GeneCIS: A Benchmark For General Conditional Image Similarity

Sagar Vaze[1], Nicolas Carion[2], Ishan Misra[2]

[1] VGG, University of Oxford [2] Meta AI Research

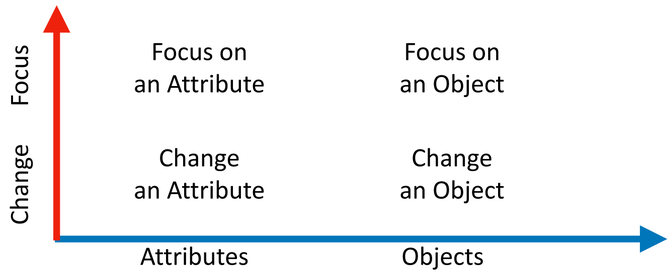
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

THE PROBLEM

- Which of the images (B, C, D) are **'most similar'** to (A)?
- Given different **conditions**, any of images are a valid answer
- Most image representations implicitly assume a single notion of similarity
- ☆ **Goal: We aim to train and evaluate models which can adapt to different similarity conditions**

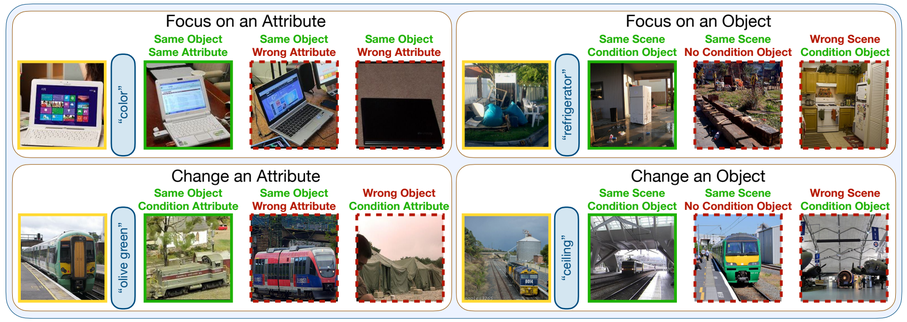
SIMILARITY CONDITIONS

- Existing conditional similarity benchmarks consider a finite set of conditions
- We consider an **open-set** of conditions with a **zero-shot benchmark**
- We consider conditions with respect to **two axes**



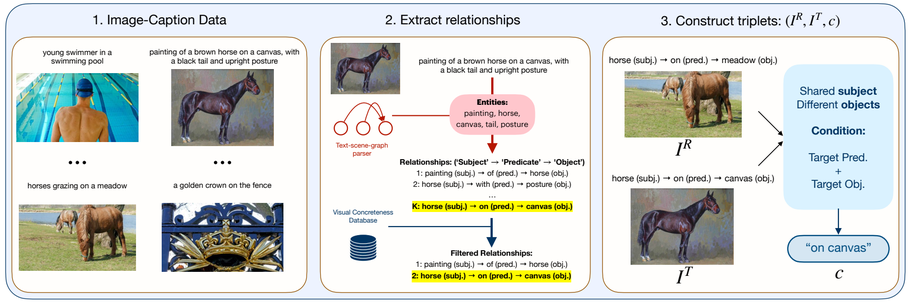
THE GeneCIS BENCHMARK

- GeneCIS contains four **conditional retrieval tasks** for zero-shot evaluation
- Model inputs: (i) Reference Image; (ii) Condition Text; (iii) Gallery of target images
- Models must select the only **conditionally similar** image in the gallery
- Gallery contains **'distractor'** images which **prevent shortcut solutions**



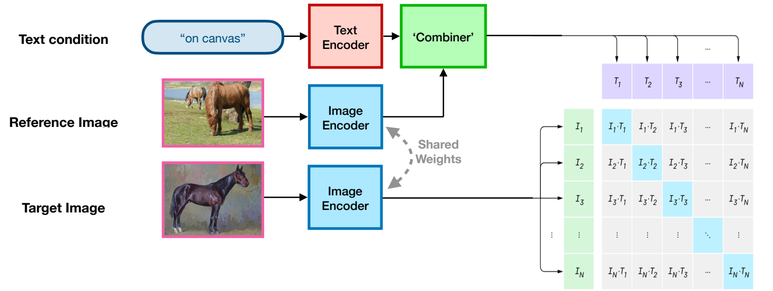
OUR METHOD: MINING TRAINING DATA

- Large-scale training data for conditional similarity is difficult to obtain
- We **mine training data** from large-scale **image-caption** datasets
- We mine triplets of (**reference image**, **target image**, **text condition**)
- We use text-scene-graph parsers to understand image contents



OUR METHOD: MODEL

- Embed text conditions and images with **CLIP encoders**
- Combiner module [1] conditions the reference image on the text condition
- Train contrastively end-to-end

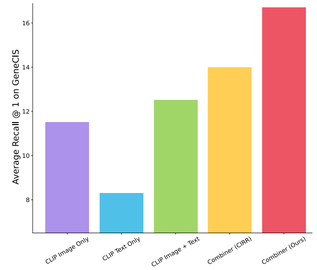


RESULTS ON GeneCIS

Our method outperforms **all baselines**

Including the same model trained with **manually annotated data from [2]**

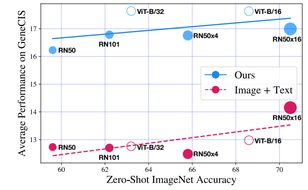
Our **zero-shot model** also outperforms **supervised baselines** on related datasets



WEAK CORRELATION WITH IMAGENET ACCURACY

GeneCIS performance is **weakly correlated** with the accuracy of the CLIP backbone

- This is different to many popular vision tasks
- Simply scaling existing methods is not fruitful



REFERENCES

[1] Conditioned and Composed Image Retrieval Combining and Partially Fine-Tuning CLIP-Based Features, Baldrati, CVPRW 2022
[2] Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models, Liu et. al, ICCV 2021

https://sgvaze.github.io/genecis/