Partially Does It: Towards Scene-Level FG-SBIR with Partial Input

Pinaki Nath Chowdhury[1,2]   Ayan Kumar Bhunia[1]   Viswanatha Reddy Gajjala[*]
Aneeshan Sain[1,2]   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.
{p.chowdhury, a.bhunia, a.sain, t.xiang, y.song}@surrey.ac.uk

Abstract

We scrutinise an important observation plaguing scene-level sketch research – that a significant portion of scene sketches are "partial". A quick pilot study reveals: (i) a scene sketch does not necessarily contain all objects in the corresponding photo, due to the subjective holistic interpretation of scenes, (ii) there exists significant empty (white) regions as a result of object-level abstraction, and as a result, (iii) existing scene-level fine-grained sketch-based image retrieval methods collapse as scene sketches become more partial. To solve this "partial" problem, we advocate for a simple set-based approach using optimal transport (OT) to model cross-modal region associativity in a partially-aware fashion. Importantly, we improve upon OT to further account for holistic partialness by comparing intra-modal adjacency matrices. Our proposed method is not only robust to partial scene-sketches but also yields state-of-the-art performance on existing datasets.

1. Introduction

The prevailing nature of touch-screen devices has triggered significant research progress on sketches [6, 16, 24, 36, 44]. The field has predominantly focused on object-level sketches to date [21, 25, 59], studying its abstractness [32], creativeness [25], and applications such as image retrieval [9, 78], and 3D synthesis/editing [29]. It was not until very recently that research efforts has undertaken a shift to scene-level analysis [37, 44, 90]. Compared to object-level sketches [21, 59], scene sketches exhibit abstraction not only on individual objects, but also on global scene configurations. Fig. 1(a) offers some examples where randomly selected sketches are overlapped on top of their corresponding photos.

We start with an important observation plaguing scene-level sketch research – that a significant portion of scene sketches are "partial". This partialness happens on two
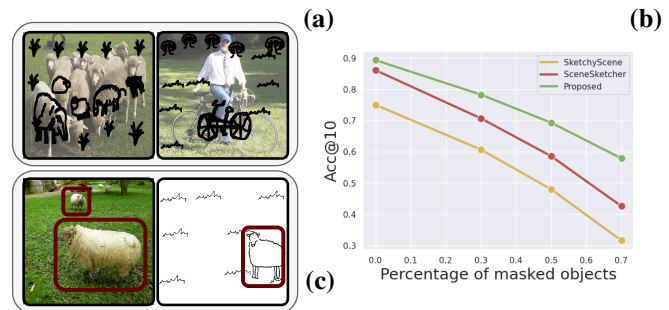

Figure 1. (a): Scene sketches exhibit abstraction on global scene configuration as shown by overlapping sketches on top of their corresponding photos. (b): Existing scene-level FG-SBIR methods collapse as scene sketches become more partial. (c): There are significant empty (white) regions. Also, the sketch of a sheep in the scene might correspond to that in the centre of photo. This calls for a solution modelling region-wise associativity.

fronts (i) a scene sketch does not necessarily contain all objects in the corresponding photo, due to subjective interpretation of scenes, i.e., holistically partial, and (ii) there exists significant empty (white) regions as a result of object-level abstraction, i.e., locally partial. This is verifiable through a quick pilot study on existing scene sketch dataset [24]: (i) scene sketches include an average of 49.7% of the objects present in photos. (ii) On average, only 13.0% of the area in any scene sketch is occupied by its individual objects (the other 87.0% being empty regions).

In this paper, we specifically tackle this "partial" problem in the context of scene-level fine-grained sketch-based image retrieval (Scene-Level FG-SBIR). We first confirm the prevalence of this problem in existing Scene-Level FG-SBIR models, where we conduct an empirical study by measuring the retrieval accuracy by progressively masking out individual objects in a scene sketch. Fig.1(b) shows that popular models [24, 90] collapse as scene sketches become more partial, calling for a solution that is robust towards partial input, and in turn yields state-of-the-art performance.

Global Average Pooling used by most existing FG-SBIR methods [4, 5] is clearly not up for the job since it is notorious for losing spatial scene configuration information.

---

*Interned with SketchX

A naive alternative is computing distances between pairs of local features from corresponding regions in sketch and photo. This however is sub-optimal since sketch and photo do not follow strict region-wise alignment (see Fig. 1). Alternatively, one can compute soft attentions independently in each domain [64], yet this largely ignores the cross-modal gap between sketch and photo. Using cross-modal co-attention [58,72] sounds a viable option but is otherwise intractable for practical applications[1]. A close contender to our approach is using graph-based matching [15, 49] of sketch and photo regions such as Liu *et al.* [44]. However, graph-based approaches have two common problems: (a) they dictate bounding-box annotations which are not always available (e.g., on [90]). (b) optimal graph construction (or sketching) strategy can be overly complicated [15, 19].

The key to solving this "partial" problem lies with modelling cross-modal region associativity. Crucially, such associativity needs to happen in a partially-aware fashion. This is because most sketch regions being empty will not be matched to any part of the photo. Partial graph matching is possible [19,68] but again not without resorting to expensive bounding-box annotations and complex scene graph construction procedures. Instead, as our first contribution, we advocate for the use of classic transportation theory (e.g., optimal transport (OT) [11]) to model this region associativity. Set-based approaches are a great fit because (i) they do not require any explicit data annotation, and (ii) naturally tackle this partial matching problem [15, 80][2].

While using OT can already model region associativity, it does not yet account for holistic partialness, i.e., differences in scene configurations. This dictates a holistic mechanism that accounts for spatial object relationships within each modal. Thus, as our second contribution, we improve upon OT by capturing intra-modal scene configurations for either modality in their respective region adjacency matrix [86]. It follows that during cross-modal comparison, we compute the differences between these two matrices and obtain a scalar for each corresponding region to use alongside OT. Simply dotting together the intra-modal adjacency matrices is however not ideal since it ignores the local partialness of sketches which results in lots of near-zero entries in the adjacency matrix, ultimately leading to an overly sparse cross-modal matrix when dotted. Instead, we perform a weighted comparison by computing the cosine distance of two region-pairs, each pair taken from sketch and photo modality respectively.

In summary, our contributions are: (i) We show a significant portion of scene-level sketches are "partial", both *holistically* and *locally*. (ii) The prevalence of this problem in existing FG-SBIR models is confirmed by showing how popular models [44, 90] collapse as scene sketches become more partial. (iii) We propose a simple solution to this "partial" problem by modelling partially-aware cross-modal region associativity using the classic transportation problem (optimal transport (OT) [11]). (iv) We improve upon OT by capturing one-to-one intra-modal spatial relationships for partial scene sketches. (v) Our method is not only robust to partial input scene sketches but also yields state-of-the-art performance on existing scene sketch datasets.

2. Related Work

Fine-grained sketch based image retrieval The ability of sketches to offer inherently fine-grained visual descriptions commences the avenues of fine-grained sketch-based image retrieval (FG-SBIR) [4, 7, 8, 52, 53, 57, 62]. This meticulous task aims to learn a pair-wise correspondence, for instance-level sketch-photo matching. Starting with graph-matching of deformable-part models [40], several deep learning approaches have surfaced with the advent of FG-SBIR datasets [63, 64]. Yu *et al.* [78] proposed a deep triplet-ranking model for instance-level matching. This framework was subsequently enhanced through a hybrid generative-discriminative cross-domain image generation [53], providing an attention based mechanism with advanced higher order retrieval loss [64], utilising textual tags [63], or pre-training strategy [5, 54]. Unlike existing frameworks in FG-SBIR [63, 78] that independently map sketch and photo to a joint embedding sketch-photo space, Sain *et al.* [58] introduced a cross-modal co-attention mechanism for FG-SBIR to give considerable improvements in retrieval accuracy. Despite offering unmatched retrieval performance, it is often inapplicable in practice for large-scale retrieval, given the fact that every gallery photo needs to be compared with the query sketch every time for new retrieval. In this work, we push for a novel distance-metric function that would work at the output of independent sketch/photo branch, and models the region-wise associativity without any costly [58] pair-wise feature matching at the intermediate convolutional feature-map level. Our retrieval framework can be thought of as *"best of both worlds"* i.e., *fast-* [78] and *slow-* [58] retrieval [26, 50]. In other words, we perform region-wise feature matching but still sketch/photo branch could be computed independently, unlike the necessity of repeated gallery image feature computation as in cross-modal co-attention mechanism [58].

Scene-level sketches As research in object-level (fine-grained) sketch-based image retrieval matured, recent works took the natural step towards the more practical but less explored setup of *scene*-level for deeper and richer reasoning about sketched visual forms [24, 44, 90]. Zou *et al.* [90] studied segmentation and colorization on scene-level sketches. Gao *et al.* [24] proposed scene-sketch to photo generation via generative adversarial approach using a sequential two-stage module. While Liu *et al.* [44] introduced FG-SBIR for scene-sketches using graph convo-

---

[1]See Appendix for further discussion

[2]Please refer to [15] for a detailed proof.

lutional networks, it largely avoided the challenging setup of partial sketches by filtering existing datasets [24] having too few foreground instances (i.e., partial sketches). Unlike [44], we propose a scene-level FG-SBIR setup which is robust to the more realistic setup of *partial* sketches that lacks aligned, position-wise correspondence between instance-level sketch-photo pairs.

Dealing with Partial Data One of the prolific areas studying incomplete or partial data is image inpainting [71, 85, 89], where the objective is to generate (or fill) the missing (or masked) region by conditioning on the overall region. In the context of sketches, there are two broad lines of work: (a) two-stage pipeline that first tries to complete the partial sketch [30, 65] by modelling a conditional distribution based on image-to-image translation followed by performing task specific objective such as recognition [43] or sketch-to-image [28] generation. (b) single step framework [9] directly handle incomplete sketches to perform task specific objectives in a single step. Similar to Bhunia *et al*. [9], ours is a single step framework capable of handling incomplete or partial sketches. While existing literature [9, 43] investigates object-level partial sketches, we focus on the novel setup of scene-level retrieval.

Application of Optimal Transport The ability to learn structural similarity *without explicit alignment information* makes optimal transport [70] in linear optimisation an important tool for several downstream tasks [35, 38, 45, 61, 84]. Rubner *et al*. [56] employed earth mover's distance, having the formulation of transportation problem, as a metric for color and texture based image retrieval. Later works extended optimal transport to the deep learning landscape for applications such as document classification [33], few-shot learning [80], domain adaptation [17], self-supervised learning [45], neural machine translation [75], and comprehending scene using 3D point cloud from monocular data [35]. In this work, for the first time, we study the application of optimal transport to design a differentiable distance metric function to model region-wise associativity without any explicit alignment label, and train a cross-modal retrieval system end-to-end through triplet-ranking objective.

Learning unsupervised region-wise correspondence Matching same or similar structure/content from two or more input data is a fundamental task in various downstream applications [47, 81] such as image stitching [10], image fusion [48], co-segmentation [23, 76], image retrieval [88], object recognition [73] and tracking [74]. Region-wise correspondence could be learned in a supervised [69, 77, 83], self-supervised [18, 39, 82], or unsupervised manner [3, 27, 34, 51, 60]. Self-supervised and unsupervised methods train without any human annotations and only use geometric [31] or semantic [46] constraints. In contrast to these works, to the best of our knowledge, we here first try to bring the power of unsupervised region-wise associativity for cross-
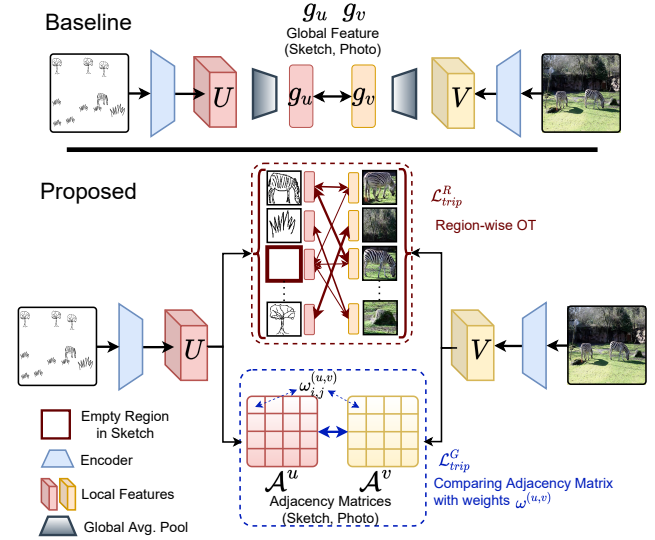

Figure 2. Illustration of our proposed method for scene-level FG-SBIR. Existing baselines typically use Global Average Pooling (GAP) on the convolutional feature-map. This loses localised region-specific feature representation necessary for "partial" scene sketches. Our proposed method models partially-aware region-wise associativity to solve this "partial" problem using: (i) set-based distance of local feature maps using optimal transport ($\mathcal{L}_{trip}^R$), (ii) a weighted cross-modal comparison of region adjacency matrix to capture holistic scene configuration ($\mathcal{L}_{trip}^G$).

modal setup in an end-to-end trainable framework.

3. Proposed Methodology

Our objective is to retrieve the scene-image(s) which satisfy the partial sketch query constraint. The existing global feature vector-based representation, usually obtained through Global Average Pooling [42], ignores this *partial associativity constraint*, thus failing to retrieve scene images from partial sketch faithfully. This work aims to model partial associativity by comparing corresponding region-wise localised features between every sketch-photo pair through a novel end-to-end trainable metric-learning loss so that the model can retrieve even from partial sketches.

3.1. Baseline Retrieval Framework

First, we briefly summarise a baseline retrieval framework that remains state-of-the-art for SBIR literature to date. Given a sketch-photo pair represented as $(\mathcal{S}, \mathcal{I})$, a feature extractor $\mathcal{F}_\theta(\cdot)$ parameterised by θ is used to get the feature map $U = \mathcal{F}_\theta(\mathcal{S}) \in \mathbb{R}^{h_S \times w_S \times c}$ and $V = \mathcal{F}_\theta(\mathcal{I}) \in \mathbb{R}^{h_I \times w_I \times c}$ for sketch and photo, respectively. $\mathcal{F}_\theta$ can be modelled by CNN [78], LSTM [36], Transformer [41], Graphs [44, 55, 65] or their combinations [5]. We flatten the backbone output feature-map $U \in \mathbb{R}^{h_S \times w_S \times c}$ and $V \in \mathbb{R}^{h_I \times w_I \times c}$ for sketch and photo as: $\mathbf{u} = \{u_1, \ldots, u_m\}$ and $\mathbf{v} = \{v_1, \ldots, v_n\}$, respectively, where $m = h_s w_s$, $n = h_I w_I$, and $u_i, v_i \in \mathbb{R}^c$. Every vector of size $\mathbb{R}^c$

from either of **u** or **v** represents a localised regions specific feature. In order to get a single feature representation, we apply Global Average Pooling (GAP) and get $g_u = \frac{1}{m} \sum_{i=1}^{m} u_i \in \mathbb{R}^c$ and $g_v = \frac{1}{n} \sum_{j=1}^{n} v_j \in \mathbb{R}^c$ for sketch and photo, respectively. For training, the distance $d(\cdot, \cdot)$ to a sketch anchor $\mathcal{S}$ from a negative photo $\mathcal{I}^-$, denoted as $\beta^- = d(g_u, g_{v^-})$ should increase while that from the positive photo $\mathcal{I}^+$, $\beta^+ = d(g_u, g_{v^+})$ should decrease. $d(a, b)$ can be either euclidean or cosine distance, but we consider dot product based cosine distance $(1 - a \cdot b)$ where $\{a, b\}$ are pre-normalised so that $||a||_2 = 1, ||b||_2 = 1$. Training is done via triplet loss with hyperparameter $\mu > 0$:

$$\mathcal{L}_{trip} = \max\{0, \mu + \beta^+ - \beta^-\} \qquad (1)$$

There are some inherent limitations of this standard baseline. *Firstly*, applying GAP on the convolutional feature-map loses any localised region-specific feature representation, leaving no chance to learn region-wise associativity. *Secondly*, region-wise associativity is a latent or hidden knowledge for which we do not have any explicit label, and this hidden knowledge is ignored here. *Thirdly*, it assumes that every paired sketch is perfectly annotated containing *all* the salient concepts/objects from the paired photo. However, in reality, most of the annotated sketches are partial and is biased towards the annotator's drawing skill and perception. Global feature vector-based representation will unnecessarily penalise mismatches for partial sketches. Therefore, this demands a further investigation on how to design a metric loss [58,78] that would implicitly discover the hidden region-wise associativity from partials sketches so that it generalises to query sketch with any degree of a partial instance during inference.

3.2. Towards Partial Associativity

Reinterpretation of Baseline: The cosine distance between global sketch feature vector $g_u \in \mathbb{R}^c$ and photo feature vector $g_v \in \mathbb{R}^c$ in our baseline could be reinterpreted as taking an average over all cosine distances computed between every localised region specific feature u_i (sketch) and v_j (photo). Formally we can write $d(g_u, g_v)$ as follows,

$$d(g_u, g_v) = (1 - g_u \cdot g_v) = 1 - (\frac{1}{m} \sum_{i=1}^{m} u_i) \cdot (\frac{1}{n} \sum_{j=1}^{n} v_j)$$
$$= \frac{1}{m\,n} \sum_{i=1}^{m} \sum_{j=1}^{n} (1 - u_i \cdot v_j) = \frac{1}{m\,n} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{i,j} \qquad (2)$$

Weighted region-wise distance: This simple average operation gives equal weightage to every pair-wise distance $c_{i,j}$ instead of prioritising those which actually have similar semantic meaning. Please see Fig. 2 for a visual illustration. Therefore, to model *region-wise associativity* for measuring the distance between partial scene sketch-photo pairs, we extend the naive averaging of cosine distances (as in Eq. 2) to a *weighted region-wise cosine distance*:

$$d_W(\mathbf{u}, \mathbf{v}) = \frac{1}{m\,n} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{i,j} x_{i,j} \qquad (3)$$

We aim to learn the weights $x_{i,j}$ representing associativity between each pair of the localised feature from **u** (sketch) and **v** (photo) feature set, respectively. In other words, we compute all the pair-wise distances but give more weightage to those having similar semantics. We do not have any explicit labels for $\mathcal{X} \in \mathbb{R}^{m \times n}$, and to model this latent knowledge, we take inspiration from Optimal Transport [2] literature. Here, $x_{i,j}$ is termed as "flow" from the source sketch region u_i to destination photo region v_j. The task of finding the optimal flow $x_{i,j}$ for the given $c_{i,j}$ is similar to the classic transportation problem (TP) [11]. Overall, our objective is to design a weighted region-wise distance metric for a partial sketch to photo matching.

Optimal Transport: In classical transportation problem, m suppliers S $= \{s_i | i = 1, \dots, m\}$ are required to supply n demanders D $= \{d_j | j = 1, \dots, n\}$. The cost of transportation $c_{i,j}$ for a unit goods between the ith supplier and j^{th} demander is $c_{i,j} = (1 - u_i \cdot v_j)$. The optimisation objective of TP is to find the least expensive *"flow"* of goods from suppliers to demanders, represented by $\tilde{\mathcal{X}} \in \mathbb{R}^{m \times n}$. Optimising $\tilde{\mathcal{X}}$ is analogous to our aim of prioritising those region-wise distances which are semantically similar, thus modelling pair-wise associativity from two feature set **u** and **v**. TP objective is written as:

$$\underset{\mathcal{X}}{\text{minimize}} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{i,j} x_{i,j}$$
$$\text{subject to } x_{i,j} \geq 0, \; i = 1, \dots, m, \; j = 1, \dots, n \qquad (4)$$
$$\sum_{j=1}^{n} x_{i,j} = s_i, \; \sum_{i=1}^{m} x_{i,j} = d_j; \; i = [1, m] \; j = [1, n]$$

The total flow from the ith region in a query sketch to every region in the target photo is represented by s_i, calculated as: $s_i = \sum_{j=1}^{n} x_{i,j}$. Similarly, d_j denotes the total flow from all m regions in the query sketch to the jth region in the target photo, computed as $d_j = \sum_{i=1}^{m} x_{i,j}$. This optimisation problem (Eq. 4) falls in the category of linear programming [11] as the objective and constraints are all affine and can be solved using the classical interior-point method. A naive solution to TP is intractable [1] . Therefore, to bring the power of optimal transport for modeling region-wise partial associativity in case of partial sketch-based scene retrieval, we need a differentiable solution for end-to-end training.

Differentiability of the Solution: Our objective is to make the flow $\mathcal{X}$ differentiable with respect to model parameters θ. Towards that goal, we rewrite Eq. 4 in a parametric convex optimisation [2] form involving model parameter θ as:

$$\underset{\mathcal{X}}{\text{minimize}}\ c(\theta)^T \mathcal{X} \tag{5}$$
$$\text{subject to } f(\mathcal{X}, \theta) \preceq 0; \ \ h(\mathcal{X}, \theta) = b(\theta)$$

where, $f(\mathcal{X}, \theta)$ is equivalent to the inequality constraint $x_{i,j} \geq 0$, and $h(\mathcal{X}, \theta) = b(\theta)$ is the equality constraint equivalent to $\sum_{j=1}^{n} x_{i,j} = s_i$ and $\sum_{i=1}^{m} x_{i,j} = d_j$ for all $i = 1, \ldots, m; j = 1, \ldots, n$.

In order to combine Eq. 5, involving three fragmented parts (one optimisation objective, one equality, and one inequality constraints) into a single differentiable equation, we augment the objective function with a weighted sum of its equality and inequality constraints, defined as Lagrangian:

$$\mathrm{L}(\mathcal{X}, \lambda, \nu, \theta) = \ c(\theta)^T \mathcal{X}$$
$$+ \lambda^T f(\mathcal{X}, \theta) + \nu^T (h(\mathcal{X}, \theta) - b(\theta)) \tag{6}$$

The vectors $\lambda \geq 0$ and ν are the dual variables or Lagrange multiplier vectors. From Eq. 6 we are looking for the optimal value of $\{\tilde{\mathcal{X}}, \tilde{\lambda}, \tilde{\nu}\}$ which corresponds to the minimum possible scalar value returned by $\mathrm{L}(\mathcal{X}, \lambda, \nu, \theta)$. For Eq. 6, assuming Slater's condition holds, then the necessary and sufficient conditions for optimality of $\mathrm{L}(\mathcal{X}, \lambda, \nu, \theta)$ is given by the Karush-Kuhn-Tucker (KKT) conditions [11], which can be algebraically represented as:

$$g(\tilde{\mathcal{X}}, \tilde{\lambda}, \tilde{\nu}, \theta) = \begin{bmatrix} \nabla_{\mathcal{X}} \mathrm{L}(\tilde{\mathcal{X}}, \tilde{\lambda}, \tilde{\nu}, \theta) \\ \mathbf{diag}(\tilde{\lambda}) f(\tilde{\mathcal{X}}, \theta) \\ h(\tilde{\mathcal{X}}, \tilde{\theta}) - b(\theta) \end{bmatrix} = 0 \tag{7}$$

where $\mathbf{diag}(\cdot)$ transforms a vector into diagonal matrix. The Jacobian $\nabla_\theta(\tilde{\mathcal{X}})$, provides the differentiation of our region-wise associativity $\mathcal{X}$ with respect to model parameters θ to allow end-to-end training. It can be derived from Eq. 7 using Dini classical implicit function theorem [20] as,

$$\nabla_\theta(\tilde{\mathcal{X}}) = -\nabla_{\mathcal{X}} g(\tilde{\mathcal{X}}, \theta)^{-1} \nabla_\theta g(\tilde{\mathcal{X}}, \theta) \tag{8}$$

Determining the equality constraints s_i and d_j: Earlier we formulated a differentiable way to calculate region-wise associativity $\mathcal{X}$, but we need to define the two important parameters s_i and d_j of Eq. 4. Our assumption is that summation of all the region-wise associativity values should be equal to the cosine similarity between the global feature vector representation of given sketch (g_u) and photo (g_v). Note, while we constrain the total (sum) value of $\mathcal{X} \in \mathbb{R}^{m \times n}$, the model is free to decide how to distribute the total value to individual region-wise associativity ($x_{i,j}$) to achieve optimality. Given Global Average Pooled vector $g_u = \frac{1}{m} \sum_{i=1}^{m} u_i$ and $g_v = \frac{1}{n} \sum_{j=1}^{n} v_j$ from sketch and photo, respectively, the summation over all $x_{i,j}$ (total value) can be formally written as:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_{i,j} = g_u \cdot g_v = \left(\frac{1}{m} \sum_{i=1}^{m} u_i\right) \cdot \left(\frac{1}{n} \sum_{j=1}^{n} v_j\right)$$
$$= \frac{1}{m\,n} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} u_i \cdot v_j\right) \text{ (distributive property)} \tag{9}$$

Therefore, following equality constraints in Eq. 4 and ignoring the constant $\left(\frac{1}{mn}\right)$

$$\sum_{j=1}^{n} x_{i,j} = \sum_{j=1}^{n} u_i \cdot v_j \implies s_i = u_i \cdot \sum_{i=1}^{n} v_j \tag{10}$$

similarly, $d_j = v_j \cdot \sum_{i=1}^{m} u_i$ Hence, our modified distance-metric function $d_W(\cdot, \cdot)$, which measures the scalar distance between two feature sets, modelling region-wise associativity between sketch and photo, is computed as:

$$d_W(\mathbf{u}, \mathbf{v}) = \frac{1}{m\,n} \sum_{i=1}^{m} \sum_{j=1}^{n} (1 - u_i \cdot v_j) \tilde{x}_{i,j}$$
$$\mathcal{L}_{trip}^{R} = \max\{0, \mu_w + \beta_R^{+} - \beta_R^{-}\} \tag{11}$$

where, $\beta_R^{+} = d_W(\mathbf{u}, \mathbf{v}^{+})$ is the distance to a sketch anchor $\mathcal{S}$ to positive photo $\mathcal{I}^{+}$. Similarly, we calculate β_R^{-} for negative photo $\mathcal{I}^{-}$. μ_w is margin hyperparameter.

3.3. Preserving scene structure consistency

While optimal transport helps to model localised region-wise associativity to measure the distance between sketch-photo pairs, it cannot preserve the global structural consistency [68] necessary for fine-grained retrieval. Region-wise associativity may fail to distinguish between scene sketches where individual objects are similar, but the global spatial arrangement is different. For instance, moving a sketched "tree" from the top-left corner to the bottom right will result in a similar distance value. Therefore, we argue that while $d_W(\cdot)$ matches the features at the local level it is suboptimal at considering the global spatial information.

We design the global structural consistency through *adjacency matrix* formulation which captures spatially-correlative maps to explicitly represent the global scene structure. Given sketch and photo feature set $\mathbf{u} = \{u_1, \ldots, u_m\}$ and $\mathbf{v} = \{v_1, \ldots, v_n\}$ respectively, we compute their respective adjacency matrix as:

$$\mathcal{A}_{i,j}^{u} = \frac{1}{m \times m} \frac{u_i \cdot u_j}{||u_i||_2 ||u_j||_2}; \ \mathcal{A}^{u} \in \mathbb{R}^{m \times m}$$
$$\mathcal{A}_{i,j}^{v} = \frac{1}{n \times n} \frac{v_i \cdot v_j}{||v_i||_2 ||v_j||_2}; \ \mathcal{A}^{v} \in \mathbb{R}^{n \times n} \tag{12}$$

Our adjacency matrix essentially computes region-wise self-similarity in sketch and photo modality. Naively comparing $(\mathcal{A}^{u}, \mathcal{A}^{v})$ assumes that *all* regions within a particular modality (sketch or photo) contribute towards self-similarity [86]. However, this assumption does not hold for partial scene sketches with empty (sparse) or uncorrelated regions. Hence, we introduce a weighting factor $\omega_{i,j}^{u,v}$ that provides lower importance while comparing $(\mathcal{A}_{i,j}^{u}, \mathcal{A}_{i,j}^{v})$ if there is a low correlation between either of (u_i, v_i), (u_i, v_j), (u_j, v_i), or (u_j, v_j). The distance function capturing global structural consistency is as follows:

$$d_G(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{m} \sum_{j=1}^{m} \omega_{i,j}^{(u,v)} ||\mathcal{A}_{i,j}^u - \mathcal{A}_{i,j}^v||_1 \quad (13)$$

$$\omega_{i,j}^{(u,v)} = (u_i \cdot v_i)(u_i \cdot v_j)(u_j \cdot v_i)(u_j \cdot v_j)$$

if $m \neq n$, we make use of bi-linear interpolation [87] to make $\mathcal{A}^v$ to be of same spatial size as $\mathcal{A}^u$ to realize matrix subtraction. Intuitively, if either of the localised partial sketch feature (u_i, u_j) be a sparse region, it will have a lower correlation with the localised photo feature set (v_i, v_j) that will give a low value of $\omega_{i,j}^{(u,v)}$ in the distance function $d_G(\cdot, \cdot)$. Such a mechanism ignores empty and uncorrelated regions in the partial sketch while focusing only on the relevant regions. Hence, we can define our final training objective as,

$$\mathcal{L}_{total} = \mathcal{L}_{trip}^R + \alpha \cdot \mathcal{L}_{trip}^G$$
$$\mathcal{L}_{trip}^G = \max\{0, \mu_g + \beta_G^+ - \beta_G^-\} \quad (14)$$

where, $\beta_G^+ = d_G(\mathcal{A}^u \mathcal{A}^{v^+})$, computed by Eqn. 13. Similarly we compute β_G^-. α, μ_g are hyperparameters.

4. Experiments

Datasets: We use two benchmark scene sketch datasets that support scene-level FG-SBIR tasks: (a) SketchyScene [90] comprise of sketch templates with paired cartoon style photos. Following Zou *et al.* [90], we adopt the standard train/test split of 2472/252 of sketch-photo pairs for scene-level FG-SBIR. On average each scene sketch has 16 instances, 6 object classes, and 7 occluded instances. (b) Unlike SketchyScene [90] that comprise of cartoon style photos, SketchyCOCO [24] includes natural photos from COCO Stuff dataset [14] with paired scene sketches. Following Liu *et al.* [44], we use $1015/210$ train/test split of sketch/photo pairs. In addition, we also evaluate how our proposed method generalise to object-level sketches in QMUL-Shoe-V2 [78]. QMUL-Shoe-V2 contains $6,730$ sketches and $2,000$ photos. Following [9] we use $6,051$ and $1,800$ respectively for training and the rest for testing.

Implementation Details: Our model is implemented in PyTorch on a 11GB Nvidia RTX 2080-Ti GPU. ImageNet pretrained InceptionV3 [66] (excluding auxiliary branch) is used as the encoder network $\mathcal{F}_\theta(\cdot)$ where we remove the global average pooling and flattening layers. We train our model for 200 epochs using Adam optimiser with learning rate 0.0001, batch size 16, margin value for triplet loss as 0.3, and value of α in Eq. 14 is set to 0.01. Our novel distance metrics for FG-SBIR do not add training parameters, hence the total number of model parameters stays the same as its backbone network. During training, to solve the linear programming problem in Eq. 7 we use a GPU accelerated *convex optimisation solver* QPTH proposed in Amos and Kolter [1] to derive gradients for backpropagation. Since gradients only needs to be computed during training, for

testing we replace the QPTH solver with a faster alternative in the OpenCV library [12] that is non-differentiable[3]. A PyTorch-like pseudo-code to solve the linear programming problem using QPTH and to compute region-wise associativity $x_{i,j}$ in Eq. 5 is provided in the Appendix.

Evaluation Metric: In line with FG-SBIR research, we use Acc.@q accuracy, i.e. percentage of sketches having true matched photo in the top-q list. To evaluate performance on the partial setup, we explicitly mask local regions in scene and object sketch regions, described later in Sec. 4.2 and 4.3 respectively. We repeat our evaluations 10 times and report the average metrics. Since our goal is to evaluate *robustness* to partial sketches, we train the network using the complete sketches but evaluate on *masked partial sketch*.

4.1. Competitors:

We compare against (i) existing state-of-the-art (SOTA) methods: **Triplet-SN** [78] employs Sketch-A-Net [79] backbone trained using triplet loss. **HOLEF** [64] extends *Triplet-SN* by adding spatial attention along with higher order ranking loss. **On-the-fly** [9] aims to model partial sketches by employing reinforcement learning (RL) based fine-tuning for on-the-fly retrieval on partial object sketches. Since it requires vectored sketch data unavailable in existing scene sketch datasets [24, 90], we compare with *On-the-fly* on object sketches [78] using vectored data. As early retrieval is not our objective, we cite result at sketch completion point. **SketchyScene** is the first work on scene-level FG-SBIR that adopts *Triplet-SN* but replace the base network from Sketch-A-Net to InceptionV3 [66] along with an auxiliary cross-entropy loss to utilise the available object category information. **SceneSketcher** [44] is a recent work that use graph convolutional network to model the scene sketch layout information. (ii) Since local-level features are largely ignored by SOTAs, we design a few Baselines that computes the distance between localised sketch and photo features: **Local-Align** use a naive approach of computing cosine distance between a pair of local features at the same location of feature map from sketch and photo. However, sketches and photo do not follow strict region-wise alignment. **Local-MIL** on the other hand, utilise multiple instance learning (MIL) paradigm [13] where the minimum cosine distance pair between the set of localised sketch and photo is considered in loss computation. While it overcomes the limitation of misaligned localised features in *Local-Align*, MIL is unstable since it leaves the remaining pairs of localised features under-constrained. **Local-Self-Atten** injects global context to local patches using self-attention mechanism [67] to aggregate contextual information followed by computing the cosine distance between lo-

---

[3]The differentiable QPTH solver in Amos and Kolter [1] use a custom primal-dual interior point method which is slower compared to the non-differentiable solver in OpenCV employing a modified Simplex algorithm.

Table 1. Scene-level Fine-Grained SBIR on SketchyScene.

	Method		Complete Sketch	p_{mask} 0.3	p_{mask} 0.5
SOTA	Triplet-SN [78]	Acc.@1	4.5	<0.1	<0.1
		Acc.@10	26.7	6.9 ± 0.5	3.7 ± 0.7
	HOLEF [64]	Acc.@1	5.3	<0.1	<0.1
		Acc.@10	29.5	7.5 ± 0.3	3.7 ± 0.7
	SketchyScene [90]	Acc.@1	32.2	8.1 ± 0.7	4.5 ± 0.6
		Acc.@10	69.3	24.7 ± 0.3	16.8 ± 1.3
Baselines	Local-Align	Acc.@1	32.7	8.3 ± 0.4	4.9 ± 0.7
		Acc.@10	70.1	31.6 ± 0.7	19.7 ± 1.2
	Local-MIL	Acc.@1	33.4	9.7 ± 0.3	5.5 ± 0.6
		Acc.@10	71.9	35.3 ± 0.5	21.2 ± 1.1
	Local-Self-Attn	Acc.@1	33.6	9.7 ± 0.5	5.9 ± 0.6
		Acc.@10	72.5	37.2 ± 0.8	21.7 ± 0.9
Variant	Ours-Local-OT	Acc.@1	34.9	15.5 ± 0.3	8.1 ± 0.8
		Acc.@10	74.5	45.3 ± 0.4	29.3 ± 0.7
	Ours-Local-MMD	Acc.@1	34.4	14.5 ± 0.2	7.5 ± 0.7
		Acc.@10	73.4	43.8 ± 0.5	25.9 ± 1.1
	Proposed Method	Acc.@1	35.7	20.6 ± 0.3	10.6 ± 0.9
		Acc.@10	75.1	49.2 ± 1.1	29.9 ± 1.2

Table 2. Scene-level Fine-Grained SBIR on SketchyCOCO.

	Method	Acc.@q	Complete Sketch	p_{mask} 0.3	p_{mask} 0.5
SOTA	Triplet-SN [78]	Acc.@1	6.2	<0.1	<0.1
		Acc.@10	32.8	24.2 ± 0.9	18.5 ±1.1
	HOLEF [64]	Acc.@1	6.8	<0.1	<0.1
		Acc.@10	35.9	25.3 ± 0.7	19.3 ± 1.5
	SketchyScene [90]	Acc.@1	27.6	19.7 ± 1.1	13.9 ± 1.4
		Acc.@10	75.0	60.7 ± 1.2	48.0 ± 1.8
	SceneSketcher [44]	Acc.@1	31.7	23.5 ± 1.5	17.2 ± 1.2
		Acc.@10	86.1	70.7 ±0.9	57.7 ± 1.3
Baselines	Local-Align	Acc.@1	31.9	23.6±1.4	17.4±1.1
		Acc.@10	86.6	70.7 ±1.1	57.9±1.2
	Local-MIL	Acc.@1	32.5	23.7±1.2	17.7±1.2
		Acc.@10	87.8	71.1 ±0.9	58.5±1.3
	Local-Self-Attn	Acc.@1	33.1	23.9±1.1	18.1±1.4
		Acc.@10	88.7	71.8 ± 0.7	59.1±1.1
Variant	Ours-Local-OT	Acc.@1	34.3	24.7±1.2	18.7±1.5
		Acc.@10	89.2	75.6 ± 1.1	65.7±1.2
	Ours-Local-MMD	Acc.@1	33.9	24.2±0.9	18.5±1.3
		Acc.@10	89.1	74.6±0.9	63.2±1.5
	Proposed Method	Acc.@1	34.5	25.1 ± 1.9	19.2 ± 1.4
		Acc.@10	89.3	78.3 ± 1.6	69.3 ± 1.7

Table 3. Object-level Fine-Grained SBIR on QMUL-Shoe-V2.

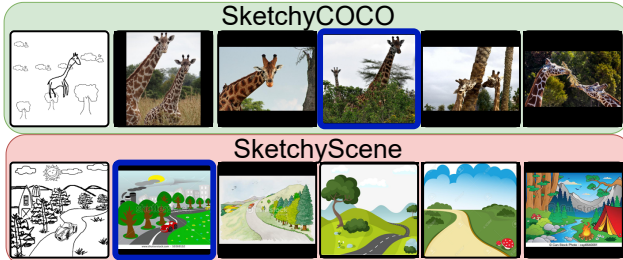|	Method	Acc.@q	Complete Sketch	p_{mask} 0.3	p_{mask} 0.5
SOTA	Triplet-SN [78]	Acc.@1	28.7	22.3±0.4	9.7±0.9
		Acc.@10	79.6	73.5±0.3	67.1±0.5
	HOLEF [64]	Acc.@1	31.2	24.6±0.6	12.9±1.0
		Acc.@10	81.4	75.1±0.5	68.4±0.9
	On-the-fly [9]	Acc.@1	34.1	29.5 ± 0.5	20.9 ± 0.9
		Acc.@10	79.6	76.3±0.3	71.9±1.2
Baselines	Local-Align	Acc.@1	33.5	25.7±0.2	14.9±0.7
		Acc.@10	79.6	75.6±0.3	69.5±0.8
	Local-MIL	Acc.@1	35.5	29.9±0.1	21.0±0.9
		Acc.@10	80.6	79.1±0.5	71.3±1.1
	Local-Self-Attn	Acc.@1	37.1	31.6±0.3	21.7±0.7
		Acc.@10	81.4	79.5±0.1	71.5±0.5
Variant	Ours-Local-OT	Acc.@1	39.7	34.7±0.3	25.7±1.0
		Acc.@10	82.9	80.5±0.1	73.4±0.5
	Ours-Local-MMD	Acc.@1	38.2	33.6±0.4	24.3±0.6
		Acc.@10	82.5	79.7±0.2	73.3±0.5
	Proposed Method	Acc.@1	39.9	35.3±0.2	25.9±0.7
		Acc.@10	82.9	80.9±0.1	73.4±0.7

Figure 3. Qualitative top-5 retrieved results on SketchyCOCO [24] and SketchyScene [90] using our proposed method. Blue denotes ground-truth photo.

calised feature maps as *Local-Align*. (iii) Naively computing a distance between localised features from sketch and photo fails to capture the geometry of the underlying localised feature space [22]. Hence, we design two Variants of our proposed method: **Ours-Local-MMD** removes region-wise associativity from our proposed method and replace optimal transport with maximum mean discrepancy (MMD) to compare localised sketch and photo features that take into account the geometry of the underlying feature space. **Ours-Local-OT** replace MMD with the more accurate [22] optimal transport to compute region-wise associativity between localised features.

4.2. Evaluation on partial scene sketches

We perform a comparative study on scene sketches from SketchyScene [90] and SketchyCOCO [24] datasets. Our experimental setup includes: Complete Sketch that evaluates on the original input scene sketch. $p_{mask} = 0.3$ use instance segmentation map of scene sketches available in both datasets to mask 30% sketched objects in a scene. Similarly, $p_{mask} = 0.5$ mask 50% sketched objects respectively.

Performance Analysis: From Tab. 1 and Tab. 2 we make the following observations: (i) Performance of all SOTAs degrade significantly when increasing p_{mask} from 0.3 to 0.5. This verifies our intuition that using a global feature vector is sub-optimal for the partial scene sketch setup. (ii) Our Baselines using local features are more resilient than SOTAs for partial scene sketches setup. It justifies the need for modelling localised features in partial scene sketches. However, abstract scene sketches and photos do not have a strict spatial alignment of localised regions assumed by *Local-Align*. It explains the inferior performance of *Local-Align* in comparison with other Baselines. *Local-MIL* considers only the minimum distance pair from a set of localised features in sketch and photo for loss computation but leaves the other pairs *unconstrained*. This leads to instability during training which explains the inferior performance than *Local-Self-Atten*. (iii) Both *Ours-Local-MMD* and *Ours-Local-OT* outperforms Baselines due to their abil-

ity to capture geometric information in the underlying lo-calised feature space. Performance of *Ours-Local-OT* is slightly superior to *Ours-Local-MMD* due to the better accuracy of optimal transport in modelling the underlying geometry of localised feature space. Finally, our proposed method, equipped with optimal transport for region-wise associativity and weighted affinity matrix for scene structure consistency outperforms all competitors. Fig. 3 shows qualitative retrieval results on scene-sketch datasets.

4.3. Evaluation on partial object sketches

Most SOTAs in FG-SBIR were developed focusing on object-level sketches [30, 59, 78]. Hence for a fair comparison and to investigate the generalisation of our method to partial object-level sketches, we compare FG-SBIR using QMUL-Shoe-V2 [78] dataset. Our experimental setup comprise of: Complete Sketch that use the original object sketch. For $p_{mask} = 0.3$, $p_{mask} = 0.5$ we mask 30% and 50% of the strokes respectively. We derive strokes using the available point coordinate and pen state information.

Performance Analysis: From Tab. 3, we observe: (i) With the exception of *On-the-fly*, performance of all SOTAs degrade for partial object sketch, as more strokes are masked from $p_{mask} = 0.3$ to $p_{mask} = 0.5$. High performance of *On-the-fly* among SOTAs is due to its reinforcement learning based fine-tuning that takes into account the complete episode of progressive complete sketch from $p_{mask} = 0.5$ to Complete Sketch before updating the weights. This provides a more principled and practically reliable approach to modelling partial object sketches. However, its RL based unstable training leads to inferior performance than our Baselines. (ii) The performance gap between SOTAs and Baselines narrows down from that observed for scene sketches. In particular, *On-the-fly* using global average pooling but explicitly trained on episodes of partial sketches outperforms *Local-Align* trained on Complete Sketches and naively computes the distance between a pair of local sketch and photo features at the same location. This suggests that for object-level sketch, a carefully trained global feature can surpass naively training localised features. (iii) A lower performance gap between *Ours-Local-OT* and the proposed method for object sketches when compared to that of scene sketch show the effect of weighted affinity matrix dilutes for the simpler scenario of object sketches.

4.4. Ablation

Significance of Adjacency Matrix: We quantitatively evaluate the significance of adjacency matrix in Tab. 4 where including weighted adjacency matrix (W-Adj) improves scene-level FG-SBIR Acc.@1 performance by 1.0% and 1.4% over *SketchyScene* in complete and partial sketch ($p_{mask} = 0.3$) respectively.

Significance of weighting factor $\omega_{i,j}^{u,v}$**:** Partial sketches, by

Table 4. Ablative study on SketchyScene: *(w/o)-Region-wise Associativity* (RwA), using Naive Adjacency matrix (N-Adj), Weighted Adjacency (W-Adj) from Eq. 13.

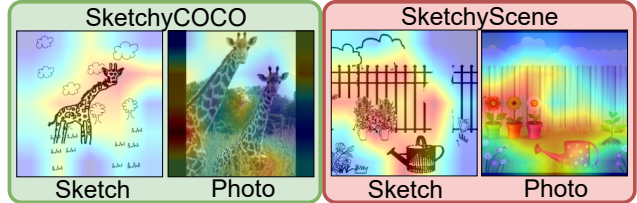| RwA	N-Adj	W-Adj	Complete Sketch	p_{mask} 0.3	Test time
✗	✗	✗	32.2	8.1	11.9ms
✓	✗	✗	34.9	15.5	12.4ms
✗	✓	–	32.9	8.9	12.0ms
✗	–	✓	33.2	9.5	12.0ms
✓	–	✓	35.7	20.6	12.4ms

Figure 4. Illustrating region-wise associativity between "partial" scene sketch and photo from SketchyCOCO [24] and SketchyScene [90]. For visualisation, we select the maximum value of region-wise associativity $(x_{i,j})$ from a local sketch region (u_i) to all photo regions $\{v_j\}_{j=1}^{n}$, and vice-versa.

definition, contain empty local regions. Naively computing the adjacency matrix (Eq. 12) would unnecessarily penalise and confuse the model. The weighting factor $\omega_{i,j}^{u,v}$ makes the model *aware* of empty regions by comparing (u_i, u_j) across modality with (v_i, v_j) in Eq. 13. From Tab. 4, adding naive adjacency matrix (N-Adj) enhance accuracy over *SketchyScene* by 0.7%, whereas a weighted adjacency matrix (W-Adj) gives 1.0% improvement. Fig. 4 highlights the corresponding sketch (u_i) with photo (v_j) regions.

Computation Time: From Tab. 4, we observe that including region-wise associativity (RwA) increases computation time by 0.5ms per iteration. Although insignificant compared to existing SOTA *Triplet-SN*, this added computation dramatically improves accuracy on complete and partial sketches. This slight increase in time is due to convex optimisation solvers is expected to diminish with faster and efficient solvers in the near future.

5. Conclusion

In this paper, we scrutinise an important observation plaguing scene-level sketch research – that a significant portion of scene sketches are "partial". We analyse how this partialness happens on two fronts: (i) significant empty (white) regions – locally partial, (ii) scene sketch does not necessarily contain all objects in the corresponding photo – holistically partial. For this, we propose a solution using a simple set-based approach using optimal transport to overcome local partialnes, and weighted cross-modal comparison of intra-modal adjacency matrices to address holistic partialness. Empirical evidence shows remarkable performance as a direct result of tackling this partial problem.

References

[1] Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *ICML*, 2017. 4, 6

[2] Shane Barratt. On the differentiability of the solution to convex optimization problems. *arXiv preprint arXiv:1804.05098*, 2019. 4

[3] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.net: Keypoint detection by hand-crafter and learned cnn filters. In *CVPR*, 2019. 3

[4] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *CVPR*, 2021. 1, 2

[5] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Vectorization and rasterization: Self-supervised learning for sketch and handwriting. In *CVPR*, 2021. 1, 2, 3

[6] Ayan Kumar Bhunia, Ayan Das, Umar Riaz Muhammad, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. Pixelor: A competitive sketching ai agent. so you think you can beat me? In *SIG-GRAPH Asia*, 2020. 1

[7] Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Subhadeep Koley, Rohit Kundu, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Doodle it yourself: Class incremental learning by drawing a few sketches. In *CVPR*, 2022. 2

[8] 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. 2

[9] 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, 3, 6, 7

[10] Moushumi Zaman Bonny and Mohammad Shorif Uddin. Feature-based image stitching algorithms. In *International workshop on computational intelligence*, 2016. 3

[11] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge Press, 2004. 2, 4, 5

[12] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc, 2008. 6

[13] Razvan C. Bunescu and Raymond J. Mooney. Multiple instance learning for sparse positive bags. In *ICML*, 2007. 6

[14] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018. 6

[15] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *ICML*, 2020. 2

[16] John Collomosse, Tui Bui, and Hailin Jin. Livesketch: Query perturbations for guided sketch-based visual search. In *CVPR*, 2019. 1

[17] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepj-dot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *ECCV*, 2018. 3

[18] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR*, 2018. 3

[19] Yihe Dong and Will Sawin. Copt: Coordinated optimal transport on graphs. In *NeurIPS*, 2020. 2

[20] Asen L. Dontchev and R. Tyrrell Rockafellar. *Implicit Functions and Solution Mappings*. Springer-Verlag, 2014. 5

[21] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 2012. 1

[22] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *AISTATS*, 2019. 7

[23] Huazhu Fu, Dong Xu, Stephen Lin, and Jiang Liu. Object-based rgbd image co-segmentation with mutex constraint. In *CVPR*, 2015. 3

[24] Chengying Gao, Qi Liu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: Image generation from free-hand scene sketches. In *CVPR*, 2020. 1, 2, 3, 6, 7, 8

[25] Songwei Ge, Vedanuj Goswami, C. Lawrence Zitnick, and Devi Parikh. Creative sketch generation. In *ICLR*, 2021. 1

[26] Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić, and Iryan Gurevych. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *arXiv preprint arXiv:2103.11920*, 2021. 2

[27] Georgios Georgakis, Srikrishna Karanam, Ziyan Wu, Jan Ernst, and Jana Kosecká. End-to-end learning of keypoint detector and descriptor for pose invariant 3d matching. In *CVPR*, 2018. 3

[28] Arnab Ghosh, Richard Zhang, Puneet K. Dokania, Oliver Wang, Alexei A. Efros, Philip H. S. Torr, and Eli Shechtman. Interactive sketch fill: Multiclass sketch-to-image translation. In *ICCV*, 2019. 3

[29] Yulia Gryaditskaya, Felix Hähnlein, Chenxi Liu, Alla Sheffer, and Bousseau. Lifting freehand concept sketches into 3d. In *SIGGRAPH Asia*, 2020. 1

[30] David Ha and Douglas Eck. A neural representation of sketch drawings. In *ICLR*, 2018. 3, 8

[31] Oshri Halimi, Or Litany, Emanuele Rodola, Alex Bronstein, and Ron Kimmel. Unsupervised learning of dense shape correspondence. In *CVPR*, 2019. 3

[32] Umar Riaz Huhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Learning deep sketch abstraction. In *CVPR*, 2018. 1

[33] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From work embeddings to document distances. In *ICML*, 2015. 3

[34] Karel Lenc and Andrea Vedaldi. Learning covariant feature detectors. In *ECCV*, 2016. 3

[35] Rémy Leroy, Pauline Trouvé, Frederic Champagnat, Bertrand Le Saux, and Marcela Carvalho. Pix2point: Learning outdoor 3d using sparse point clouds and optimal transport. In *MVA*, 2021. 3

[36] Lei Li, Changqing Zou, Youyi Zheng, Qingkun Su, Honbo Fu, and Chiw-Lan Tai. Sketch-r2cnn: An attentive network for vector sketch recognition. *arXiv preprint arXiv:1811.08170*, 2018. 1, 3

[37] Mengtian Li, Zhe Lin, Radomir Mech, Ersin Yumer, and Deva Ramanan. Photo-sketching: Inferring contour drawings from images. In *WACV*, 2019. 1

[38] Pelhua Li. Tensor-sift based earth movers distance for contour tracking. *Mathematical Imaging & Vision*, 2013. 3

[39] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *NeurIPS*, 2019. 3

[40] Yi Li, Timothy M Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval. In *BMVC*, 2014. 2

[41] Hangyu Lin, Yanwei Fu, Yu-Gang Jiang, and Xiangyang Xue. Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt. In *CVPR*, 2020. 3

[42] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv perprint arXiv:1312.4400*, 2013. 3

[43] Fang Liu, Xiaoming Deng, Yu-Kun Lai, Yong-Jin Liu, Cuixia Ma, and Hongan Wang. Sketchgan: Joint sketch completion and recognition with generative adversarial network. In *CVPR*, 2019. 3

[44] Fang Liu, Changqing Zhou, Xiaoming Deng, Ran Zuo, Yu-Kun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. Scenesketcher: Fine-grained image retrieval with scene sketches. In *ECCV*, 2020. 1, 2, 3, 6, 7

[45] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2021. 3

[46] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *CVPR*, 2020. 3

[47] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *IJCV*, 2020. 3

[48] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 2020. 3

[49] Hermina Petric Maretic, Mireille EL Gheche, Giovanni Chierchia, and Pascal Frossard. Got: An optimal transport framework for graph comparison. In *NeurIPS*, 2019. 2

[50] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, 2021. 2

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

[52] 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

[53] Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC*, 2017. 2

[54] Kaiyue Pang, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *CVPR*, 2020. 2

[55] Yonggang Qi, Guoyao Su, Pinaki Nath Chowdhury, Mingkang Li, and Yi-Zhe Song. Sketchlattice: Latticed representation for sketch manipulation. In *ICCV*, 2021. 3

[56] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. In *IJCV*, 2000. 3

[57] 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. 2

[58] 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, 4

[59] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Trans. Graph.*, 2016. 1, 8

[60] Nikolay Savinov, Akihito Seki, Lubor Ladicky, Torsten Sattler, and Marc Pollefeys. Quad-networks: Unsupervised learning to rank for interest point detection. In *CVPR*, 2017. 3

[61] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. Deep network flow for multi-object tracking. In *CVPR*, 2017. 3

[62] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Hospedales Timothy M. Learning to sketch with shortcut cycle consistency. In *CVPR*, 2018. 2

[63] Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In *BMVC*, 2017. 2

[64] 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, 6, 7

[65] Guoyao Su, Yonggang Qi, Kaiyue Pang, Jie Yang, and Yi-Zhe Song. Sketchhealer: A graph-to-sequence network for recreating partial human sketches. In *BMVC*, 2020. 3

[66] Christian Szegedy, Vincet Vanhoucke, Sergey Ioffee, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 6

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 6

[68] Titouan Vayer, Laetitia Chapel, Remi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. In *ICML*, 2019. 2, 5

[69] Yannick Verdie, Kwang Moo Yi, Pascal Fua, and Vincet Lepetit. Tilde: A temporally invariant learned detector. In *ICCV*, 2015. 3

[70] Cédric Villani. *Optimal transport: old and new*. Springer-Verlag Berlin Heidelberg, 2009. 3

[71] Tengfei Wang, Hao Ouyang, and Qifeng Chen. Image in-painting with external-internal learning and monochromic bottleneck. In *CVPR*, 2021. 3

[72] 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

[73] Paul Wohlart and Vincet Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *CVPR*, 2015. 3

[74] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE-TPAMI*, 2015. 3

[75] Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. Vocabulary learning via optimal transport for neural machine translation. In *ACL*, 2021. 3

[76] Jun Yang, Bo Wang, Maozheng Wang, and Yunsheng Ke. Unsupervised co-segmentation of 3d shapes based on components. In *CSSE*, 2919. 3

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

[78] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen Change Loy. Sketch me that shoe. In *CVPR*, 2016. 1, 2, 3, 4, 6, 7, 8

[79] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *IJCV*, 2017. 6

[80] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *CVPR*, 2020. 2, 3

[81] Kaihua Zhang, Mingliang Dong, Bo Liu, Xiao-Tong Yuan, and Qingshan Liu. Deepacg: Co-saliency detection via semantic-aware contrast gromov-wasserstein distance. In *CVPR*, 2021. 3

[82] Linguang Zhang and Szymon Rusinkiewicz. Learning to detect features in texture images. In *CVPR*, 2018. 3

[83] Xu Zhang, Felix X Yu, Svebor Karaman, and Shih-Fu Chang. Learning discriminative and transformation covariant local feature detectors. In *CVPR*, 2017. 3

[84] Qi Zhao, Zhi Yang, and Hai Tao. Differential earth mover's distance with its application to visual tracking. *IEEE-TPAMI*, 2008. 3

[85] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, 2019. 3

[86] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *CVPR*, 2021. 2, 5

[87] Denny Zhou, Mao Ye, Chen Chen, Tianjian Meng, Mingxing Tan, Xiaodan Song, Quoc Le, Qiang Liu, and Dale Schuurmans. Go wide, then narrow: Efficient training of deep thin networks. In *ICML*, 2020. 6

[88] Wengang Zhou, Houqiang Li, Yijuan Lu, and Qi Tian. Large scale image search with geometric coding. In *ACM MM*, 2011. 3

[89] Yuqian Zhou, Connelly Barnes, Eli Shechtman, and Sohrab Amirghodsi. Transfill: Reference-guided image inpainting by merging multiple color and spatial transformations. In *CVPR*, 2021. 3

[90] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. Sketchyscene: Rickly-annotated scene sketches. In *ECCV*, 2018. 1, 2, 6, 7, 8