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

Subhadeep Koley[1,2], Ayan Kumar Bhunia[1], Aneeshan Sain[1], Pinaki Nath Chowdhury[1], Tao Xiang[1,2], Yi-Zhe Song[1,2]

[1]SketchX, CVSSP, University of Surrey

[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence

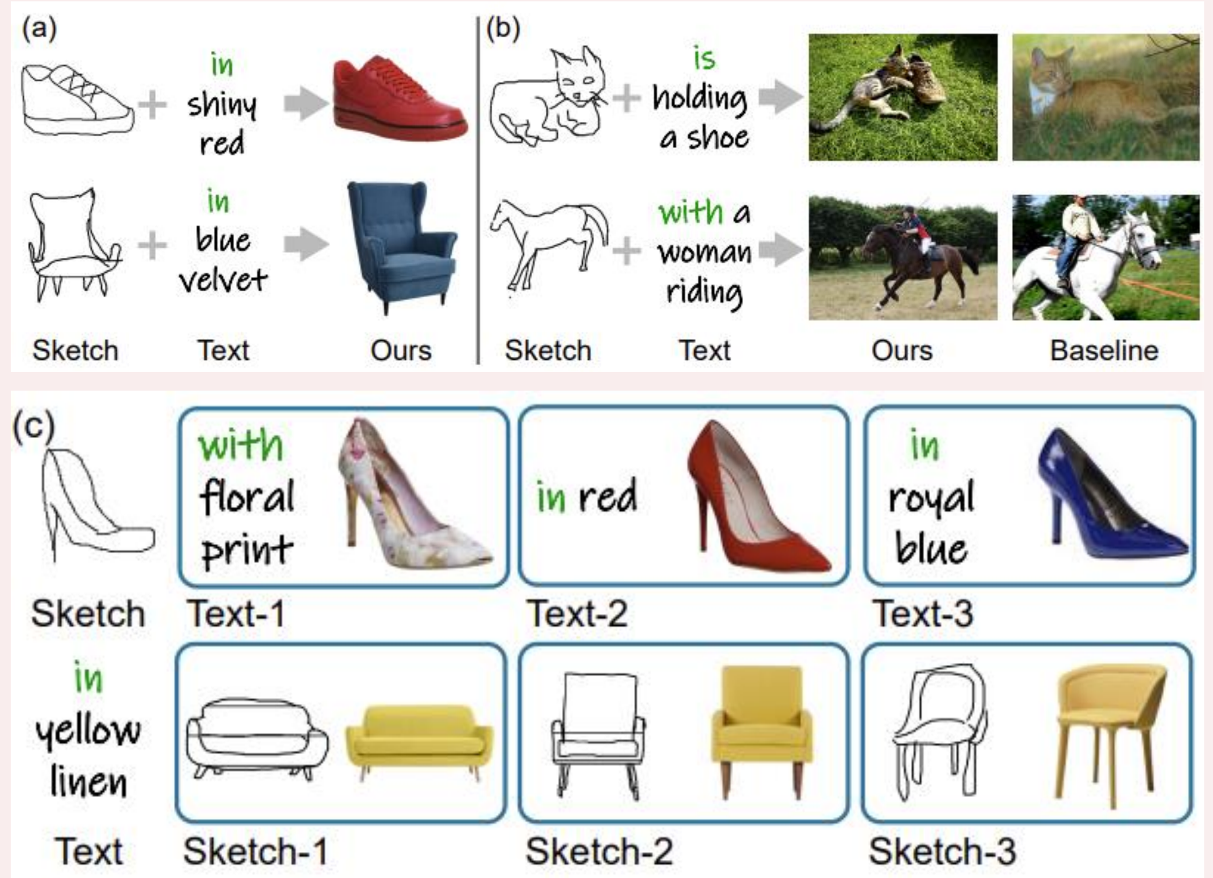
CVPR JUNE 17-21, 2024 SEATTLE, WA

SketchX — UNIVERSITY OF SURREY — CVSSP Centre for Vision, Speech and Signal Processing

Got 2 Minutes? Start here ⬇

Summary

- Addresses the challenge of **fine-grained image retrieval** by leveraging the synergy between freehand sketches and text captions.
- Effectively combining sketches and text using **pre-trained CLIP model**, *eliminating the need* for extensive fine-grained textual descriptions.
- Unlocks **novel applications** like *object-sketch-based scene retrieval*, *domain attribute transfer*, & *sketch+text-based fine-grained generation*.

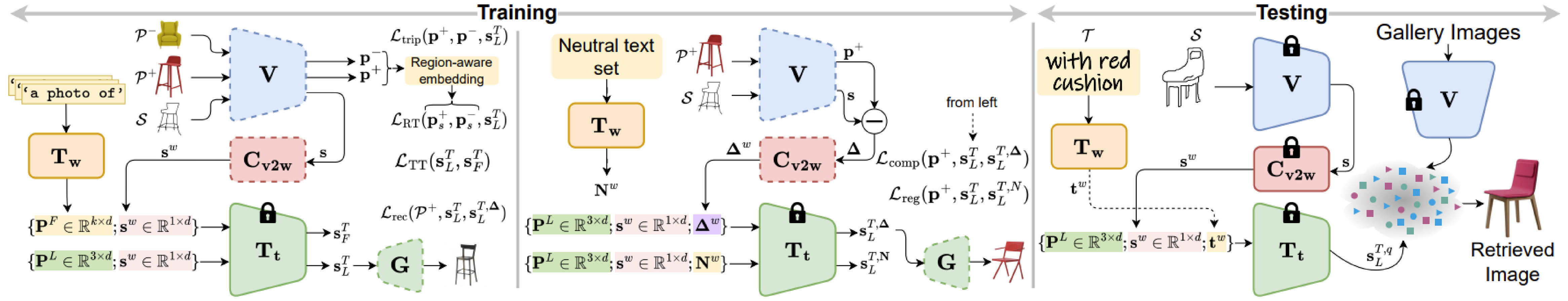


Problems

- Existing sketch+text-based retrieval methods have predominantly focused on **scene-level or category retrieval**.
- Existing fine-grained sketch-photo datasets **lack paired fine-grained textual description**.
- Textual caption generation via SoTA captioners often results in **noisy and inaccurate description** in case of abstract freehand sketches.

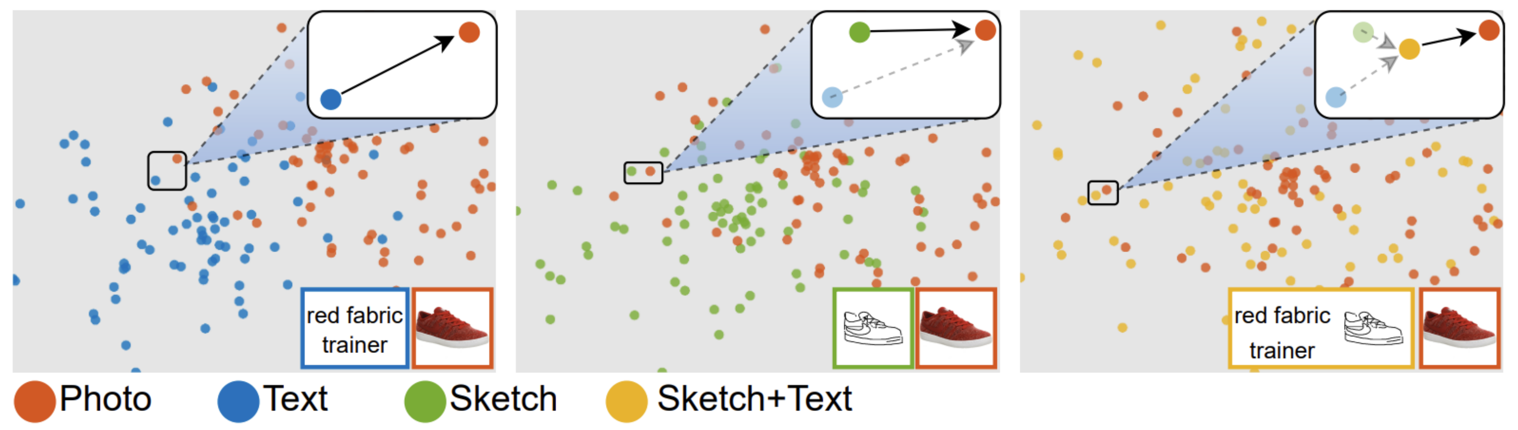
Solutions

- We convert input sketches into fine-grained textual equivalents, referred to as a **"pseudo-word token"**. This token, when combined with text input, forms a fine-grained textual query that seamlessly integrates both sketch and text features.
- We hypothesise that the fine-grained description embedded in a **photo (P)** can be approximated by that of a **sketch (S)** plus **text (T)**, leading to **T = P - S**. This relationship illustrates how the absence of **T** can be approximated by the difference signal between **P** and **S**.
- We enforce fine-grained matching between composed query and paired photo embedding via **region-aware triplet loss** and an **auxiliary generative loss**.



Proposed Model

- **Salient Components**
  1. Novel compositionality constraint to imitate the missing textual description.
  2. Neutral text to preserve the grammatical structure of CLIP's input text-space.
  3. Generalisable continuous prompt-learning over handcrafted textual prompts.
  4. Fine-grained matching via region aware triplet loss and auxiliary generative loss.
- While our method **does not** rely on paired textual captions during training, **users can provide optional captions** during inference.
- We compute the sketch-photo difference signal embedding Δ^w, which could be considered as a **pseudo word token imitating the difference between sketch and photo**, which ideally would be substituted with real query text during inference.
- Although Δ^w enforces compositionality, this mere numeric signal does not exist in CLIP's input text manifold and might break its grammatical syntax.
- To restrict the adverse effect of Δ^w, we regularise the training via a **"neutral-text" set** containing a list of 3-5 word generic description of a freehand sketch.
- We impose a **text-to-text generalisation loss** that enforces the learned prompts to be similar to a set of handcrafted language prompts.



- Finally to maintain fine-grained appearance features, we use a UNet generator to enforce generator guidance.

Experiments & Results