What Can Human Sketches Do for Object Detection?

Pinaki Nath Chowdhury Ayan Kumar Bhunia Aneeshan Sain Subhadeep Koley
Tao Xiang Yi-Zhe Song
SketchX, CVSSP, University of Surrey, United Kingdom.
{p.chowdhury, a.bhunia, a.sain, s.koley, t.xiang, y.song}@surrey.ac.uk

Abstract

Sketches are highly expressive, inherently capturing subjective and fine-grained visual cues. The exploration of such innate properties of human sketches has, however, been limited to that of image retrieval. In this paper, for the first time, we cultivate the expressiveness of sketches but for the fundamental vision task of object detection. The end result is a sketch-enabled object detection framework that detects based on what you sketch – that "zebra" (e.g., one that is eating the grass) in a herd of zebras (instance-aware detection), and only the part (e.g., "head" of a "zebra") that you desire (part-aware detection). We further dictate that our model works without (i) knowing which category to expect at testing (zero-shot) and (ii) not requiring additional bounding boxes (as per fully supervised) and class labels (as per weakly supervised). Instead of devising a model from the ground up, we show an intuitive synergy between foundation models (e.g., CLIP) and existing sketch models build for sketch-based image retrieval (SBIR), which can already elegantly solve the task – CLIP to provide model generalisation, and SBIR to bridge the (sketch→photo) gap. In particular, we first perform independent prompting on both sketch and photo branches of an SBIR model to build highly generalisable sketch and photo encoders on the back of the generalisation ability of CLIP. We then devise a training paradigm to adapt the learned encoders for object detection, such that the region embeddings of detected boxes are aligned with the sketch and photo embeddings from SBIR. Evaluating our framework on standard object detection datasets like PASCAL-VOC and MS-COCO outperforms both supervised (SOD) and weakly-supervised object detectors (WSOD) on zero-shot setups. Project Page: https://pinakinathc.github.io/sketch-detect

1. Introduction

Sketches have been used from prehistoric times for humans to express and record ideas [35, 76]. The level of expressiveness [28, 41] they carry remains unparalleled today even in the face of language [14, 82] – recall that moment that you want to resort to pen and paper (or Zoom Whiteboard) to sketch down an idea?

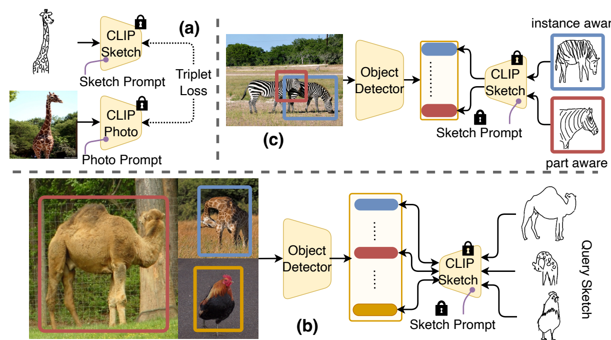

Figure 1. We train an object detector using SBIR models. (a) First, we train an FG-SBIR model using existing sketch–photo pairs that generalise to unseen categories. (b) To train the object detector module, we tile multiple object-level photos from SBIR datasets [75] and use its paired sketch encoding via a pre-trained sketch encoder to align the region embedding of detected boxes. (c) Inclusion of sketches for object detection opens several avenues like detecting a specific object for query sketch (e.g., detect a "zebra" eating grass) or part of an object (e.g., "head" of "zebra").

Sketch research has also flourished over the past decade [16, 70, 90, 99], with a whole spectrum of works on traditional tasks such as classification [30] and synthesis [15, 31, 54], and those more sketch-specific such as modelling visual abstraction [1, 59], style transfer [74] and continuous stroke fitting [17], to cute applications such as turning a sketch into a photo classifier [5, 37].

The expressiveness of sketches, however, has been only explored in the form of sketch-based image retrieval (SBIR) [19, 70, 94], especially the fine-grained [2, 6, 9] variant (FG-SBIR). Great strides have been made, with recent systems already reaching maturity for commercial adaptation [6] – a great testimony to how cultivating sketch expressiveness can make a real impact.

In this paper, we ask the question – what can human sketches do for the fundamental vision tasks of object detection? The envisaged outcome is, therefore, a sketch-enabled object detection framework that detects based on what you sketch, i.e., how *you* want to express yourself. Sketching a "zebra eating the grass" (in Fig. 1) should detect "that" zebra from a herd of zebras (instance-aware detection), *and* it will also give you the freedom to be specific with parts (part-aware detection), so if the "head" of a "zebra" is what you would rather desire, then just sketch the very head.

Instead of devising a sketch-enabled object detection model from the ground up, we show that an intuitive synergy between foundation models (e.g., CLIP [64]) and off-the-shelf SBIR models [8,100] can already, rather elegantly, solve the problem – CLIP to provide model generalization, and SBIR to bridge the (sketch→photo) gap. In particular, we adapt CLIP to build sketch and photo encoders (branches in a common SBIR model) by learning independent prompt vectors [42, 71] separately for both modalities. More specifically, during training, the learnable prompt vectors are prepended into the input sequence of the first transformer layer of CLIP's ViT backbone [22] while keeping the rest frozen. As such, we inject model generalization into the learned sketch and photo distributions. Next, we devise a training paradigm to adapt the learned encoders for object detection, such that the region embeddings of detected boxes are aligned with the sketch and photo embeddings from SBIR. This allows our object detector to train *without* requiring additional training photos (bounding boxes and class labels) from auxiliary datasets.

To make our sketch-based detector more interesting (general-purpose [13, 57]), we further dictate it also works in a zero-shot manner. For that, following [10], we extend object detection from a pre-defined fixed-set setup to an open-vocab setup. Specifically, we replace the classification heads in object detectors with prototype learning [49], where the encoded query sketch features act as the support set (or prototypes). Next, the model is trained under the weakly supervised object detection (WSOD) setting [10, 78], using a multi-category cross-entropy loss over the prototypes of all possible categories or instances. However, while SBIR is trained using object-level (single object) sketch/photo pairs, object detection works on image-level (multiple categories). Hence, to train object detectors using SBIR, we also need to bridge the gap between object and image-level features. Towards this, we use a data augmentation trick that is embarrassingly simple yet highly effective for robustness towards corruption and generalisation to out-of-vocab [101, 102] – we randomly select $n = \{1, \ldots, 7\}$ photos from SBIR datasets [30, 75] and arbitrarily tile them on a blank canvas (similar to CutMix [101]).

In summary, our contributions are (i) for the first time cultivating the expressiveness of human sketches for object detection, (ii) sketch-based object detector that detects what you intend to express in your sketch, (iii) an object detector that is both instance-aware and part-aware, in addition to performing conventional category-level detection. (iv) a novel prompt learning setup to marry CLIP and SBIR to build the sketch-aware detector that works without needing bounding box annotations (as supervised [67]), class labels (as weakly supervised [10]), and in a zero-shot manner. (v) results outperform both supervised (SOD) and weakly supervised object detectors (WSOD) on zero-shot setup.

2. Related Works

Sketch for Visual Understanding Hand-drawn sketches serve as a useful query modality for visual understanding tasks that involve human perception and structural cues. Sketches not only convey a visual description [35] but also exhibit artistic styles [103]. This makes sketch a vital querying modality for the creative industry, like artistic image editing [98] and animation [93]. Unlike photos that are passively captured by a camera, sketches are actively drawn by humans, which makes them a good visual representation [4, 7] enriched with human participation. Apart from the widely explored sketch-based image retrieval [2, 9, 70, 74], sketch as a query has shown potential in several vision understanding tasks like incremental learning [5] image and video synthesis [31, 44, 54], representation learning [1, 88], image-inpainting [92], 3D shape retrieval [96], 3D shape modelling [15], medical image analysis [43, 90], object localisation [69, 86] and segmentation [38, 63].

Studying sketch as a query for object detection by Tripathi *et al*. [86], several limitations surfaced with respect to problem definition as well as architectural designs. Firstly, instead of fine-grained matching, sketch was used to specify object category (easier via text/keyword [29,64]), thus overlooking the potential of sketch to model fine-grained details. Secondly, it requires both bounding-box and sketch annotation, which increases the annotation budget without significant improvement in performance over traditional object detection setups. Thirdly, due to an expensive annotation, only fewer than 50% object categories in existing object detection datasets [23, 52] are available for training. Finally, using an early fusion strategy [95] of sketch with object detection results in recomputing object regions for each new sketch – leading to a slower detection framework with increasing query sketches. In this paper, we propose a fine-grained sketch-based object detection framework that uses only object-level sketch photo pairs without any bounding-box annotations for training and is scalable with multiple fine-grained query sketches, even under zero-shot setup.

Supervised Object Detection Object detection jointly localises and identifies objects in an image. Traditional object detectors rely on large datasets such as PASCAL VOC [23] and MS-COCO [52], containing thousands of examples per object category which are quite time-consuming to annotate, unlike our pipeline, that leverage sketch-photo pairs as annotation. Existing literature on object detection is bifurcated as: (i) fast yet less accurate single-shot [51, 55, 65, 66, 108]; (ii) slow but more accurate two-stage object detectors [26, 27, 32, 67]. To fully exploit the fine-grained cues provided by sketch, our proposed method aligns with the accurate two-stage detectors that predict object regions using selective search in RCNN [27], ROI pooling [26] in Fast-RCNN, and Region Proposal Network (RPN) with ROIAlign in Faster-RCNN [67]. While there has been several attempts with sophisticated architectural modifications [46, 108, 109], Faster-RCNN [67] still acts a fundamental building block for several downstream tasks like scene-graph generation [97], visual grounding [58], and relationship prediction [110]. Therefore, we resort to the more traditional Faster-RCNN based two-stage pipeline.

Weakly Supervised Object Detection (WSOD) Collecting bounding box annotation per object category is already a time-consuming process, which is aggravated even further by fine-grained object detection (e.g., recognising animal species). To overcome this, existing WSOD adopt two schools of thought: (a) formulate this as a multiple instance learning (MIL) [10,20,21,40,48,79,85,104] problem that interpret an image as a bag of proposals or regions. The image is labelled positive if one of the regions tightly contains the object of interest; otherwise negative. (b) CAM-based methods [105, 106] that use class activation maps to predict proposals. Specifically, an image is fed to a backbone network to generate a feature vector from which the bounding box of each class is predicted by thresholding activation maps with the highest probability. Although CAM-based methods are faster, we use MIL-based technique as it can detect *multiple instances* within the *same* category [78].

Data Augmentation in Computer Vision Data augmentation improves the sufficiency and diversity of training data. Approaches vary from simple image rotation and flipping to more advanced techniques of image erasing [11] like CutOut [18], Hide-and-Seek [81] and image mixing like MixUp [102] and CutMix [101]. Aiming to generalise sketch-based object detection to complex scenes while training exclusively on existing object-level sketch photo pairs [75], we employ a CutMix [101] like data augmentation trick – a method that replaces removed sub-regions with a patch from another image to synthesise new images.

Sketch-Based Object Representation Sketch with its intrinsic ability to model fine-grained visual details makes it an ideal modality for object retrieval, giving rise to avenues like *category-level* [16,19,70,94,99] and fine-grained (FG) *instance-level* [2, 6, 9, 72] sketch-based image retrieval (SBIR). Contemporary research also explored zero-shot SBIR [19, 73, 99], cross-domain translation [61] and approaches like reinforcement learning based on-the-fly retrieval [9], self-supervised [3, 62], etc. Apart from object-level images, retrieving sketched objects from scene images was studied using graph convolutional networks [53] and optimal transport [12]. Similar to cross-category FG-SBIR [8, 61], here we explore fine-grained sketch photo association for object detection by adapting large vision-language models like CLIP [64] using prompt engineering [107].

3. Proposed Method

Overview We propose a new paradigm training object detection without bounding box annotation or image-level class labels. Instead, we use sketch-based image retrieval for supervision. This leads to several emergent behaviours (i) fine-grained object detection – specify focused region-of-interest using fine-grained visual cues in sketch. (ii) category-level object detection – specify the category of detected instances via sketch. (iii) part-level object detection – detect specified parts (e.g., "head" and "legs" of a "horse").

3.1. Background

Our framework has two key modules – Object Detection and Sketch-Based Image Retrieval (category-level and fine-grained). For completeness, we give a brief background.

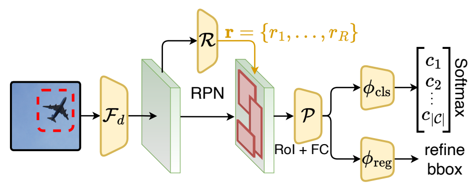

Figure 2. Faster-RCNN [67] use image encoder $\mathcal{F}_d$ and RPN $\mathcal{R}$ to generates box proposals. Feature maps of proposals, computed via RoI pool $\mathcal{P}$, predicts class probabilities and box regression.

Baseline Supervised Object Detection We briefly introduce a supervised object detection (SOD) framework, Faster-RCNN [67] that remains state-of-the-art [58, 97, 110]. Given a photo $\mathbf{p} \in \mathbb{R}^{H \times W \times 3}$, a backbone feature extractor (VGG [80], or ResNet [33]) $\mathcal{F}_d(\cdot) : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{H' \times W' \times 512}$ computes feature map $f_\mathbf{p} \in \mathbb{R}^{H' \times W' \times 512}$. Next, a two-stage process is followed: (i) Given backbone feature map $f_\mathbf{p}$, a region proposal network (RPN) $\mathcal{R} : \mathbb{R}^{H' \times W' \times 512} \rightarrow \mathbb{R}^{R \times 5}$ generates rectangular boxes (i.e., proposals) $\mathbf{r} = \{r_1, \dots, r_R\}$, where $\mathbf{r} \in \mathbb{R}^{R \times 4}$ and "objectness measure" – a scalar $[0, 1]$ probability of the box r_j having an object. (ii) Using proposals $\mathbf{r} \in \mathbb{R}^{R \times 4}$ we pool the feature map $f_\mathbf{p}$ via RoI pool [26] to get intermediate feature of size $\mathbb{R}^{7 \times 7 \times 512}$, followed by a fully-connected layer (FC) to get $f_\mathbf{r} \in \mathbb{R}^{R \times 512}$ as, $f_\mathbf{r} = \mathcal{P}(f_\mathbf{p}, \mathbf{r})$. The patch feature $f_\mathbf{r}$ is branched into two streams – a classification branch $\phi_{\text{cls}} : \mathbb{R}^{R \times 512} \rightarrow \mathbb{R}^{R \times (|\mathcal{C}|+1)}$ outputs probability distribution (per RoI) over $\mathcal{C}$ pre-defined classes and a catch-all background class; a box regressions $\phi_{\text{reg}} : \mathbb{R}^{R \times 512} \rightarrow \mathbb{R}^{R \times 4}$ refines initial box predictions $\mathbf{r} \in \mathbb{R}^{R \times 4}$.

Baseline SBIR Framework We recap a baseline SBIR framework. Given a sketch/photo pair $(\mathbf{s}, \mathbf{p})$, we use a sketch/photo feature extractor to get the feature map $f_\mathbf{s} = \mathcal{F}_\mathbf{s}(\mathbf{s}) \in \mathbb{R}^{512}$ and $f_\mathbf{p} = \mathcal{F}_\mathbf{p}(\mathbf{p}) \in \mathbb{R}^{512}$. Category-level SBIR requires $(\mathbf{s}, \mathbf{p})$ from the same category, whereas fine-grained SBIR requires instance-level sketch/photo matching. For training, the cosine distance $\delta(\cdot, \cdot)$ to a sketch anchor ($\mathbf{s}$) from a negative photo ($\mathbf{p}^-$), denoted as $\beta^- = \delta(f_\mathbf{s}, f_{\mathbf{p}^-})$ should increase while that from the positive photo ($\mathbf{p}^+$), $\beta^+ = \delta(f_\mathbf{s}, f_{\mathbf{p}^+})$ should decrease. Training is done via triplet loss with hyperparameter $\mu > 0$,

$$\mathcal{L}_{\text{trip}} = \max\{0, \mu + \beta^+ - \beta^-\} \tag{1}$$

To extend FG-SBIR across multiple categories (cross-category FG-SBIR), we train with Eq. (1) using "hard-triplets" – $(\mathbf{s}, \mathbf{p}^+, \mathbf{p}^-)$ have same category, and a class discriminator loss across categories using cross-entropy loss,

$$\mathcal{L}_{\text{cat}} = -c^i_\mathbf{q} \log \frac{\exp(\mathcal{F}_\mathbf{c}(f^i_\mathbf{q}))}{\sum_{\forall j} \exp(\mathcal{F}_\mathbf{c}(f^j_\mathbf{q}))} \tag{2}$$

where, query $\mathbf{q} = \{\mathbf{s}, \mathbf{p}\}$, $c^i_\mathbf{q}$ represent class label of i^{th} sample, $\mathcal{F}_\mathbf{c} : \mathbb{R}^{512} \rightarrow \mathbb{R}^{|\mathcal{C}|}$ predicts softmax probabilities.

3.2. Weakly Supervised Object Detection

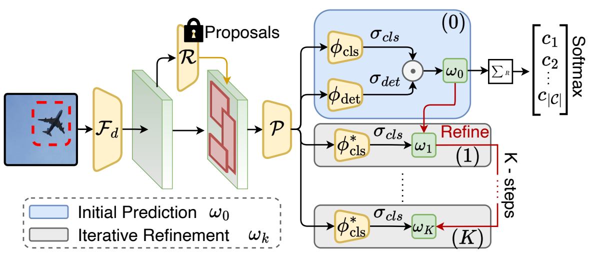

Figure 3. Weakly supervised setup (no bounding box) trains using image-level class labels with classification heads (ϕ_{cls}). Initial prediction ω_0 is refined in K steps with ϕ^*_{cls} to predict ω_k.

To avoid collecting expensive bounding box annotation, weakly supervised object detection (WSOD) trains using image-level class labels – objects of a class are present or not. To avoid using bounding box annotation, we either use a pre-trained region proposal network[1] ($\mathcal{R}$) or heuristic-based selective search [87], or edge boxes [111] that generate box proposals $\mathbf{r} = \{r_1, \dots, r_R\}$. The patch features $f_\mathbf{r} = \mathcal{P}(f_\mathbf{p}, \mathbf{r})$ is branched into a classification head $x_c = \phi_{\text{cls}}(f_\mathbf{r}) \in \mathbb{R}^{R \times (|\mathcal{C}|+1)}$ and a detection head $x_d = \phi_{\text{det}}(f_\mathbf{r}) \in \mathbb{R}^{R \times (|\mathcal{C}|+1)}$. The classification head ϕ_{cls} scores individuals proposals into pre-defined $\mathcal{C}$ classes and a catch-all background class via softmax across ($|\mathcal{C}| + 1$) class labels

$$\sigma_{\text{cls}}(x^{(i,j)}_c) = \frac{\exp(x^{(i,j)}_c)}{\sum^{|\mathcal{C}|+1}_{k=1} \exp(x^{(i,k)}_c)} \tag{3}$$

The detection head ϕ_{det} measures the contribution of each patch i ($r_i \in \mathbf{r}$) of being classified to class j (in $\mathcal{C}+1$), (i.e., a patch score for each class) via softmax across R regions

$$\sigma_{\text{det}}(x^{(i,j)}_d) = \frac{\exp(x^{(i,j)}_d)}{\sum^R_{k=1} \exp(x^{(k,j)}_d)} \tag{4}$$

[1]Pre-trained [45] RPN is highly generalisable to unseen datasets [29] due to its generic objective that learns to predict "objectness" measure.

We train using image-level labels $\mathbf{Y} = [y_0, y_1, \dots, y_{|\mathcal{C}|}]^T \in \mathbb{R}^{(|\mathcal{C}|+1)\times 1}$, where $y_c = 1$ or 0 indicates if instance of class $c \in \mathcal{C}$ is present in the image or not. The combined score (element-wise product) of class score σ_{cls} for each patch and a patch score for each class σ_{det} is computed as, $\omega_0 = \sigma_{\text{cls}}(x_c) \odot \sigma_{\text{det}}(x_d)$. Since we only have image-level class labels, from the combined score $\omega_0 \in \mathbb{R}^{R\times(|\mathcal{C}|+1)}$, we take the sum over all patches $\hat{y}_c = \sum^R_{i=1} \omega^{i,c}_0$ to get the probability of instances from the c^{th} class present in the image or not. Training happens via multi-class cross entropy,

$$\mathcal{L}_{\text{ws}} = -\sum^{|\mathcal{C}|+1}_{c=1} y_c \log \hat{y}_c + (1 - y_c) \log(1 - \hat{y}_c) \tag{5}$$

Unlike SOD, using bounding box annotation to refine proposals, WSOD uses only image-level class labels that fail to naively refine proposals. Hence, we use an iterative refinement classifier $\omega_k = \phi^*_{\text{cls}}(f_\mathbf{r})$, where $\omega_k \in \mathbb{R}^{R\times(|\mathcal{C}|+1)}$ to predict a *refined* class score for each RoI, as shown in Fig. 3. The refinement classifier ϕ^*_{cls} is supervised via pseudo scores labels l_{k-1} from $(k-1)^{th}$ iteration as, (i) we compute the patches with highest scores in each class $r^c_* = \arg\max_r(\omega^{(r,c)}_{k-1})$. (ii) All regions $r_i \in \mathbf{r}$ that has high overlap with a top scoring patch r^c_* should be the same class label c as, $l^{r,c}_{k-1} = 1$ if $\text{IoU}(r_i, r^c_*) \geq 0.5$. (iii) If a region $r_i \in \mathbf{r}$ has low overlap with any top scoring patch r^c_*, we assign it to background class $l^{r,0}_{k-1} = 1$. (iv) If a class c is not in image $\mathbf{p}$ we assign $l^{r,c}_{k-1} = 0$. The refinement loss is

$$\mathcal{L}^k_{\text{ref}} = \frac{1}{R} \sum^R_{i=1} \sum^{|\mathcal{C}|}_{c=1} \omega^{(i,j)}_{k-1}\, l^{(i,j)}_{k-1} \, \log \omega^{(i,j)}_k \tag{6}$$

Both SOD and WSOD restrict detection to pre-defined $\mathcal{C}$ classes. In the next section, we overcome this fixed-set limitation using prototype learning with SBIR.

3.3. Localising Object Regions with Query Sketch

We replace the fixed-set classifier in WSOD with scalable open-set prototype learning [49]. Each head $\{\phi_{\text{cls}}, \phi_{\text{det}}, \phi^*_{\text{cls}}\}$ in WSOD that predict scores $\mathbb{R}^{R\times 512} \rightarrow \mathbb{R}^{R\times(|\mathcal{C}|+1)}$ is modified to compute their respective embedding vectors $e = \{e_{\text{cls}}, e_{\text{det}}, e^*_{\text{cls}}\}$ as $\mathbb{R}^{R\times 512} \rightarrow \mathbb{R}^{R\times 512}$. Next, we compute a support set (prototypes for category-level/instance-level sketch) $\mathcal{S} = [e_\mathbf{bg}, f^1_\mathbf{s}, f^2_\mathbf{s}, \dots, f^{|\mathcal{C}|}_\mathbf{s}]^T \in \mathbb{R}^{512\times(|\mathcal{C}|+1)}$ by encoding query sketches $\{\mathbf{s}_1, \dots, \mathbf{s}_{|\mathcal{C}|}\}$ with a pre-trained sketch encoder ($\mathcal{F}_\mathbf{s}$) and a learned catch-all background embedding $e_\mathbf{bg} \in \mathbb{R}^{512}$, as shown in Fig. 4. The scores $\{x_c, x_d, \omega_k\}$ (analogous to Sec. 3.2) are computed using $\mathcal{S}$ and embedding vectors e of detected regions

$$x_c = e_{\text{cls}} \cdot \mathcal{S}; \quad x_d = e_{\text{det}} \cdot \mathcal{S}; \quad \omega_k = e^*_{\text{cls}} \cdot \mathcal{S} \tag{7}$$

Carefully choosing a sketch encoder $\mathcal{F}_\mathbf{s}$ leads to several properties: (i) pre-training $\mathcal{F}_\mathbf{s}$ on category-level SBIR computes $\mathcal{S}$ that detect regions $\mathbf{r}$ with the same category as

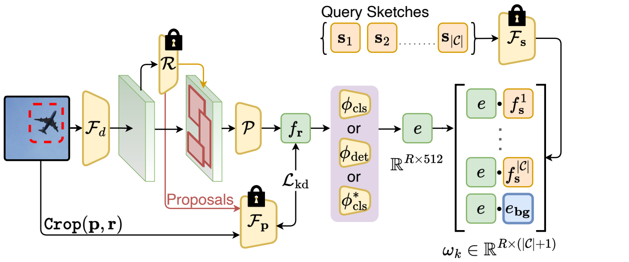

Figure 4. The object detection modules $\{\mathcal{F}_d, \mathcal{P}, \phi_{cls}, \phi_{det}, \phi^*_{cls}\}$ are learned using pre-trained sketch ($\mathcal{F}_\mathbf{s}$) and photo ($\mathcal{F}_\mathbf{p}$) encoders.

query sketches – category-level object detection. (ii) pre-training $\mathcal{F}_\mathbf{s}$ on cross-category FG-SBIR computes $\mathcal{S}$ where only instance-level aligned regions $\mathbf{r}$ are detected – fine-grained object detection. (iii) Extending fine-grained object detection with a generalisable (out-of-vocab) sketch encoder $\mathcal{F}_\mathbf{s}$ helps to detect object parts (e.g., "head" of a "horse") given query sketches – part-level object detection. We train object detection modules $\{\mathcal{F}_d, \mathcal{P}, \phi_{\text{cls}}, \phi_{\text{det}}, \phi^*_{\text{cls}}\}$, using Eq. (5) and Eq. (6) in WSOD (Sec. 3.2).

While the sketch encoder ($\mathcal{F}_\mathbf{s}$) trains object detector via prototypes for each category/instance sketch, we further enhance training efficiency with additional supervision from the photo encoder ($\mathcal{F}_\mathbf{p}$), as shown in Fig. 4. Specifically, we impose a $L1$-based feature matching loss (analogous to feature distillation [34]) between patch features $f_\mathbf{r}$ from proposals $\mathbf{r}$ in object detector and the photo feature computed for cropped photo regions $\texttt{Crop}(\mathbf{p}, \mathbf{r})$ using pre-trained $\mathcal{F}_\mathbf{p}$ as, $\mathcal{L}_{\text{kd}} = ||f_\mathbf{r} - \mathcal{F}_\mathbf{p}(\texttt{Crop}(\mathbf{p}, \mathbf{r}))||_1$. The final loss is,

$$\mathcal{L}_{\text{tot}} = \underbrace{\mathcal{L}_{\text{ws}} + \sum_{k=1}^{K} \mathcal{L}^k_{\text{ref}}}_{Eq.\ (5)\text{ and } Eq.\ (6)} + \lambda \underbrace{||f_\mathbf{r} - \mathcal{F}_\mathbf{p}(\texttt{Crop}(\mathbf{p}, \mathbf{r}))||_1}_{\mathcal{L}_{\text{kd}}} \quad (8)$$

where the hyperparameter $\lambda = 1$. Although, in theory, we can use our baseline SBIR (in Sec. 3.1), training object detection requires learning a generalised (out-of-vocab) SBIR for category-level and fine-grained sketch/photo matching under wide variations like illumination, complex background, occlusions, unseen categories etc.

3.4. Prompt Learning for Generalised SBIR

To train object detection using SBIR with high generalisation and open-vocab capabilities, we introduce prompt learning [107] using CLIP [64] for SBIR (both category-level and cross-category fine-grained). CLIP [64] consists an image and text encoder (e.g., ViT [22], or ResNet [33]) trained on large $400M$ text/image pairs. This leads to a highly generalisable model that works zero-shot across multiple tasks and datasets. However, adapting CLIP for sketches is tricky since naive fine-tuning leads to model collapse. Hence, we use prompt learning, a set of P learnable vector $\mathbf{v_s} \in \mathbb{R}^{P \times 768}$ for sketch and $\mathbf{v_p} \in \mathbb{R}^{P \times 768}$

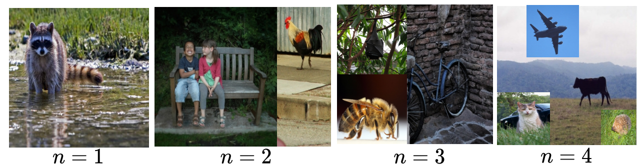

Figure 5. Bridge object and image-level gap with synthetic photos by tiling $n = \{1, \ldots 7\}$ object-level photos in SBIR datasets.

for photo, injected into the first layer of ViT to induce CLIP to learn downstream sketch/photo distribution. Importantly, prompting CLIP preserves the desired generalisation ability [107] since the knowledge learned by CLIP is distilled into prompt's weights while keeping the ViT weights frozen. Our new sketch encoder is defined by adapting CLIP's image encoder using sketch prompt ($\mathbf{v_s}$) as, $\mathcal{F}_\mathbf{s}(\cdot) = \mathcal{F}_{\text{clip}}(\cdot, \mathbf{v_s})$ and using $\mathbf{v_p}$ for photo encoder as, $\mathcal{F}_\mathbf{p}(\cdot) = \mathcal{F}_{\text{clip}}(\cdot, \mathbf{v_p})$. Since ViT weights are frozen, training our CLIP-based SBIR is parameter-efficient – we train only $\mathbf{v_s} \in \mathbb{R}^{P \times 768}$ and $\mathbf{v_p} \in \mathbb{R}^{P \times 768}$. This allows training with less data, and faster convergence. For category-level SBIR, $(\mathbf{v_s}, \mathbf{v_p})$ learns category inducing prompts using triplet loss (in Eq. (1)). Learning cross-category FG-SBIR, is slightly more complicated that trains $(\mathbf{v_s}, \mathbf{v_p})$ using hard-triplet in Eq. (1), and a modified class discriminative loss Eq. (2) using CLIP's text encoder as,

$$\mathcal{L}_{\text{cat}} = -c^i_\mathbf{q} \log \frac{\exp(f^i_\mathbf{q} \cdot f^i_\mathbf{t})}{\sum_{\forall j} \exp(f^i_\mathbf{q} \cdot f^j_\mathbf{t})} \quad (9)$$

where, $f^i_\mathbf{t} \in \mathbb{R}^{512}$ is computed by CLIP's text encoder as, $f^i_\mathbf{t} = \mathcal{F}^{(\mathbf{t})}_{\text{clip}}(\text{“}\texttt{a photo of a } [c^i_\mathbf{q}]\text{”})$ for category $c^i_\mathbf{q}$. Equipped with our novel prompt-based SBIR, we train open-vocab category-level object detection, fine-grained object detection, and part-level object detection.

3.5. Bridging Object-Level and Image-Level

While SBIR is trained using object-level (single object) sketch/photo pairs, object detection works on image-level (multiple objects) data. To train object detectors using SBIR, we need to bridge this object and image-level gap. Our solution is embarrassingly simple – synthesise a canvas of size $(H \times W)$ by randomly tiling $n = \{1, \ldots, 7\}$ object-level photos in SBIR datasets [30, 75]. Despite its simplicity, our augmentation trick, analogous to CutMix [101], improves robustness against input corruptions and out-of-distribution generalisation [101, 102]. The paired sketches for photos in canvas are used to construct the support set $\mathcal{S}$. Note, we train our object detector without the need to "see" the evaluation data distribution or use any annotation (bounding box or image-level class labels). We call this setup – extremely weakly supervised object detection (EW-SOD) – no need to "see" the downstream data distribution.

4. Experiments

Dataset We train our object detector using existing cross-category FG-SBIR dataset – Sketchy [75] that contains 125 categories, each with 100 photos. Every photo in [75] has at least 5 instance-level paired sketches. To evaluate fine-grained object detection, we use SketchyCOCO [25] comprising of natural images in MS-COCO [52] with instance-level paired sketches. Following Liu *et al*. [53], we select $1,225$ sketch/photo pairs from SketchyCOCO [25] with at least one foreground sketched object. We filter the overlapping categories of in SketchyCOCO [25] from Sketchy [75] to measure true zero-shot performance. For category-level object detection, we train on category-level sketch/photo pairs in QuickDraw-Extended [19] having $330k$ sketches and $204k$ photos from 110 categories. Following [86], we evaluate on a subset of standard object detection datasets like PASCAL-VOC [23] and MS-COCO [52] that have 20 and 56 overlapping categories in QuickDraw [30].

Implementation Details Our model is implemented in PyTorch on a 11GB Nvidia RTX 2080-Ti GPU. First, we train a generalised cross-category FG-SBIR with image size (224×224) by adapting CLIP with ViT [22] backbone (ViT-B/32 weights) using prompt learning [39]. The prompts ($P = 3$) are trained with triplet loss [100], margin $\mu = 0.3$, Adam optimiser with learning rate $1e - 4$ for 60 epochs, and batch size 64. Our object detection pipeline is build using Detectron2 [91]. We use FasterRCNN [67], pretrained on Visual Genome [45] and remove the RoIPooling [26] and subsequent layers to keep only the pretrained backbone ResNet+FPN ($\mathcal{F}_d$) [33, 50] and Region Proposal Network ($\mathcal{R}$) that generates 1000 proposals. An alternative is to use handcrafted region proposals like selective search [87], but we observed slight performance drop. The object detector trains using SGD with batch size 8 and initial learning rate $5e - 3$, multiplied by 0.1 at $150k$ and $250k$ iterations. We train in a two-step process: (i) keeping $\mathcal{F}_d$ and $\mathcal{R}$ fixed, we train the RoI pooling and FC layers ($\mathcal{P}$), classification head (ϕ_{cls}), detection head (ϕ_{det}), and refinement head (ϕ^*_{cls}) for $240k$ iterations. (ii) We freeze only $\mathcal{R}$ and finetune all modules for $80k$ iterations. Non-maxima suppression with IoU ≥ 0.3 is applied to get final predictions.

Evaluation Metric (i) For fine-grained object detection, we measure $AP_{.3}$, $AP_{.5}$, and $AP_{.7}$ that computes the average precision (AP) at IoU values 0.3, 0.5, and 0.7. (ii) For category-level object detection, we use measure $AP_{.5}$ and *CorLoc* that computes percentage of images for which the most confident predicted box has IoU ≥ 0.5 with at least one of the ground-truth boxes for every class. (iii) For cross-category FG-SBIR, we measures $Acc.@q$ – percentage of sketches having true matched photo in the top-q list, and (iv) mean average precision (mAP), and precision considering top 200 retrievals P@200 for category-level SBIR.

4.1. Competitors

For object detection, we compare against, (i) supervised object detection (SOD) using both bounding box in addition to sketch/photo annotations: **Mod-FRCNN** adapts Faster-RCNN [67] for unseen class by concatenating query sketch feature with the RoI pooled feature followed by a binary classifier. **MatchNet** [36] extends *Mod-FRCNN* using co-attention to generate region proposals conditioned on query sketch along with squeeze-and-co-excitation to adaptively re-weight importance distribution of candidate proposals. **CoAttOD** [86] improves upon *MatchNet* by mitigating the sketch/photo domain misalignment using cross-modal attention. (ii) Weakly supervised object detection (WSOD) trains only on image-level sketch annotations without any additional bounding boxes: **WSDDN** [10] repurposed object detection as a region classification via multiple instance learning (MIL) paradigm. To inject query sketch to *WSDDN*, we use cross-attention with RoI pooled feature followed by a binary classifier for detection. **OICR** [84] improves *WSDDN* with an iterative MIL to refine initial predictions scores to improve discriminatory power for detection. **PCL** [83] generates multiple positive instance in an image via clustering and assigning proposals to the label of corresponding object class for each cluster. **ICMWSD** [68] addresses the problem of prior WSOD that focus on the most discriminative part of an object using context information. In particular, *ICMWSD* obtains a "dropped feature" by dropping the most discriminative parts, followed by maximising the loss of the "dropped feature" that force the network to look in the surrounding context regions. (iii) We adapt <**Method**> in *WSOD* to **E-<Method>** that exclusively training on SBIR datasets [30, 75] by synthesising canvas with randomly tiling $n = \{1, \ldots, 7\}$ object-level photos and using their paired sketches to construct the support $\mathcal{S}$. We call this setup – extreme weakly supervised object detection (EWSOD).

For zero-shot category-level SBIR, we compare against: **GRL** [19] combines similar semantic information (word2vec [56]) of class labels with visual sketch information and trains using a gradient reversal layer [24] to reduce sketch/photo domain gap. **VKD** [89] is similar to ours using prototype-learning but employ selective knowledge distillation and ViT [22] backbone. For zero-shot cross-category FG-SBIR: **CDG** is a SOTA domain generalisation method [77] adapted to cross-category FG-SBIR [60] using categories as domain and intra-category sketch/photo pairs as label. **CCD** [60] models a universal manifold of prototypical visual sketch traits that dynamically embeds sketch/photo, to generalise to unseen categories.

4.2. Generalisibility of Cross-Category FG-SBIR

Due to the significant impact of SBIR on training object detectors, it is imperative to learn a powerful cross-

Table 1. Quantitative performance of zero-shot category-level SBIR (CL-SBIR) and cross-category FG-SBIR (CC-FGSBIR).

Train		CL-SBIR [19]			CC-FGSBIR [75]	
		mAP	P@200		Acc.@1	Acc.@5
100%	GRL	9.01	6.75	CDG	20.1	46.4
	VKD	15.0	29.8	CCD	22.6	49.0
	Ours	**18.2**	**36.1**	**Ours**	**27.6**	**59.5**
70%	GRL	6.3	5.7	CDG	14.6	39.5
	VKD	9.4	17.3	CCD	16.3	41.4
	Ours	**13.1**	**23.2**	**Ours**	**21.0**	**47.7**
50%	GRL	3.2	2.7	CDG	7.9	25.4
	VKD	4.8	6.3	CCD	9.2	32.2
	Ours	**9.6**	**11.4**	**Ours**	**14.7**	**40.1**

Figure 6. Qualitative retrieval results for cross-category FG-SBIR.

category FG-SBIR that is highly generalisable. In other words, the accuracy of SBIR puts a bottleneck on object detection performance. Tab. 1 compares category-level SBIR (CL-SBIR) and cross-category FG-SBIR (CC-FGSBIR) on QuickDraw-Extended [19] and Sketchy [75] respectively, using 100%, 70%, and 50% of the training set.

Performance Analysis From Tab. 1 we make the following observations: (i) with decreasing train-set categories, the performance gap (ratio of proposed / SOTA) between the proposed method versus GRL (for CL-SBIR) and CDG (for CC-FGSBIR) increases from $2.1/1.4$ at 100% data to $3.0/4.2$ at 50% data. This shows the high generalisation potential when using prompt-based CLIP models for sketch/photo matching. (ii) Performance gap of proposed versus SOTAs for $100\% \rightarrow 50\%$ is more significant in CC-FGSBIR as compared to CL-SBIR. Hence, it is more difficult to discriminate unseen intra-category sketch/photo pairs than recognise a novel categories. (iii) Performance of all competitors in CL-SBIR and CC-FGSBIR are staggeringly inferior to proposed CLIP-based approach. Such a strong SBIR is necessary to unlock training object detection in EWSOD setup (cross-dataset and weakly supervised).

4.3. Category-Level Object Detection

We benchmark on a subset of standard object detection PASCAL-VOC [23] and MS-COCO [52] datasets that have overlapping categories with QuickDraw [30] sketches. Unlike traditional object detection that detects all instances for known classes in an image, category-level object detection specifies the category of interest by drawing a query sketch.

Performance Analysis From Tab. 2 we observe: (i) best SOD method outperform the best WSOD by an average $AP_{.5}$ margin of $1.7\%/0.1\%$ in VOC/MS-COCO. This shows that although WSOD performs less than SOD (using additional bounding box annotation), the performance gap is not as significant as generally observed in prior works on seen setup using text as query [29, 68, 84]. In other words, using sketch gives nearly similar performance for zero-shot setup for SOD and WSOD. (ii) EWSOD methods further drops $AP_{.5}$ of best WSOD method by $5.1\%/1.6\%$. This highlights the lack of generalisation of object detectors to the shift in data distribution when trained on SBIR photos and tested on VOC/MS-COCO. (iii) Despite being trained on the challenging EWSOD setup, our proposed method outperforms best SOD by $14.7/10.9$, WSOD by $16.4/11.0$, and EWSOD by $21.5/13.2$ in zero-shot setup. This shows the extreme generalisation potential of training object detetction using a strong CLIP-based SBIR.

Table 2. Quantitative performance of category-level object detection on VOC 2007 and MS-COCO using $AP_{.5}$ and *CorLoc*.

	Method	VOC 2007 [23]		MS-COCO [52]	
		$AP_{.5}$	CorLoc	$AP_{.5}$	CorLoc
SOD	Mod-FRCNN	30.1	51.2	7.4	65.8
	MatchNet	31.4	51.7	12.4	68.1
	CoAttOD	34.6	53.9	15.0	**71.3**
WSOD	WSDDN	20.9	40.1	11.9	67.3
	OICR	24.7	42.3	12.2	67.7
	PCL	26.1	45.5	13.8	68.6
	ICMWSD	32.9	52.6	14.9	69.5
EWSOD	E-WSDDN	17.7	37.9	10.1	66.7
	E-OICR	21.2	40.5	10.4	67.0
	E-PCL	22.3	41.1	11.8	67.3
	E-ICMWSD	27.8	46.3	12.7	67.9
	Proposed	**49.3**	**69.4**	**25.9**	70.3

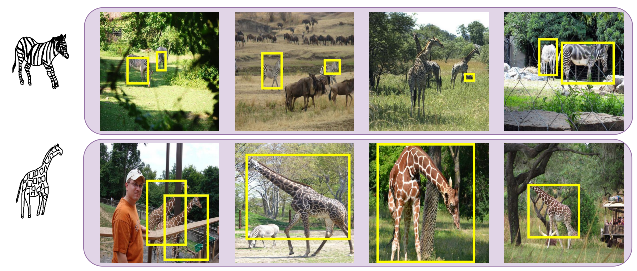

Figure 7. Category-Level Object Detection using query sketches with images from MS-COCO [52] and PASCAL-VOC [23].

4.4. Fine-Grained Object Detection

Unlike category-level object detection that detects all instances of sketched category, the goal of fine-grained object detection is to detect only a specific instance for the input query sketch with instance-level alignment.

Performance Analysis From Tab. 3 we observe: (i) Methods in SOD have nearly similar performance as WSOD and drops for EWSOD, similar to that in category-level detection in Tab. 2. (ii) Compared to SOD, the performance of WSOD drops more for $AP_{.5} \rightarrow AP_{.7}$. This

Table 3. SketchyCOCO detection fine-grained.

	Method	$AP_{.3}$	$AP_{.5}$	$AP_{.7}$
SOD	Mod-FRCNN	2.5	3.5	3.1
	MatchNet	9.3	11.0	10.5
	CoAttOD	10.4	12.1	11.7
WSOD	WSDDN	8.1	10.2	9.4
	OICR	8.9	10.9	10.0
	PCL	9.2	11.5	10.6
	ICMWSD	10.3	11.9	10.8
EWSOD	E-WSDDN	6.4	8.5	7.6
	E-OICR	7.1	9.1	8.3
	E-PCL	7.3	9.4	8.7
	E-ICMWSD	8.5	10.2	9.4
	Proposed	**15.0**	**17.1**	**16.3**

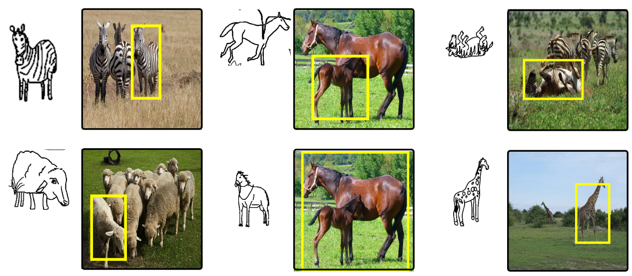

Figure 8. Cross-Category Fine-Grained Object Detection using query sketches with images from SketchyCOCO [25].

is since WSOD methods use less accurate selective search [87] and edge boxes [111] for region proposals compared to the more accurate RPN [67] in SOD. (iii) Our proposed method outperforms SOD, WSOD, and EWSOD in zero-shot setup, thereby proving its fine-grained generalisation.

4.5. Part-Level Object Detection

Encouraged with the generalised fine-grained discriminative power of the proposed method in Tab. 3, we go a step further and ask: can we only detect a *part* (e.g., only 'head') of an instance? Due to lack of annotation, a quantitative evaluation of part-level object detection is infeasible. Nonetheless, we conduct a qualitative study by manually editing sketches to create partial sketches of a single part (e.g., only "head" of "horse"). Fig. 9 presents some results (for more see supplementary). We observe that (i) our proposed method can uniquely detect the sketched 'head' region of different objects. (ii) Detection performance is lower for ambiguous part sketches like 'leg' (e.g., front-leg, back-leg etc.) (iii) Since detection depends on region proposals from RPN, our model fails to detect tiny sketched parts. Tiny object detection [47] is a known challenge for traditional object detection [67].

4.6. Ablation

Selective Search v/s Edge Boxes v/s RPN Unlike the proposed method using pre-trained [45] RPN to generate 1000 box proposals, WSOD methods mostly use selective search [87] (SS) or edge boxes [111] (EB) that do not need pre-training using box annotation from visual genome [45].

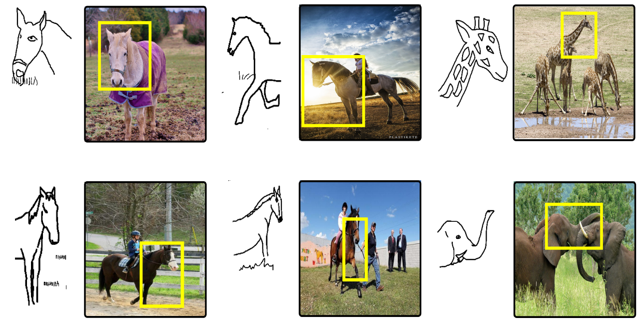

Figure 9. Unlike traditional object detection that detects an entire object (e.g., a "horse"), sketches can express fine-grained RoI to detect a specified part of an object (e.g., the "head" of a "horse").

Hence, for a fair comparison, we replace RPN with SS/EB drops $AP_{.5}$ performance by 1.3/2.6 on SketchyCOCO [25].

Influence of Classifier Refinement We observe $AP_{.5}$ improve by 2.4 and 1.1 for $K = 1 \rightarrow 2$ and $K = 2 \rightarrow 3$ respectively but a small drop of 0.2 for $K = 3 \rightarrow 4$.

Influence of Supervision from Photo Encoder in SBIR Although we can train an object detector using only pre-trained sketch encoder (trained on SBIR) via prototype learning, removing supervision from the photo encoder in SBIR drops 4.5 in $A_{.5}$ on SketchyCOCO [25].

4.7. Limitation and Future Works

Introducing fine-grained object detection using sketch opens several possibilities that we do not consider. Given multiple query sketches, currently we tread them as independent query embeddings. However, a user might be interested in detecting complex scenes (a "dog" on the right of a "person") with *multiple* objects that have meaningful *spatial* alignment. Future works can extend fine-grained object detection to semantic segmentation using complex sketches from the recently introduced FS-COCO [14] dataset.

5. Conclusion

We cultivate the expressiveness that human sketch bring for object detection. The proposed sketch-enabled object detection framework detects what *you* intend to express in *your* sketch – an object detector that is both instance-aware and part-aware. Accordingly, we design a novel prompt learning setup to marry CLIP and SBIR, to train a sketch-aware detector, that works without needing bounding box annotation, or class labels. To make our detector general-purpose, we further dictate it to work in a zero-shot manner. While SBIR is trained using object-level (single object) sketch/photo pairs, object works on image-level (multiple categories). We bridge this object and image-level gap using a data augmentation trick that improves robustness towards corruption and generalisation to out-of-vocab. The resulting framework outperforms both supervised, and weakly supervised object detectors on zero-shot setup.

References

[1] Stephan Alaniz, Massimiliano Mancini, Anjan Dutta, Diego Marcos, and Zeynep Akata. Abstracting sketches through simple primitives. In *ECCV*, 2022. 1, 2

[2] 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, 3

[3] 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. 3

[4] 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 sketch? *ACM TOG*, 2020. 2

[5] 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. 1, 2

[6] 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, 3

[7] Ayan Kumar Bhunia, Subhadeep Koley, Amandeep Kumar, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch2Saliency: Learning to Detect Salient Objects from Human Drawings. In *CVPR*, 2023. 2

[8] Ayan Kumar Bhunia, Aneeshan Sain, Parth Shah, Animesh Gupta, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Adaptive fine-grained sketch-based image retrieval. In *ECCV*, 2022. 2, 3

[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, 2, 3

[10] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016. 2, 3, 6

[11] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020. 3

[12] 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. 3

[13] 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. 2

[14] 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, 8

[15] Pinaki Nath Chowdhury, Tuanfeng Wang, Duygu Ceylan, Yi-Zhe Song, and Yulia Gryaditskaya. Garment ideation: Iterative view-aware sketch-based garment modeling. In *3DV*, 2022. 1, 2

[16] John Collomosse, Tu Bui, and Jin Hailin. Livesketch: Query perturbations for guided sketch-based visual search. In *CVPR*, 2019. 1, 3

[17] Ayan Das, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. Sketchode: Learning neural sketch representation in continuous time. In *ICLR*, 2022. 1

[18] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[19] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, 2019. 1, 3, 6, 7

[20] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *CVPR*, 2017. 3

[21] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 1997. 3

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvian Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2, 5, 6

[23] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 2, 6, 7

[24] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 6

[25] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: Image generation from freehand scene sketches. In *CVPR*, 2020. 6, 8

[26] Ross Girshick. Fast-rcnn. In *ICCV*, 2015. 2, 3, 6

[27] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2, 3

[28] Todd Goodwin, Ian Vollick, and Aaron Hertzmann. Isophote distance: A shading approach to artistic stroke thickness. In *NPAR*, 2007. 1

[29] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 2, 4, 7

[30] David Ha and Douglas Eck. A neural representation of sketch drawings. In *ICLR*, 2018. 1, 2, 5, 6, 7

[31] 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*, 2022. 1, 2

[32] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5, 6

[34] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019. 5

[35] Aaron Hertzmann. Why do line drawings work? a realism hypothesis. *Perception*, 2020. 1, 2

[36] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. In *NeurIPS*, 2019. 6

[37] Conghui Hu, Da Li, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Sketch-a-classifier: Sketch-based photo classifier generation. In *CVPR*, 2018. 1

[38] Conghui Hu, Da Li, Yongxin Yang, Timothy M. Hospedales, and Yi-Zhe Song. Sketch-a-segmenter: Sketch-based photo segmenter generation. *IEEE-TIP*, 2020. 2

[39] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 6

[40] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *CVPR*, 2017. 3

[41] John M. Kennedy. A psychology of picture perception: Images and information. *Jossey-Bass Publishers*, 1974. 1

[42] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *arXiv preprint arXiv:2210.03117*, 2022. 2

[43] Kazuma Kobayashi, Lin Gu, Ryuichiro Hataya, Takaaki Mizuno, Mototaka Miyake, Hirokazu Watanabe, Masamichi Takahashi, Yasuyuki Takamizawa, Yukihiro Yoshida, Satoshi Nakamura, Nobuji Kouno, Amina Bolatkan, Yusuke Kurose, Tatsuya Harada, and Ryuji Hamamoto. Sketch-based Medical Image Retrieval. *arXiv preprint arXiv:2303.03633*, 2023. 2

[44] 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. 2

[45] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 4, 6, 8

[46] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 3

[47] Chunggi Lee, Seonwook Park, Heon Song, Jeongun Ryu, Sanghoon Kim, Haejoon Kim, Sérgio Pereira, and Donggeun Yoo. Interactive multi-class tiny-object detection. In *CVPR*, 2022. 8

[48] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, 2016. 3

[49] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *CVPR*, 2021. 2, 4

[50] Tsung-Yi Lin, Piotr , Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 6

[51] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2

[52] 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. 2, 6, 7

[53] 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. 3, 6

[54] Feng-Lin Liu, Shu-Yu Chen, Yu-Kun Lai, Chunpeng Li, Yue-Ren Jiang, Hongbo Fu, and Lin Gao. DeepFaceVideoEditing: Sketch-based deep editing of face videos. *ACM TOG*, 2022. 1, 2

[55] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 2

[56] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013. 6

[57] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *ECCV*, 2022. 2

[58] Panagiotis Mouzenidis, Antonios Louros, Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. Multi-modal variational faster-rcnn for improved visual object detection in manufacturing. In *ICCV*, 2021. 3

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

[60] 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. 6

[61] 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. 3

[62] 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. 3

[63] Anran Qi, Yulia Gryaditskaya, Tao Xiang, and Yi-Zhe Song. One sketch for all: One-shot personalized sketch segmentation. *IEEE-TIP*, 2022. 2

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 5

[65] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[66] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[67] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection. In *NeurIPS*, 2015. 2, 3, 6, 8

[68] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, G. Schwing, Alexander, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *CVPR*, 2020. 6, 7

[69] Pau Riba, Sounak Dey, Ali Furkan Biten, and Josep Llados. Localizing infinity-shaped fishes: Sketch-guided object localization in the wild. *arXiv preprint arXiv:2109.11874*, 2021. 2

[70] Leo Sampaio Ferraz Ribeiro, Tui Bui, John Collomosse, and Moacir Ponti. Sketchformer: Transformer-based representation for sketched structure. In *CVPR*, 2020. 1, 2, 3

[71] 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

[72] 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. 3

[73] 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. 3

[74] 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

[75] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM TOG*, 2016. 1, 2, 3, 5, 6, 7

[76] Bilge Sayim and Patrick Cavanagh. What line drawings reveal about the visual brain. *Front. Hum. Neurosci.*, 2011. 1

[77] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *ICLR*, 2018. 6

[78] Feifei Shao, Long Chen, Jian Shao, Wei Ji, Shaoning Xiao, Lu Ye, Yueting Zhuang, and Jun Xiao. Deep learning for weakly-supervised object detection and localization: A survey. *Neurocomputing*, 2022. 2, 3

[79] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *CVPR*, 2019. 3

[80] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[81] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 3

[82] Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In *BMVC*, 2017. 1

[83] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE-TPAMI*, 2018. 6

[84] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, 2017. 6, 7

[85] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *ECCV*, 2018. 3

[86] Aditay Tripathi, Rajath R. Dani, Anand Mishra, and Anirban Chakraborty. Sketch-guided object localization in natural images. In *ECCV*, 2020. 2, 6

[87] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 2013. 4, 6, 8

[88] Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM TOG*, 2022. 2

[89] Kai Wang, Yifan Wang, Xing Xu, Xin Liu, Weihua Ou, and Huimin Lu. Prototype-based selective knowledge distillation for zero-shot sketch based image retrieval. In *ACM MM*, 2022. 6

[90] Xi Wang, Kathleen Ang, and Faramarz Samavati. Sketch-based editing and deformation of cardiac image segmentation. *PRISM*, 2022. 1, 2

[91] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. *https://github.com/facebookresearch/detectron2*, 2019. 6

[92] Minshan Xie, Menghan Xia, and Tien-Tsin Wong. Exploiting aliasing for manga restoration. In *CVPR*, 2021. 2

[93] Jun Xing, Li-Yi Wei, Takaaki Shiratori, and Koji Yatani. Autocomplete hand-drawn animations. *ACM TOG*, 2015. 2

[94] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, Zhanyu Ma, and Jun Guo. Sketchmate: Deep hashing for million-scale human sketch retrieval. In *CVPR*, 2018. 1, 3

[95] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*, 2022. 2

[96] Rui Xu, Zongyan Han, Le Hui, Jianjun Qian, and Jin Xie. Domain disentangled generative adversarial network for zero-shot sketch-based 3d shape retrieval. In *AAAI*, 2022. 2

[97] Jianwei Yang, Jiasen Lu, Stefan Leel, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, 2018. 3

[98] 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

[99] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 1, 3

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

[101] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 2, 3, 5

[102] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2, 3, 5

[103] Lvmin Zhang, Jinyue Jiang, and Yi Ji. Smartshadow: Artistic shadow drawing tool for line drawings. In *ICCV*, 2021. 2

[104] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *CVPR*, 2018. 3

[105] Xialin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018. 3

[106] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 3

[107] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 3, 5

[108] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Object as points. *arXiv preprint arXiv:1904.07850*, 2019. 2, 3

[109] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, 2019. 3

[110] Yaohui Zhu and Shuqiang Jiang. Deep structured learning for visual relationship detection. In *AAAI*, 2018. 3

[111] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 4, 8