Telling the What while Pointing to the Where: Multimodal Queries for Image Retrieval

Soravit Changpinyo Jordi Pont-Tuset Vittorio Ferrari Radu Soricut
Google Research
{schangpi,jponttuset,vittoferrari,rsoricut}@google.com

Abstract

Most existing image retrieval systems use text queries as a way for the user to express what *they are looking for. However, fine-grained image retrieval often requires the ability to also express* where *in the image the content they are looking for is. The text modality can only cumbersomely express such localization preferences, whereas pointing is a more natural fit. In this paper, we propose an image retrieval setup with a new form of multimodal queries, where the user simultaneously uses both spoken natural language (the* what*) and mouse traces over an empty canvas (the* where*) to express the characteristics of the desired target image. We then describe simple modifications to an existing image retrieval model, enabling it to operate in this setup. Qualitative and quantitative experiments show that our model effectively takes this spatial guidance into account, and provides significantly more accurate retrieval results compared to text-only equivalent systems.*

1. Introduction

Gargantuan amounts of pictures are taken and shared every day, at an ever accelerating pace. Finding the picture that one has in mind should be easier and faster than painfully scrolling through hundreds of pictures in a digital-camera roll. Building effective *image retrieval* systems for finding specific images among large collections is, therefore, of paramount importance. To speed the search up, image retrieval systems build an *index* that represents a collection of images by automatically analyzing their content [83, 53, 21, 66, 43, 62, 70, 71, 17, 37, 44, 39, 12].

A *query* is a description of what a user is looking for in an image, a translation of their mental model of the target image into a concrete form that can be understood by a retrieval system. At a coarse level, a query can be a list of specific classes of objects (*e.g.*, cars, people) the user wants to be contained by the target image [67]. At a finer-grained level is a natural language description of its contents [70, 71, 17, 37, 44, 39, 12]. The latter is the most common paradigm in the recent literature, partly due to the availability of captioning datasets that can be used as training and testing data [42, 10, 78, 55]. These types of queries generally focus on *what* is present in the image, but fall short of expressing *where* in the image the user expects it.

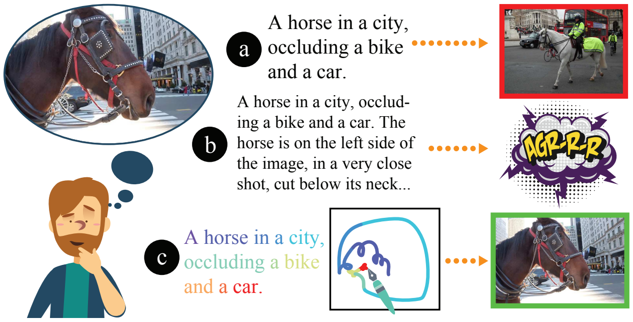
a
A horse in a city, occluding a bike and a car.
b
A horse in a city, occluding a bike and a car. The horse is on the left side of the image, in a very close shot, cut below its neck...
AGRRR
c
A horse in a city, occluding a bike and a car.

Figure 1: **Different types of textual queries** to represent the *what* and the *where* in the target image: (a) spatial information is usually lacking in textual descriptions and (b) it is cumbersome to express in written form, while (c) it is natural using mouse traces synchronized with the text.

As an example, consider the image in Fig. 1. One textual query could be "A horse in a city, occluding a bike and a car" (Fig. 1a). The retrieved image, while not the one the user had in mind, is a perfect match for this description: the *what* in the image is similar to the intended target. Expressing the *where* part using the textual query is not only cumbersome for the user to write, but also hard for the retrieval system to process (Fig. 1b).

In this paper, we propose a new query modality where the user describes the characteristics of the desired target image simultaneously using spoken natural language, the *what*, and mouse traces over an empty canvas, the *where* (Fig. 1c). Roughly pointing to an object's location comes naturally to humans [18, 13] and is an effective way of communicating the image layout the user has in mind. When the localization information is also temporally aligned with the natural language query, it becomes a natural grounding signal that can be exploited to make retrieval more precise.

We propose an image retrieval model that takes this new type of multimodal query as input. We start from an image-

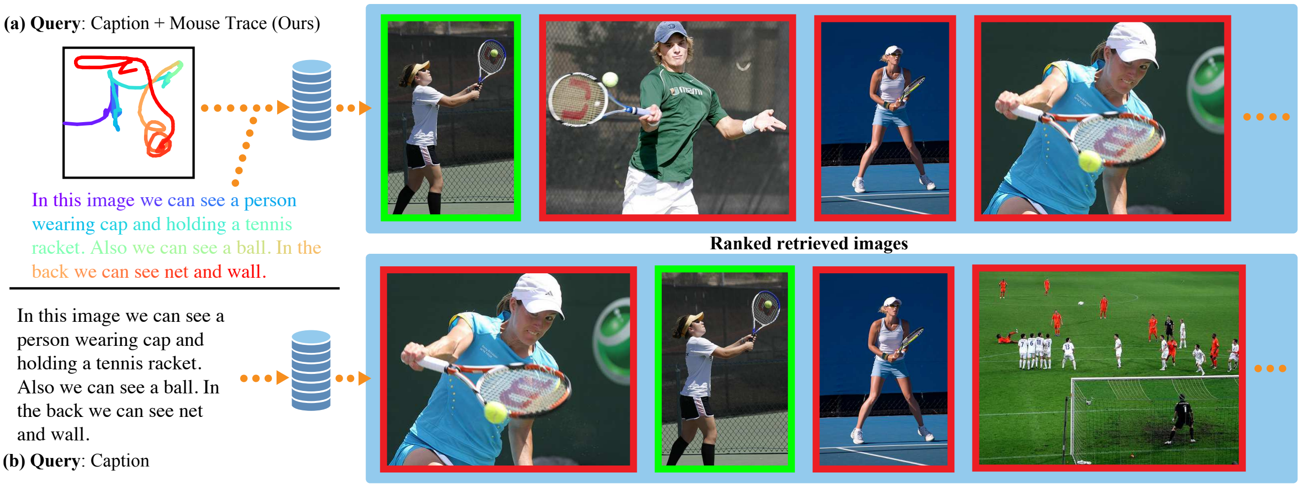
(a) Query: Caption + Mouse Trace (Ours)
In this image we can see a person wearing cap and holding a tennis racket. Also we can see a ball. In the back we can see net and wall.
Ranked retrieved images
In this image we can see a person wearing cap and holding a tennis racket. In Also we can see a ball. In the back we can see net and wall.
(b) Query: Caption

Figure 2: **Qualitative results**: Querying with (a) text and mouse traces, versus (b) only text. The target image is marked in green. Adding mouse traces to express the spatial location of the image content allows us to get a better retrieval result even given the same textual query. In this particular case, notice that the exact position of the racket and the ball allow the model to detect the correct target image.

to-text matching model that is repurposed as an image retriever by ranking image-text pairs according to their affinity, as in previous literature [32, 17, 37, 80]. We then augment the text input to also take the rough position in the blank canvas of each of the words into account (Fig. 4).

The data for training and evaluating such a model comes from Localized Narratives [56], a captioning dataset where annotators describe the images with their voice while simultaneously moving their mouse over the objects they are describing. The mouse traces are effectively grounding each word of the caption in the image. To use this data in an image retrieval scenario, we take the caption and corresponding mouse trace as input query, and the image on which the annotation is generated as target image.

Our experimental evaluation shows that this query modality provides a +7% absolute better recall (43% relative error rate decrease) for the top image compared to a model using only text queries. As we show in Fig. 2, having the rough location of the objects mentioned in the query restricts the space of plausible images and thus allows for more effective retrieval results.

In summary, our main contributions are:

(a) A novel query modality for fine-grained image retrieval that allows for a more natural specification of localization preferences.
(b) One concrete implementation of this idea that is simple and broadly applicable through a strong transformer-based model capable of incorporating the mouse traces.
(c) An experimental setup that suggests that Localized Narratives can be used to measure progress on this task.
(d) Empirical image retrieval results that demonstrate significant accuracy gains when the user is empowered with the ability to *point to the where*.

2. Related Work

Query Modality for Image Retrieval. The closest line of work to ours is **text**-based image retrieval (discussed in detail below), in which a natural language description serves as input to an image retrieval system. We augment this input with mouse traces drawn on an empty canvas to express where in the image the content should appear.

Other works also augment the text query with a certain **structure** that indicates the *where*, either limited to a closed vocabulary [28, 48, 25, 19, 61, 30] or derived automatically from the natural language descriptions [35, 38, 63, 72] (challenging in itself [77, 41, 79]). In contrast, our mouse traces cover all words and are drawn as input.

Drawing **sketches** on an empty canvas [62, 66, 43, 2, 81] was also used to represent an abstraction of an object category. We argue that expressing the *what* in natural language is significantly more intuitive and faster than drawing a sketch (*e.g*. compare using "horse" versus drawing one with enough detail to differentiate it from a zebra).

In content-based image retrieval, the query is an **image** and the target image depicts either (i) the same object [83, 58, 53, 21], typically from another viewpoint, at another time of the day, *etc*. (instance-level); or (ii) another object of the same category [5, 64, 52] (category-level). One can also add some natural language text that describes the desired modifications to the input image [22, 69, 11]. However, querying by image is a rather inflexible way to express what the user has in mind, as it already has its content fixed (both the what and the where).

We believe that our query modality makes the most efficient use of both natural language and mouse traces: the former to express a fine-grained *what* naturally and fast, and the latter to specify the *where* effectively and intuitively.

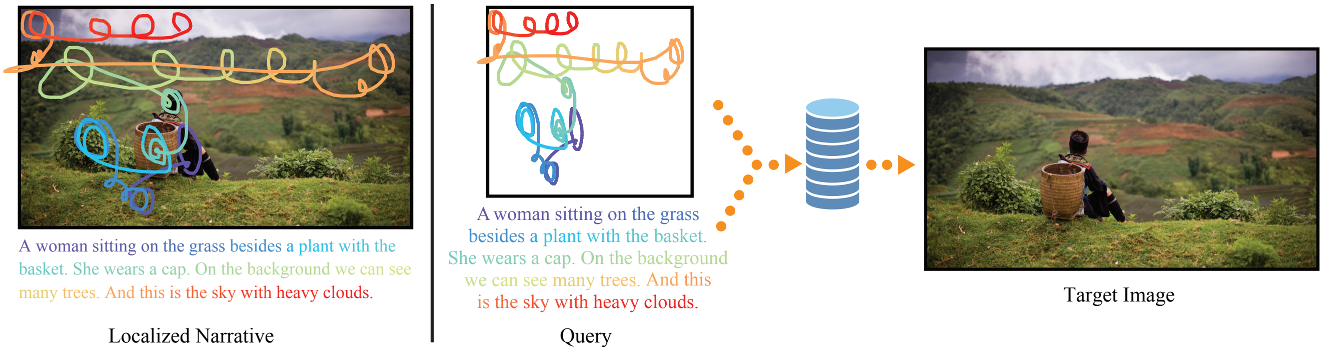
A woman sitting on the grass besides a plant with the basket. She wears a cap. On the background we can see many trees. And this is the sky with heavy clouds.
Localized Narrative
A woman sitting on the grass besides a plant with the basket. She wears a cap. On the background we can see many trees. And this is the sky with heavy clouds.
Query
Target Image

Figure 3: Localized Narratives annotations (left) can be transformed into training and testing data for image retrieval (right) by using the mouse traces as if they were drawn on a blank canvas, forming part of the query.

Approach to Caption-Based Image Retrieval. We focus on the most relevant works to ours, given the vast literature [84, 9] . Typical methods learn deep representations of images and texts, fuse them, and score the fused representations. To this end, a variety of factors have contributed to retrieval performance, including image and text features and encoders, types of cross-modal interaction, approaches to hard negative mining, loss functions, and pre-training data sources. [3, 7, 24] investigate the effects of these factors.

Convolutional and recurrent neural networks with late fusion were popular in earlier works [70, 32, 29, 33, 54, 47, 17, 26, 82], whereas recent works use transformers [44, 39, 12, 46, 80, 50, 49], graph neural networks [40, 73, 15], or architectures with more complex cross-modal interactions [37, 74, 6]. The latter often leverage region-based "bottom-up" visual features [1, 55, 37]. Moreover, multiple losses are explored, often requiring image-text triplets and hard negative mining [71, 17, 16, 51, 82, 76, 8]. Finally, pre-training image retrieval systems with large-scale image-text data sources has been shown to be extremely beneficial [20, 44, 40, 39, 12, 57, 4, 59, 27].

Our base image-text matching model (Sec. 4) follows most recent work [44, 39, 12] that uses transformers [68] with region-based Faster R-CNN visual features [60] trained on Visual Genome [35]. In addition, we explore the use of the Conceptual Captions [65], following [44], and Localized Narratives [56] as additional pre-training data sources. Finally, we adopt late image-text fusion [32, 17] due to its simplicity, scalability, and effectiveness over early-fusion-based approaches in scenarios where large-scale pre-training data and contrastive learning with a large batch size are used [4, 59, 27].

Building on top of our strong caption-based image retrieval system, our approach to connecting text tokens and image regions via box representations (Sec. 4, the orange boxes in Fig. 4) is largely inspired by position/location embeddings that are used extensively in recent work from both the computer vision and NLP communities [68, 14, 44, 56].

3. New Query Modality

Description. We propose a new query modality for image retrieval in which the user provides a mouse trace on a blank canvas and a natural language description that are synchronized with each other. This allows the user to seamlessly specify *what* they want (through language) and *where* they want it (through mouse traces, Fig. 1). We argue that pointing is a more natural means for taking into account the user's spatial preferences than existing approaches (Sec. 2).

***What+Where* Image Retrieval Setting**. As a second contribution, we construct the setting of *what+where* image retrieval, and leverage the recent Localized Narratives [56] dataset for this purpose. It is a collection of image-caption pairs, where each caption word is grounded in the image by a mouse trace segment (Fig. 3 left). They were obtained by annotators describing the images with their voice while simultaneously moving their mouse over the objects they were describing.

We transform the original Localized Narratives into useful annotations for image retrieval, by forming a query-image pair for each Localized Narrative as follows. We first strip away the image and keep only the caption and synchronized mouse trace, as if it had been drawn on an empty canvas. This forms our input query. Then we place the underlying image in our database, forming the intended target for that query (Fig. 3 right).

In the remainder of the paper we describe an image retrieval model that can operate in this setting (Sec. 4), and then experimentally show that this leads to more accurate results with respect to the user's intent (Sec. 5).

4. Technical Model

In this section, we describe an approach that enables a strong image retrieval system to operate in the *what+where* setting (Sec. 3). We first describe our base image retrieval system based on image-text matching (Sec. 4.1). We then

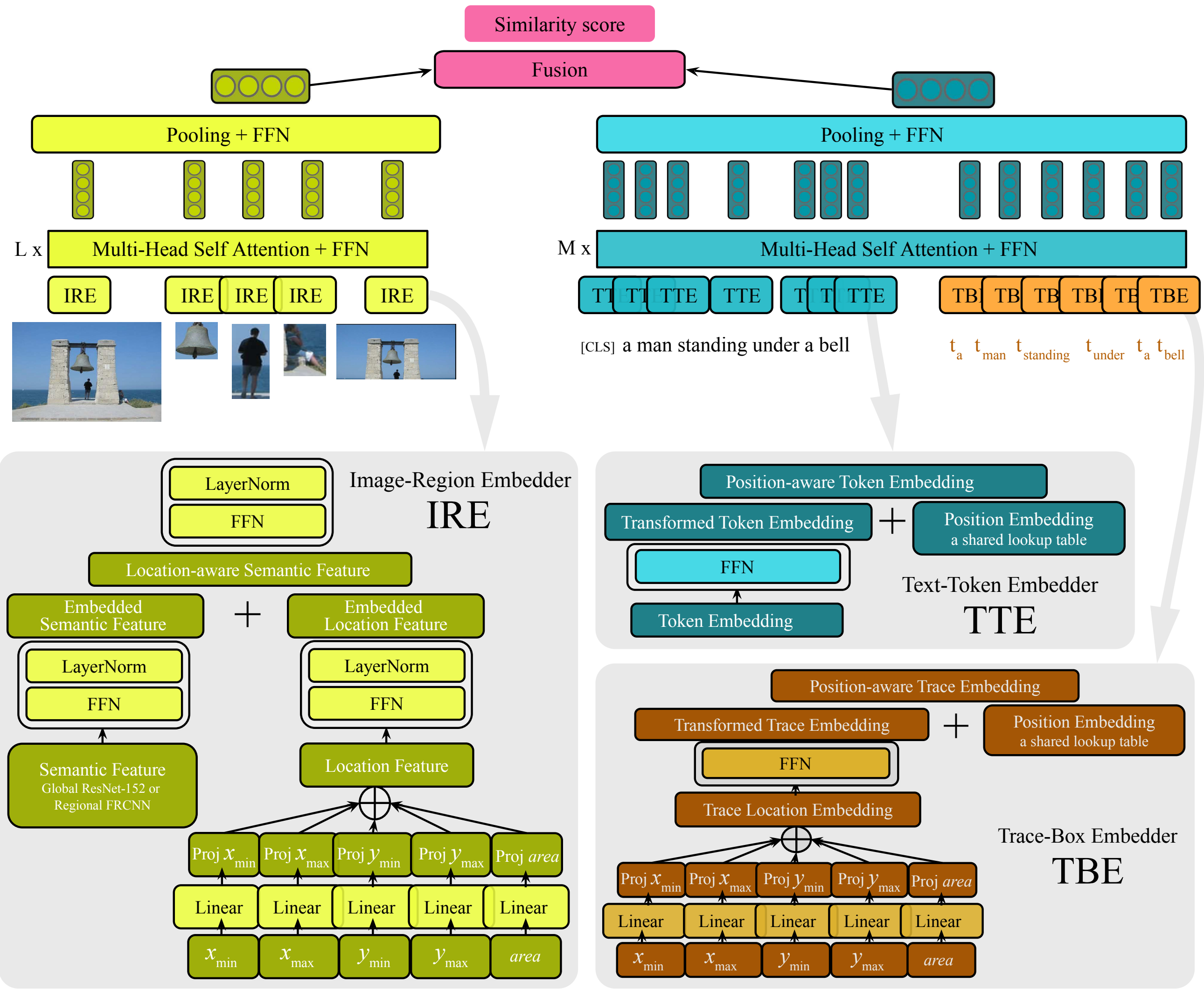

Figure 4: **Model**: Our model performs early fusion of text token representations (blue) and the box representations (orange) using transformers. Similarly, the model embeds the global and regional image embeddings (yellow). During the late fusion, the model combines the two streams and computes the similarity score between the image embedding and the text+traces embedding.

propose a modification to incorporate the extra input in the form of bounding boxes (Sec. 4.2) and show how we derive them from mouse trace segments (Sec. 4.3).

4.1. Base image retrieval model

As in much of the previous work (Sec. 2), we turn the standard text-based image retrieval problem into learning image-text matching. Let us denote by $x = (x_1, \ldots, x_N)$ a set of feature vectors representing the image (*e.g*. the output of a CNN or an object detector run on the image) and $y = (y_1, \ldots, y_K)$ a set of feature vectors representing the text (*e.g*. random or pre-trained character/subword/wordpiece/word embeddings of text tokens). We fix both N and K in our experiments and use padding and masking as necessary.

Our base model learns a similarity function

$$s(x, y) = p\big(f(x), g(y)\big), \tag{1}$$

where f, g, and p are an image *tower*, a text *tower*, and an image-text *fuser*, respectively. Each tower reduces a set of feature vectors into a fixed-length vector and the fuser combines them to produce the final score. In this paper, we choose the dot product as the image-text fuser p and use the symmetric batched contrastive loss for parameter estimation, treating all other image-text pairs within the batch of size B as negative examples:

$$L = \frac{1}{2}(L_{x \rightarrow y} + L_{y \rightarrow x}) \tag{2}$$

$$L_{a \rightarrow b} = \sum_{i}^{B} \log \frac{\exp\left(s(a^{(i)}, b^{(i)})\right)}{\sum_{j=1}^{B} \exp\left(s(a^{(i)}, b^{(j)})\right)} \tag{3}$$

At training time, we learn the parameters of f, g, and p from a collection of image-text pairs. At test time, given a query text y', we use the learned p to compute a similarity score between y' and each of the images x in the database. We then output a ranking of all database images sorted by their score, which represents our retrieval result.

Figure 4 (without the trace inputs and the trace box embedder, in orange) illustrates our base model. We adopt a two-stream model in which the image tower f and the text tower g do not share weights. Each tower consists of three components: (i) an embedder, (ii) a contextualizer, and (iii) a pooler. Both towers use a 6-layer Transformer architecture [68] for (ii) and mean pooling for (iii). We use the vanilla architecture, where each transformer layer consists of a multi-head self-attention and feed-forward fully-connected network. We refer the reader to [68] for details about the Transformer architecture. Below, we describe the first component of each tower.

The Image Region Embedder (IRE). The input of the IRE is a fixed-length feature vector representing the whole image (CNN output) or a region of the image (one of an object detector's region outputs). The IRE transforms each of these feature vectors into an embedded semantic feature vector, and their corresponding 5D geometric feature of box coordinates (x_{min}, x_{max}, y_{min}, y_{max}) and box area into an embedded location feature. Adding the two together gives a location-aware semantic feature vector of the region, which goes through a 2-layer Multi-Layer Perceptron (MLP) before it is used as input of the image transformer.

The Text Token Embedder (TTE). Given a fixed-length vector representing a text token (a character, a subword, a word, *etc*.), the TTE applies a 2-layer MLP and adds a position embedding to the output, resulting in a token embedding that is position aware.

We will use what we described here as our base image retrieval model throughout the paper, unless stated otherwise. The end of Section 2 discusses our modeling choices with respect to prior work. Additionally, we verify that our implementation is strong, achieving a Recall@1 of 36.9 on the task of zero-shot image retrieval on Flickr30k [78, 55] with Conceptual Captions [65] as pre-training data, outperforming ViLBERT [44], a leading early-fusion, larger model.

4.2. Incorporating mouse traces

Our high-level idea is to inject the traces to our base model by introducing the trace-box embedding (TBE) module whose encoded 1D text positions and 2D image locations act as a glue between text tokens and image regions.

Given mouse traces t as an additional input, we modify our similarity function in (1) by injecting it into the text stream of the model:

$$s(x, y) = p\big(f(x), h(y, t)\big), \tag{4}$$

where h is a text-trace fuser/embedder, and f and p are the same as in (1).

Similarly to the setting in Section 4.1, at training time we learn the parameters of f, h, and p from a collection of positive image-text-trace triplets. At test time, given a query text y' and its corresponding query trace t', we use the learned p to compute a similarity score between (y', t') and each of the images in the database and output a ranking of the images. Note that our setting assumes the existence of traces both during training and testing, as we envisage these new "text+trace" queries to be cast by users using an interface analog to the one used for Localized Narratives annotation [56].

Figure 4 depicts our full model, with the components described in Section 4.1 unchanged. The extra component, the mouse trace input t, is encoded in the form of a sequence of boxes by the Trace Box Embedder (TBE, bottom right of Fig. 4), described below, and then fuse it with the text query.

The Trace Box Embedder (TBE). Analogous to the location input of IRE, each of the trace boxes is represented using a 5D vector consisting of coordinates and area (x_{min}, x_{max}, y_{min}, y_{max}, *area*). Since these boxes correspond to parts of the text query, they also have the notion of 1D time-location "position" in the query. Thus, we add a position embedding to the transformed trace embedding vector, resulting in a trace embedding vector that is both location-aware (visually) and position-aware (textually).

Fusing texts and traces. We concatenate all the outputs of TTE (Sec. 4.1) and TBE, and use the result as input to the text-trace transformer. We believe this is both simple and powerful, as the transformer self-attention layers allow text tokens and trace boxes to attend to each other freely. Note that it is this early fusion of text and traces that is capable of modeling where in the image certain parts of the query are expected to be relevant.

4.3. From mouse traces to bounding boxes

A Localized Narrative annotation has each utterance in the caption associated with a mouse trace segment, which *grounds* the utterance on the image. In other words, it defines the rough position in the image where the semantic content from the utterance (the *what*) is located (the *where*).

The mouse trace segment for a certain utterance corresponds to the sequence of image points the mouse traversed during the time interval (t_1, t_2) while the annotator spoke the utterance. We observe that the mouse traces *around* the time when an utterance was spoken can still refer to the same utterance, so we explore adding *temporal padding* t_p to better define the trace segment. That is, we consider the trace segment in the time interval ($t_1 - t_p$, $t_2 + t_p$).

As our model inputs bounding boxes that locate the query in the image (Fig 4), we convert the mouse trace segments to boxes. We start from the tightest box (Fig. 5, yellow) that fully contains the trace segment defined by the

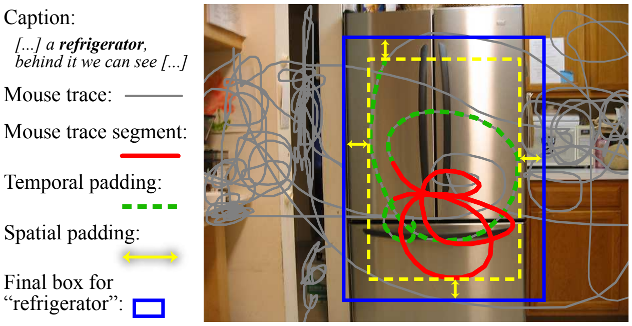
Caption:
[...] a refrigerator, behind it we can see [...]
Mouse trace:
Mouse trace segment:
Temporal padding:
Spatial padding:
Final box for "refrigerator":

Figure 5: **From a mouse trace segment to its box**: We first prolong the mouse trace segment along the temporal dimension (green), and then we add spatial 2D padding (blue).

time segment $(t_1 - t_p, t_2 + t_p)$, and we enlarge it in all dimensions by a certain *spatial padding* s_p (Fig. 5, blue).

5. Experiments

5.1. Setup

Overview. The main goal of our experiments is to test whether incorporating mouse traces into the query improves the accuracy of image retrieval. We will test this hypothesis in multiple scenarios, including several vision-and-language pre-training settings, inspired by [44, 45, 39, 12].

Datasets. Table 1 summarizes the main datasets used in our experiments. We use one dataset as our main task with multiple evaluation sets and two datasets as pre-training data sources. For the **main task**, we use *Flickr30k Localized Narratives (Flickr30k LocNar)* [75]. This comes with the same set of 31,783 images as Flickr30k, but we use the Localized Narratives captions and their synchronized mouse traces instead of the original captions without mouse traces. We either train or fine-tune our models on the training split (29,783 images), perform model selection on the validation split (1,000 images), and report our quantitative results on the test split (1,000 images, Sec. 5.2). We further evaluate on different splits of Flickr30k and on two out-of-domain datasets: *COCO Localized Narratives (COCO LocNar)* and *ADE20K Localized Narratives (ADE20K LocNar)* [75], without additional fine-tuning (Sec. 5.3).

For **pre-training**, we use the training splits of *Conceptual Captions (CC)* [65] and *Open Images Localized Narratives (OID LocNar)* [75]. The former contains 3.3M pairs of (image, alt-text) harvested from the web. The latter is a subset of the 9M images in the Open Images dataset [36, 34] that is annotated with Localized Narratives. We use these annotations to pre-train both the image and the language branches of our model (Fig. 4). We explore two pre-training data sources due to their complementary strengths: CC is larger-scale with more semantically specific terms (*e.g.* croissant *vs*. food), while the style of descriptions in OID LocNar is more similar to our target task of Flickr30k LocNar. Furthermore, the existence of mouse traces in the OID LocNar enables us to explore incorporating traces *during pre-training* (using the model in Sec. 4.2).

Stage	Dataset	Size	#Tok/cap
Main	Flickr30k LocNar	31,783	57.1
Pretrain	Conceptual Captions (train)	3.3M	10.3
Pretrain	Open Images LocNar (train)	507K	35.5

Table 1: **Main datasets** used in our experiments. LocNar is short for Localized Narratives. #tok/cap is the average number of tokens per caption.

Settings. We consider the from-scratch (no pre-training) setting and multiple pre-training settings: (i) on CC only, (ii) on OID LocNar only (with and without mouse traces), and (iii) on CC followed by OID LocNar (with and without mouse traces). Setting (iii) is based on our intuition (which will be verified in the experiments) that the domain of OID LocNar is closer to that of Flickr30k LocNar.

In each of these settings, we then compare the retrieval performance of the model with text-only queries (Sec. 4.1) against that of the model with text+trace queries (Sec. 4.2) on the Flickr30k LocNar. Note that when pre-training is involved, we make use of all available pre-trained weights and randomly initialize the rest (*e.g.* the TBE weights when the mouse traces are not used during pre-training).

Evaluation metrics. We use Recall@K (denoted as R@K for K=1,5,10): the percentage of images in the test set for which the target image falls within the top-K of the model's output ranking, when using its corresponding text(+trace) as the input query. We also report mean Average Precision (mAP) in our main experiments. Since we observe a consistent trend with that from R@K, we focus on R@K on the other experiments.

Implementation details. We use subtokens and random embeddings to represent text units (*e.g.* "standing" → "stand", "ing") . We use a vocabulary size of 10,000. We represent an image with two types of features: A 2048D global feature vector of ResNet152 [23] and top 16 regional feature vectors by a Faster-RCNN [60] trained on Visual Genome [35] with a ResNet101 backbone [23]. Our box coordinates and area of a region are represented with relative numbers between 0 and 1, such that the 5D location information $x_{\min}$, $x_{\max}$, $y_{\min}$, $y_{\max}$, and *area* of the whole image is 0.0, 0.0, 1.0, 1.0, 1.0, respectively. We concatenate the two sets of features and permute the 16 regional vectors during training. We use Adam [31] and contrastive learning treating all other image-text pairs in each batch as negatives (Sec. 4.1). We tune an initial learning rate but always use a linear warm-up of 20 epochs and multiply the learning rate by 0.95 every 25 epochs after that.

Scenario		Recall@K=			mAP
Pre-train?	Query	1	5	10	
	text	63.5	87.4	92.8	74.0
	text+trace	68.2	88.8	94.4	77.7
✓	text	83.4	97.6	98.5	89.7
✓	text+trace	**90.6**	**98.2**	**99.4**	**94.0**

Table 2: **Main results.** The image retrieval performance on the Flickr30k LocNar 1K test set.

Pre-training		Final	Recall@K=		
data	query	query	1	5	10
CC	text	text	74.2	93.9	96.2
	text	text+trace	79.5	95.1	97.8
OID LocNar	text	text	81.5	97.6	99.0
	text	text+trace	83.9	97.1	98.5
	text+trace	text+trace	**90.6**	98.2	**99.4**
CC → OID LocNar	text	text	83.4	97.6	98.5
	text	text+trace	83.5	97.2	98.2
	both	text+trace	90.2	**98.4**	99.0

Table 3: **Pre-training with different data sources and query modalities** affects image retrieval performance on the Flickr30k LocNar 1K test set.

5.2. Main Results

Table 2 compares the image retrieval performance on Flickr30k LocNar of the models using the text-only queries and the ones using text+trace queries, i.e. our new *what+where* setting (Sec. 3).

Regardless of whether we perform pre-training, incorporating the mouse trace (the "where") leads to significant gains in absolute R@1: +4.7% without pre-training (Row 1 *vs*. Row 2), and +7.2% with pre-training (Row 3 *vs*. Row 4). Overall, the best result is obtained when we both pre-train and inject the trace to our model; we improve over the baseline model by an absolute +27.1%, +10.8%, +6.6% in R@{1,5,10}, and +20.0% in mAP (Row 1 *vs*. Row 4).

Our results suggest that *the top retrieved image will be much more accurate if the user gets to "point to the where"*. Furthermore, *pre-training and our new query modality are complementary*: the main benefit of pre-training with the text-only query modality is on improving "telling the what".

5.3. Detailed Results and Ablation Studies

Pre-training data sources. In Table 3, we observe that OID LocNar is superior to CC as a pre-training data source for this task, supporting our intuition that the domain of the OID LocNar is closer to that of Flickr30k LocNar. However, they are complementary when the trace is not involved.

Pre-training query modality. In Table 3, the largest benefit of pre-training is observed when text+traces are used during both the pre-training and final stages; in the case of OID LocNar, this leads to the best R@1 of 90.6. When this is not possible (*i.e*. pre-training data does not come with traces as in CC), we still observe significant improvements in R@1 when using text+trace in the final stage. This suggests that text+trace queries are *generally* superior, working robustly across pre-training scenarios.

Pre-training data	Query	Recall@K=		
		1	5	10
CC	text	21.0	42.2	54.0
OID LocNar	text	79.0	95.7	98.3
CC → OID LocNar	text	79.1	95.7	97.9
OID LocNar	text+trace	88.0	97.7	99.1
CC → OID LocNar	text+trace	86.7	98.0	98.8

Table 4: **Zero-shot image retrieval** performance on the Flickr30k LocNar 1K test set. Best viewed together with Table 3.

Eval data	Query	Recall@K=			mAP
		1	5	10	
ADE20K	text	47.4	73.8	84.6	59.5
	text+trace	**60.3**	**84.1**	**90.7**	**70.7**
COCO	text	73.7	94.3	97.6	82.5
	text+trace	**82.4**	**96.6**	**98.4**	**88.7**

Table 5: **Out-of-domain evaluation.** Image retrieval performance on ADE20K LocNar val (2K images) and COCO LocNar val (averaged over 5-fold 1K images).

Zero-shot image retrieval. We test our models when they have not seen any image of the test domain (Flickr30k LocNar), i.e. only trained on the pre-training data and evaluated on the Flickr30k LocNar test set (Tab. 4). Together with Table 3, we see that fine-tuning on Flickr30k LocNar is beneficial in all cases. Notably, the zero-shot performance of the CC model is much lower than the one fine-tuned on OID LocNar, indicating a big domain gap between Conceptual Captions and Localized Narratives-style datasets.

Out-of-domain evaluation. We take the best text-only and text+trace models (last two rows of Tab. 2) as-is and evaluate their performance on two additional datasets, ADE20K LocNar and COCO LocNar (without fine-tuning on their training sets, Tab. 5). The text+trace modality is still far superior to the text-only one (+12.9% on ADE20K and +8.7% in R@1 on COCO). We stress that these datasets are in a different domain than the training sets (Open Images and Flickr30k). Thus, our improvements cannot be achieved simply by overfitting on the training domain.

Statistical significance. We re-split the union of the training and test subsets of Flickr30k LocNar (keeping val intact) and then re-train and re-evaluate our best text+trace model (last row in Tab. 2). Over 5 re-splits, the R@1 is 90.6% ± 0.9, which suggests that our gain of +7.2% over the text-only model is statistically very significant.

Trace-only query modality. In Tab. 6, our trace-only query achieves a R@1 of 14.5 without pre-training (Row 3). When compared to the text+trace and text-only queries

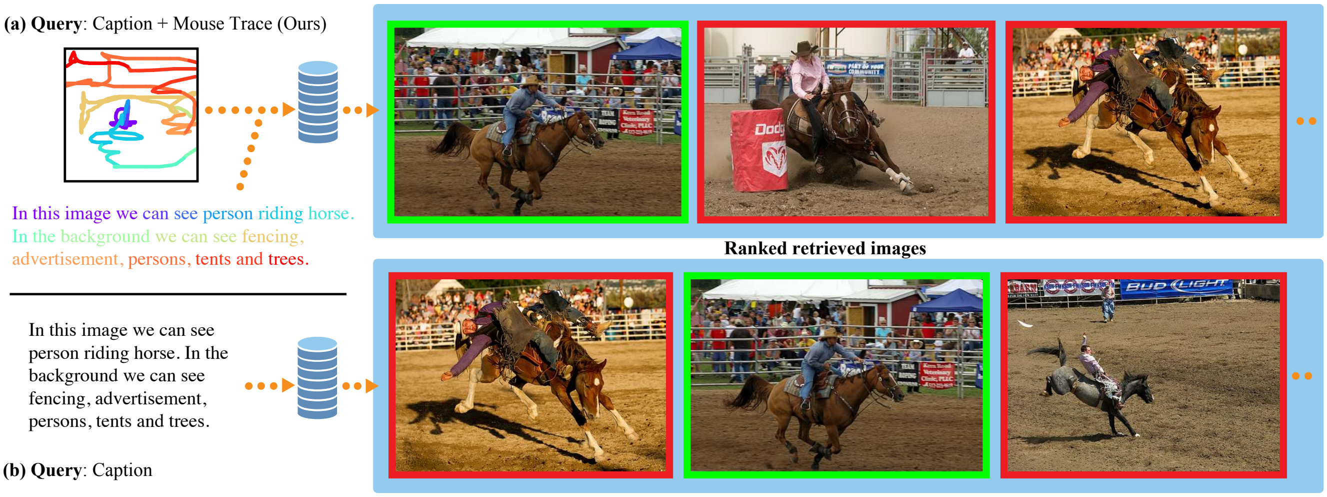
(a) Query: Caption + Mouse Trace (Ours)
In this image we can see person riding horse. In the background we can see fencing, advertisement, persons, tents and trees.
Ranked retrieved images
In this image we can see person riding horse. In the background we can see fencing, advertisement, persons, tents and trees.
(b) Query: Caption

Figure 6: **Qualitative results**: Comparison between our best method (a) to that without trace supervision (b). In green, the target image that corresponds to the query on the left.

Image		Text		Trace	Recall@K=		
sem	loc	tok	pos		1	5	10
✓	✓	✓	✓	✓	68.2	88.8	94.4
✓	✓	✓	✓		63.5	87.4	92.8
✓	✓			✓	14.5	31.7	42.7
✓		✓	✓	✓	66.8	89.4	94.5
✓	✓	✓		✓	65.1	87.8	93.9

Table 6: **Benefits of retrieval components** on the image retrieval performance on the Flickr30k LocNar 1K test set. The image features consist of semantic (sem) and 2D location (loc) embeddings. The text features consist of token (tok) and 1D position (pos) embeddings. See Sec. 4 and Fig. 4 for details of these components.

(Row 1-2), this shows that while text plays a major role, both elements are important to achieve strong performance.

Position and location embeddings. Table 6 also investigates the benefits of 1D word position (TTE) and 2D image region location (IRE) embeddings, both of which are connected to TBE in Figure 4. We find that they are important as their absences lead to degradation in the top retrieved image (Row 1 *vs*. Row 4-5).

Drawing traces on an empty canvas. Analog to all modern text-to-image retrieval works that leverage image captioning datasets, our experiments are limited by the fact that our trace queries were drawn while the annotator was looking at the target image. What if the traces were drawn on an empty canvas? We select 7 images from the Flickr30k LocNar test set on which our best text+trace model retrieved the correct image in the top rank, but our best text-only model did not. We then ask an annotator to briefly look at these 7 images, and then draw a trace for each image on an empty canvas, while reading the original caption (*without seeing the image*). In this scenario, our text+trace model retrieves the correct image in 6 out of 7 cases, suggesting that our model can maintain high accuracy even when the traces are not exactly aligned with the image regions.

Architecture. In the supplementary material (Sec. B), we experiment with the number of layers of text (M) and image (L) transformer encoders of our model (Fig. 4). We find that the benefit of the text+trace query modality over the text-only one generalizes to all our ablation studies.

Qualitative results. Figure 6 shows qualitative results, comparing our best model with text+trace query and our best model with text-only query. Note that the exact positions of the fence and the advertisement allows the model to distinguish between images with very similar content. More qualitative results are in Figure 2 and in the supplementary material (Sec. C).

6. Conclusions

In this paper, we propose a new query modality for content-based image retrieval systems where the user describes the characteristics of the desired target image simultaneously using spoken natural language (the "what") and mouse traces over an empty canvas (the "where"). We present an image retrieval model that takes this new type of multimodal query as input. We train and evaluate our model using Localized Narratives, where the caption and its corresponding mouse trace is used as input query, and the corresponding image as target. Our experimental evaluation shows that this query modality provides a 43% relative error rate decrease for the top image compared to the model that only uses text-based queries.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 3

[2] Tu Bui, Leonardo Ribeiro, Moacir Ponti, and John Collomosse. Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. *Computers & Graphics*, 71:77–87, 2018. 2

[3] Andrea Burns, Reuben Tan, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. Language features matter: Effective language representations for vision-language tasks. In *ICCV*, 2019. 3

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 3

[5] Ken Chatfield, Karen Simonyan, and Andrew Zisserman. Efficient on-the-fly category retrieval using convnets and gpus. In *ACCV*, 2014. 2

[6] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *CVPR*, 2020. 3

[7] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, 2021. 3

[8] Tianlang Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. In *ECCV*, 2020. 3

[9] Wei Chen, Yu Liu, Weiping Wang, Erwin Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S. Lew. Deep image retrieval: A survey. *arXiv*, 2021. 3

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv*, 2015. 1

[11] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *CVPR*, 2020. 2

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV*, 2020. 1, 3, 6

[13] Herbert Clark. Coordinating with each other in a material world. *Discourse Studies*, 7:507–525, 10 2005. 1

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3

[15] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, 2021. 3

[16] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017. 3

[17] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2017. 1, 2, 3

[18] Chaz Firestone and Brian J. Scholl. "Please tap the shape, anywhere you like" shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological science*, 25(2):377–386, 2014. 1

[19] Ryosuke Furuta, Naoto Inoue, and Toshihiko Yamasaki. Efficient and interactive spatial-semantic image retrieval. *Multimedia Tools and Applications*, 78(13):18713–18733, 2019. 2

[20] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, 2014. 3

[21] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, 2016. 1, 2

[22] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Schmidt Feris. Dialog-based interactive image retrieval. In *NeurIPS*, 2018. 2

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

[24] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *arXiv*, 2021. 3

[25] Ryota Hinami, Yusuke Matsui, and Shin'ichi Satoh. Region-based image retrieval revisited. In *ACM Multimedia*, 2017. 2

[26] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017. 3

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 3

[28] Justin Johnson, Ranjay Krishna, Michael Stark, Li Jia Li, David A. Shamma, Michael S. Bernstein, and Fei Fei Li. Image retrieval using scene graphs. In *CVPR*, 2015. 2

[29] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 3

[30] Mert Kilickaya and Arnold WM Smeulders. Structured visual search via composition-aware learning. In *WACV*, 2021. 2

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[32] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv*, 2014. 2, 3

[33] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015. 3

[34] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun

Feng, Dhyanesh Narayanan, and Kevin Murphy. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://g.co/dataset/openimages*, 2017. 6

[35] 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*, 123(1):32–73, 2017. 2, 3, 6

[36] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020. 6

[37] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018. 1, 2, 3

[38] Ang Li, Jin Sun, Joe Yue-Hei Ng, Ruichi Yu, Vlad I. Morariu, and Larry S. Davis. Generating holistic 3d scene abstractions for text-based image retrieval. In *CVPR*, 2017. 2

[39] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 1, 3, 6

[40] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019. 3

[41] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, 2017. 2

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1

[43] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017. 1, 2

[44] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1, 3, 5, 6

[45] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 6

[46] Xiaopeng Lu, Tiancheng Zhao, and Kyusong Lee. VisualSparta: Sparse transformer fragment-level matching for large-scale text-to-image search. *arXiv*, 2021. 3

[47] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, 2015. 3

[48] Long Mai, Hailin Jin, Zhe Lin, Chen Fang, Jonathan Brandt, and Feng Liu. Spatial-semantic image search by visual feature synthesis. In *CVPR*, 2017. 2

[49] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *arXiv*, 2020. 3

[50] Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. Transformer reasoning network for image-text matching and retrieval. In *ICPR*, 2020. 3

[51] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, 2017. 3

[52] Shawn Newsam, Baris Sumengen, and BS Manjunath. Category-based image retrieval. In *ICIP*, 2001. 2

[53] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017. 1, 2

[54] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 3

[55] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30K Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017. 1, 3, 5

[56] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 2, 3, 5

[57] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. ImageBERT: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv*, 2020. 3

[58] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *TPAMI*, 41(7), 2018. 2

[59] 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. *arXiv*, 2021. 3

[60] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 3, 6

[61] Luca Rossetto, Ralph Gasser, and Heiko Schuldt. Query by semantic sketch. *arXiv*, 2019. 2

[62] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies. *ACM Transactions on Graphics*, 35(4):1–12, 2016. 1, 2

[63] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Workshop on Vision and Language*, 2015. 2

[64] Gaurav Sharma and Bernt Schiele. Scalable nonlinear embeddings for semantic category-based image retrieval. In *ICCV*, 2015. 2

[65] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 3, 5, 6

[66] 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. 1, 2

[67] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010. 1

[68] 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. 3, 5

[69] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-An empirical odyssey. In *CVPR*, 2019. 2

[70] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014. 1, 3

[71] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 1, 3

[72] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *WACV*, 2020. 2

[73] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *WACV*, 2020. 3

[74] 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. 3

[75] Website. Localized Narratives Data and Visualization. https://google.github.io/localized-narratives, 2020. 6

[76] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, and Heng Tao Shen. Universal weighting metric learning for cross-modal matching. In *CVPR*, 2020. 3

[77] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 2

[78] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 1, 5

[79] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018. 2

[80] Bowen Zhang, Hexiang Hu, Vihan Jain, Eugene Ie, and Fei Sha. Learning to represent image and text with denotation graph. In *EMNLP*, 2020. 2, 3

[81] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *ECCV*, 2018. 2

[82] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, 2018. 3

[83] Liang Zheng, Yi Yang, and Qi Tian. SIFT meets CNN: A decade survey of instance retrieval. *TPAMI*, 40(5):1224–1244, 2018. 1, 2

[84] Wengang Zhou, Houqiang Li, and Qi Tian. Recent advance in content-based image retrieval: A literature survey. *arXiv*, 2017. 3