of the pseudo-word token with user-given textual queries during inference, we surpass others with an Acc.@5 of 22.7 (33.4) on FS-COCO (SketchyCOCO).

5.2. Downstream Tasks

Sketch+Text-based Fine-Grained Image Generation. Apart from composed retrieval, our method is also suitable for sketch+text composed object image generation. Here we replace the low-quality UNet decoder with a StyleGAN2 [29] generator (pre-trained on specific classes). The composed query $\mathbf{s}_L^{T,q}$ (acting as a latent vector [29]) upon passing through the frozen StyleGAN2 generates output photos. We pass $\mathbf{s}_L^{T,q}$ via a learnable FC-layers to convert it to the dimensions required by StyleGAN2's affine transformation layer [29]. Fig. 4 shows a few cases of such fine-grained generation. Notably, the *semantic geometry* (*e.g.*, shape, structure, etc.) of generated images is driven by input sketches, whereas the high-level appearance (*e.g.*, colour, shade, etc.) is mostly governed by textual descriptions. Overall our method archives a lower FID [29, 34] score of 33.4(88.5) on ShoeV2 (ChairV2) test-set images compared to 35.85(90.21) of the current SoTA [31].

Figure 4. Qualitative results for *sketch+text composed fine-grained generation* with pre-trained StyleGAN2 [29] models.

Object Sketch-based Scene Image Retrieval. It aims to retrieve *scene* images from a *single-object* sketch and additional captions. Here, we use 7000(3000) and 1015(210) train(test) sketch-photo pairs from FS-COCO and SketchyCOCO respectively. Since they lack single-object sketches, we source them from Sketchy, which is their superset in terms of classes [10]. Here, a retrieval is deemed correct if it contains *all* objects that were queried in the sketch and text. As FS-COCO/SketchyCOCO images are derived from MS-COCO [38], we use their segmentation-map *labels* from MS-COCO to create ground truth object-lists per test-set image. During inference, we use the readily available captions of test-set images of FS-COCO/SketchyCOCO, but remove the object-*name* queried via sketch (Fig. 5). Here we compose the query like in the fine-grained composed retrieval setup and use Acc.@q as evaluation metric. Tab. 2 shows our method to surpass other baseline and SoTAs with an average Acc.@5 gain of 10.9 on FS-COCO [10].

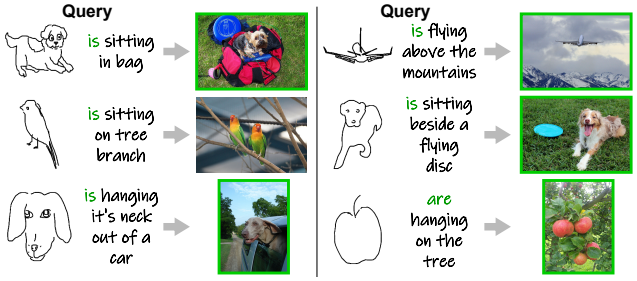

Figure 5. Qualitative result for *object sketch-based scene image retrieval* on FS-COCO [10]. GT photos are green-bordered.

Domain Attribute Transfer. Here we perform SBIR from a specified domain (*e.g.*, origami), where the domain name is additionally provided by a textual domain label. Here we use ImageNet-R [26] dataset. This being an image-only dataset, we source freehand sketches from the 104 common classes from the Sketchy [58] dataset with a train:test split of 90:10. Here, a retrieval is deemed correct, if its class and domain name match those of the query sketch and domain label. Due to non-uniform and noisy domain labels, we use four domains here *viz.*, tattoo, origami, sculpture, and painting. We compose the query as $\{\mathbf{P}^L; \mathbf{s}^w; [\texttt{in}]; [\texttt{domain}]\}$, where $[\texttt{domain}]$ denotes the word embedding of the query domain label. Following [57], we use recall@q (r@q) as the evaluation metric, which denotes the ratio of positive

retrieved images in the top-q list to all relevant images for a given query. As strict domain constraints complicate cross-modal composed image retrieval, baseline methods perform poorly here, with B-CLIP (S+T) attaining an average r@50 of 12.9 across four test domains (Tab. 2). Due to their respective CLIP [51] feature composition strategies, Combiner [1], Pic2Word [57] and SEARLE [2] depict reasonable performance on ImageNet-R, while our method achieves a notable average r@10 of 15.3.

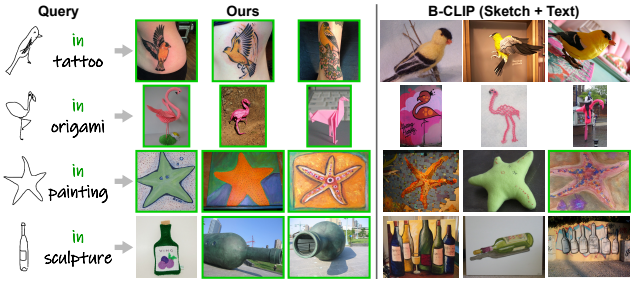

Figure 6. Top-3 *domain attribute transfer* results comparison on ImageNet-R [26]. GT photos are green-bordered.

Table 2. Results for *domain attribute transfer* and *object sketch-based scene image retrieval*.

Methods	Domain transfer		Object sketch-based scene retrieval			
	ImageNet-R		FS-COCO		SketchyCOCO	
	r@10	r@50	Acc.@5	Acc.@10	Acc.@5	Acc.@10
B-CLIP (S+T)	2.2	12.9	4.3	7.0	10.1	26.7
Combiner [1]	8.3	19.8	10.8	21.1	16.7	30.9
TASK-former [59]	7.6	18.2	11.6	23.1	18.2	35.8
SceneTrilogy [11]	9.7	20.5	14.2	25.3	21.0	40.8
Pic2Word [57]	10.8	22.7	–	–	–	–
SEARLE [2]	12.1	25.4	–	–	–	–
Proposed	**15.3**	**27.1**	**21.2**	**40.4**	**32.4**	**53.3**
Avg. Improvement	*+6.8*	*+7.1*	*+10.9*	*+21.2*	*+15.9*	*+19.4*

Table 3. Ablation on design.

Methods	ShoeV2		ChairV2		FS-COCO	
	Acc.@5	Acc.@10	Acc.@5	Acc.@10	Acc.@5	Acc.@10
w/o $\mathcal{L}_{\text{TT}}$	41.8	71.9	68.3	77.7	17.1	38.2
w/o $\mathcal{L}_{\text{rec}}$	40.7	72.8	70.9	75.5	21.4	42.7
w/o $\mathcal{L}_{\text{RT}}$	40.2	71.6	69.1	76.2	10.5	21.7
w/o compositionality	32.5	48.2	45.7	48.9	18.4	32.3
Ours-full	**47.3**	**79.1**	**73.5**	**81.4**	**22.7**	**43.5**

5.3. Ablation on Design

• **How well does s^w capture sketch semantics?** To judge the efficacy of pseudo-word token s^w in representing visual content of a sketch, we evaluate our trained model on retrieving an input sketch *solely* from its pseudo-word token *without* additional textual description. Acc.@1 of 98.35% on ShoeV2 *test-set* sketches shows the pseudo-word token to capture visual sketch features fairly well. Although representing fine-grained query sketches with one pseudo-word token might be sub-optimal, experimenting with two and three such tokens delivered similar Acc.@1 of 98.79% and 99.11% respectively on ShoeV2. We thus stick to a single-word token for computational ease.

• **Contribution of $\mathcal{L}_{\text{RT}}$ and $\mathcal{L}_{\text{rec}}$:** Region-aware local features are pivotal in bridging the huge domain gap between sparse-binary sketches and pixel-dense photos. Although less reflected on ShoeV2/ChairV2 results, a notable Acc.@5 drop of 12.2 on FS-COCO (Tab. 3) for **w/o $\mathcal{L}_{\text{RT}}$** verifies its importance in the scene-level setup. Furthermore, a 13.9% drop in Acc.@5 (ShoeV2) for **w/o $\mathcal{L}_{\text{rec}}$** shows that fine-grained matching remains incomplete without the proposed generator guidance.

• **Impact of $\mathcal{L}_{\text{TT}}$:** Removing Text-to-Text loss plummets Acc.@5 by 11.6% on ShoeV2 (Tab. 3), highlighting its vital role in aligning learned and English language prompts seen by CLIP [51] during its training. This removal likely pushes the final language embedding towards sparser parts of CLIP language manifold [2], hindering effective communication with *real-language* tokens during inference [2].

• **Why Compositionality Constraint?** Introduced via the novel idea of *neutral text*, compositionality constraint helps preserve internal grammar of CLIP language manifold, to allow optional user-provided query texts during inference. On removing that, Acc.@5 drops the lowest across all datasets (Tab. 3), proving its importance in our framework.

• **On combining sketch and text:** Our composed retrieval pipeline places the query feature *closer* to its paired photo in the latent retrieval space (Fig. 7), than *only* text/sketch-based retrieval. The relative distances (Fig. 7 top insets) depict that the individual text and sketch feature together *pushes* the composed feature towards the paired photo.

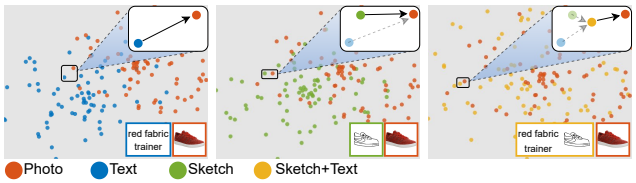

Figure 7. t-SNE plots showing the feature distances for text-based, sketch-based, and composed retrieval. Compared to sketch/text-based retrieval, combining sketch and text pushes the composed embedding closer to the ground truth photo in the latent manifold.

6. Conclusion and Future Works

In conclusion, our exploration into the fine-grained representation power of both sketch and text, coupled with the orchestration of their synergistic interplay, marks a significant stride in the realm of image retrieval. By harmonising sketches and text, we offer users a retrieval experience that transcends traditional category-level distinctions. The introduction of a novel compositionality framework, driven by pre-trained CLIP models, eliminates the need for extensive fine-grained textual annotations. Last but not least, our system extends its utility to diverse domains such as sketch+text-based fine-grained image generation, object-sketch-based scene retrieval, domain attribute transfer, etc.

References

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective Conditioned and Composed Image Retrieval Combining CLIP-Based Features. In *CVPR*, 2022. 2, 6, 7, 8

[2] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-Shot Composed Image Retrieval with Textual Inversion. In *ICCV*, 2023. 2, 3, 4, 5, 6, 7, 8

[3] 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. 2, 5

[4] Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketching without Worrying: Noise-Tolerant Sketch-Based Image Retrieval. In *CVPR*, 2022. 1

[5] Ayan Kumar Bhunia, Aneeshan Sain, Parth Hiren Shah, Animesh Gupta, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Adaptive Fine-Grained Sketch-Based Image Retrieval. In *ECCV*, 2022. 2

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. *NeurIPS*, 2020. 2, 4, 5

[7] Adrian Bulat and Georgios Tzimiropoulos. LASP: Text-to-Text Optimization for Language-Aware Soft Prompting of Vision & Language Models. In *CVPR*, 2023. 4, 5

[8] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning Aligned Cross-Modal Representations from Weakly Aligned Data. In *CVPR*, 2016. 1

[9] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Partially Does It: Towards Scene-Level FG-SBIR with Partial Input. In *CVPR*, 2022. 2, 7

[10] Pinaki Nath Chowdhury, Aneeshan Sain, Yulia Gryaditskaya, Ayan Kumar Bhunia, Tao Xiang, and Yi-Zhe Song. FS-COCO: Towards Understanding of Freehand Sketches of Common Objects in Context. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7

[11] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. SceneTrilogy: On Human Scene-Sketch and its Complementarity with Photo and Text. In *CVPR*, 2023. 1, 2, 3, 6, 7, 8

[12] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic Embeddings for Cross-Modal Retrieval. In *CVPR*, 2021. 2

[13] Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. "This is my unicorn, Fluffy": Personalizing frozen vision-language representations. In *ECCV*, 2022. 2, 3

[14] John Collomosse, Tu Bui, Michael J Wilber, Chen Fang, and Hailin Jin. Sketching with Style: Visual Search with Sketches and Aesthetic Context. In *ICCV*, 2017. 2

[15] John Collomosse, Tu Bui, and Hailin Jin. LiveSketch: Query Perturbations for Guided Sketch-based Visual Search. In *CVPR*, 2019. 2

[16] 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. 6

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 3

[18] Anjan Dutta and Zeynep Akata. Semantically Tied Paired Cycle Consistency for Zero-Shot Sketch-based Image Retrieval. In *CVPR*, 2019. 1

[19] Aviv Eisenschtat and Lior Wolf. Linking Image and Text with 2-Way Nets. In *CVPR*, 2017. 2

[20] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *BMVC*, 2018. 2

[21] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *ICLR*, 2023. 2, 3

[22] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. SketchyCOCO: Image Generation from Freehand Scene Sketches. In *CVPR*, 2020. 5, 6

[23] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. FashionVLP: Vision Language Transformer for Fashion Retrieval with Feedback. In *CVPR*, 2022. 2

[24] Cusuh Ham, Gemma Canet Tarres, Tu Bui, James Hays, Zhe Lin, and John Collomosse. CoGS: Controllable Generation and Search from Sketch and Style. In *ECCV*, 2012. 3

[25] Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. FashionViL: Fashion-Focused Vision-and-Language Representation Learning. In *ECCV*, 2022. 2

[26] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *ICCV*, 2021. 6, 7, 8

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*, 2021. 2

[28] Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*, 2015. 2

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *CVPR*, 2020. 7

[30] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal Prompt Learning. In *CVPR*, 2023. 3

[31] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Picture that Sketch: Photorealistic Image Generation from Abstract Sketches. In *CVPR*, 2023. 7

[32] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. How to Handle Sketch-Abstraction in Sketch-Based Image Retrieval? In *CVPR*, 2024. 1

[33] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Text-to-Image Diffusion Models are Great Sketch-Photo Matchmakers. In *CVPR*, 2024. 2

[34] Subhadeep Koley, Ayan Kumar Bhunia, Deeptanshu Sekhri, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. It's All About Your Sketch: Democratising Sketch Control in Diffusion Models. In *CVPR*, 2024. 7

[35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023. 6

[36] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV*, 2020. 2

[37] Hangyu Lin, Yanwei Fu, Peng Lu, Shaogang Gong, Xiangyang Xue, and Yu-Gang Jiang. TC-Net for iSBIR: Triplet Classification Network for Instance-level Sketch Based Image Retrieval. In *ACM MM*, 2019. 2

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 5, 7

[39] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep Sketch Hashing: Fast Free-hand Sketch-Based Image Retrieval. In *CVPR*, 2017. 2

[40] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 6

[41] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch Aggregation with Learnable Centers for Open-Vocabulary Semantic Segmentation. In *ICML*, 2023. 4

[42] Aryan Mikaeili, Or Perel, Mehdi Safaee, Daniel Cohen-Or, and Ali Mahdavi-Amiri. SKED: Sketch-guided Text-based 3D Editing. In *CVPR*, 2023. 3

[43] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tanick, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*, 2018. 3

[44] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[45] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM TOG*, 2022. 3

[46] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *ICML*, 2010. 4

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[48] Kaiyue Pang, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Cross-domain Generative Learning for Fine-Grained Sketch-Based Image Retrieval. In *BMVC*, 2017. 5

[49] 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. 2

[50] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional Image-Text Embedding Networks. In *ECCV*, 2018. 2

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 2, 3, 4, 5, 6, 8

[52] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 5

[53] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Cross-Modal Hierarchical Modelling for Fine-Grained Sketch Based Image Retrieval. In *BMVC*, 2020. 2

[54] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. StyleMeUp: Towards Style-Agnostic Sketch-Based Image Retrieval. In *CVPR*, 2021. 1, 2

[55] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. CLIP for All Things Zero-Shot Sketch-Based Image Retrieval, Fine-Grained or Not. In *CVPR*, 2023. 2, 3, 4, 6

[56] Aneeshan Sain, Ayan Kumar Bhunia, Subhadeep Koley, Pinaki Nath Chowdhury, Soumitri Chattopadhyay, Tao Xiang, and Yi-Zhe Song. Exploiting Unlabelled Photos for Stronger Fine-Grained SBIR. In *CVPR*, 2023. 6

[57] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval. In *CVPR*, 2023. 2, 3, 4, 6, 7, 8

[58] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies. *ACM TOG*, 2016. 3, 5, 6, 7

[59] Patsorn Sangkloy, Wittawat Jitkrittum, Diyi Yang, and James Hays. A Sketch Is Worth a Thousand Words: Image Retrieval with Text and Sketch. In *ECCV*, 2022. 1, 2, 3, 4, 6, 7, 8

[60] Jifei Song, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma. In *BMVC*, 2017. 1

[61] 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. 2

[62] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing Text and Image for Image Retrieval - An Empirical Odyssey . In *CVPR*, 2019. 2

[63] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-Guided Text-to-Image Diffusion Models. In *SIGGRAPH Asia*, 2023. 3

[64] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval. In *ICCV*, 2019. 2

[65] Kilian Q Weinberger and Lawrence K Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *JMLR*, 2009. 4

[66] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *CVPR*, 2021. 2

[67] Jiaqing Xu, Haifeng Sun, Qi Qi, Jingyu Wang, Ce Ge, Lejian Zhang, and Jianxin Liao. DLA-Net for FG-SBIR: Dynamic Local Aligned Network for Fine-Grained Sketch-Based Image Retrieval. In *ACM MM*, 2021. 2

[68] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. Deep Plastic Surgery: Robust and Controllable Image Editing with Human-Drawn Sketches. In *ECCV*, 2020. 2

[69] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A Zero-Shot Framework for Sketch-based Image Retrieval. In *ECCV*, 2018. 1

[70] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch Me That Shoe. In *CVPR*, 2016. 2, 3, 5

[71] Hua Zhang, Peng She, Yong Liu, Jianhou Gan, Xiaochun Cao, and Hassan Foroosh. Learning Structural Representations via Dynamic Object Landmarks Discovery for Sketch Recognition and Retrieval. *IEEE TIP*, 2019. 2

[72] Lvmin Zhang and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023. 3

[73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional Prompt Learning for Vision-Language Models. In *CVPR*, 2022. 3, 4, 5

[74] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. *IJCV*, 2022. 4, 5