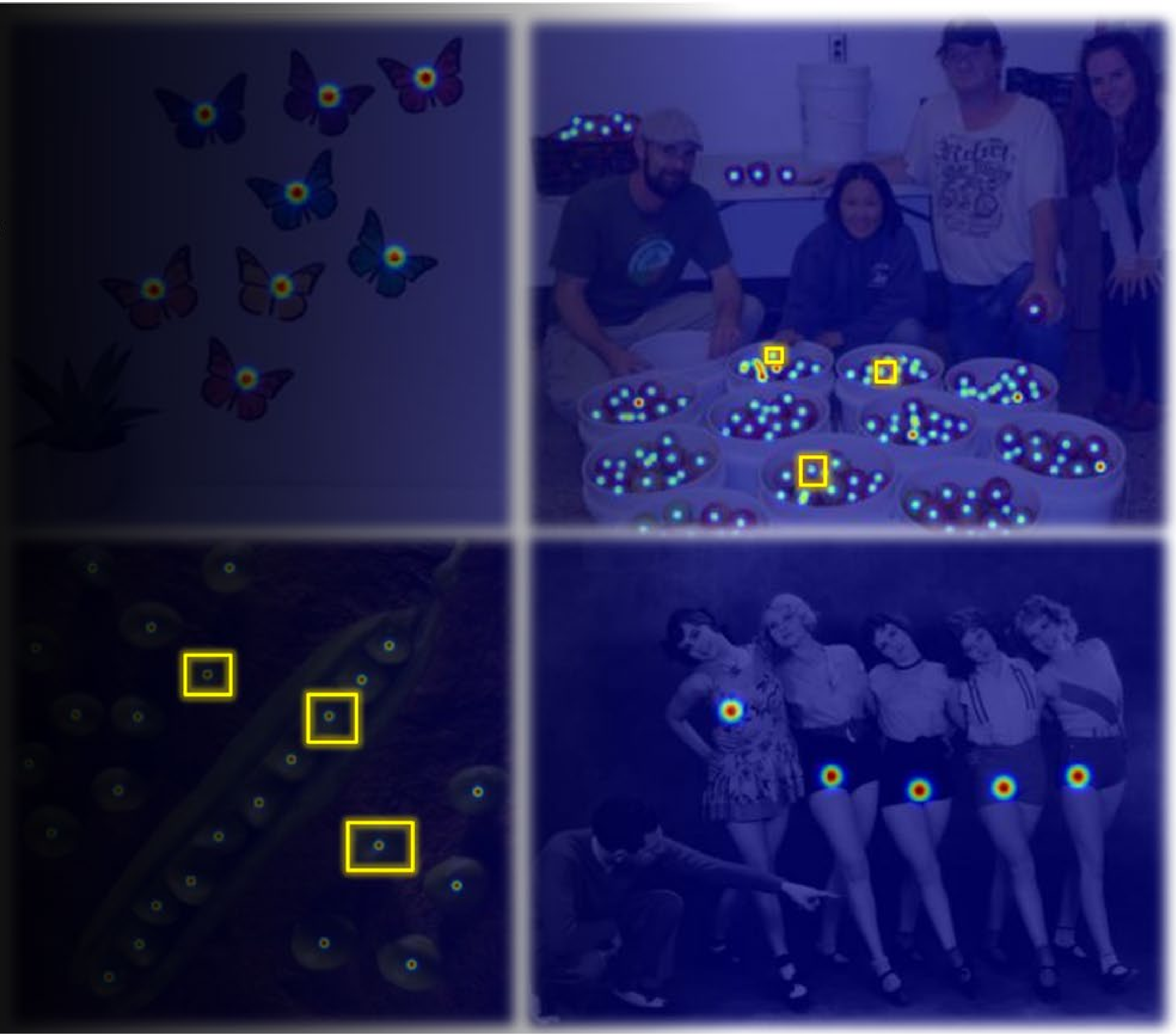
CountGD: Multi-Modal Open-World Counting

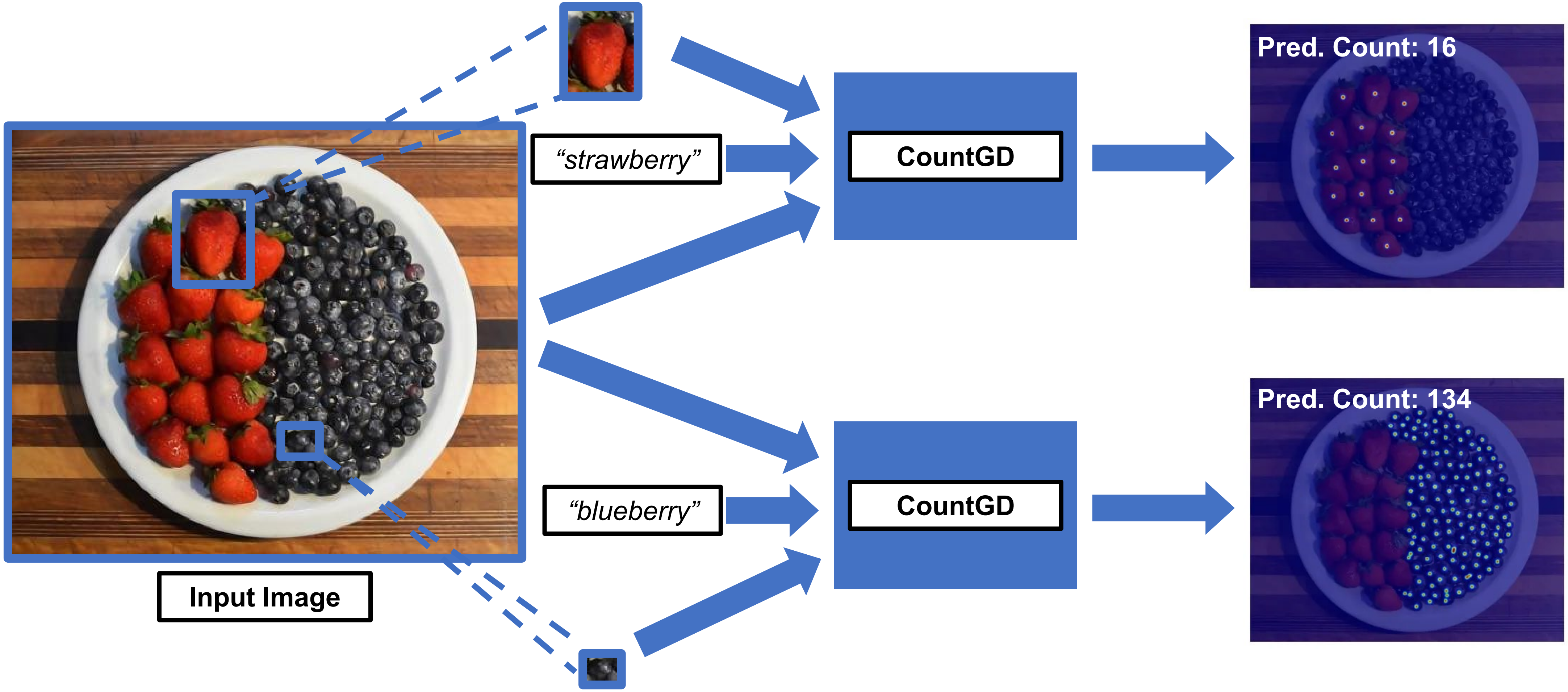
NeurIPS 2024

Niki Amini-Naieni, Tengda Han, & Andrew Zisserman

Project Page | ArXiv | Code | App



Main Idea



“strawberry”

“blueberry”

CountGD

CountGD

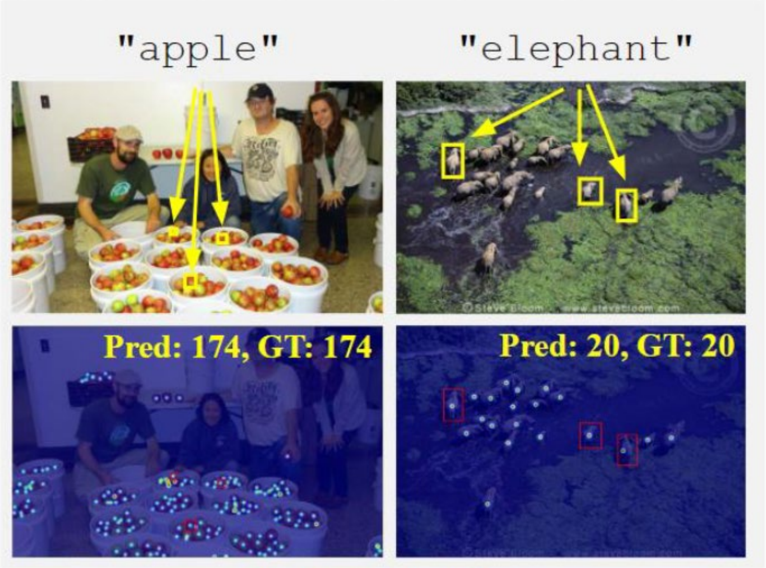
Input Image

Pred. Count: 16

Pred. Count: 134

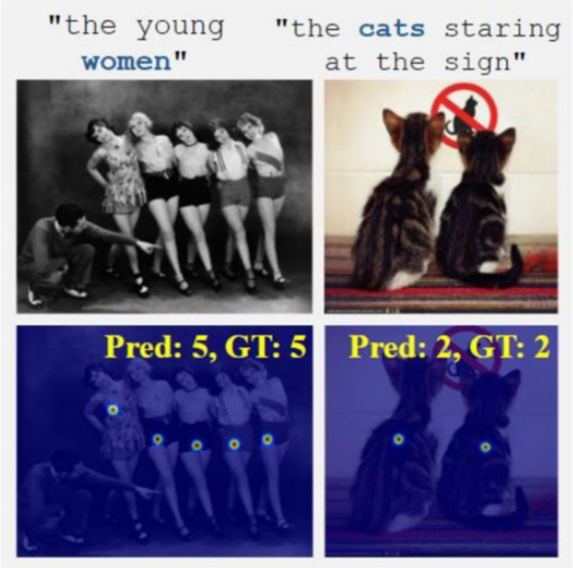
Contributions

1. We introduce the first open-world counting model, CountGD, where the prompt can be specified by a text description or visual exemplars or both.
2. We show the performance of CountGD significantly improves the state-of-the-art on multiple counting benchmarks.
3. We carry out a preliminary study into different interactions between the text and visual exemplar prompts, including the cases where they reinforce each other and where one restricts the other.



Visual Exemplars & Text

Input text is in quotes, and input visual exemplars are boxec

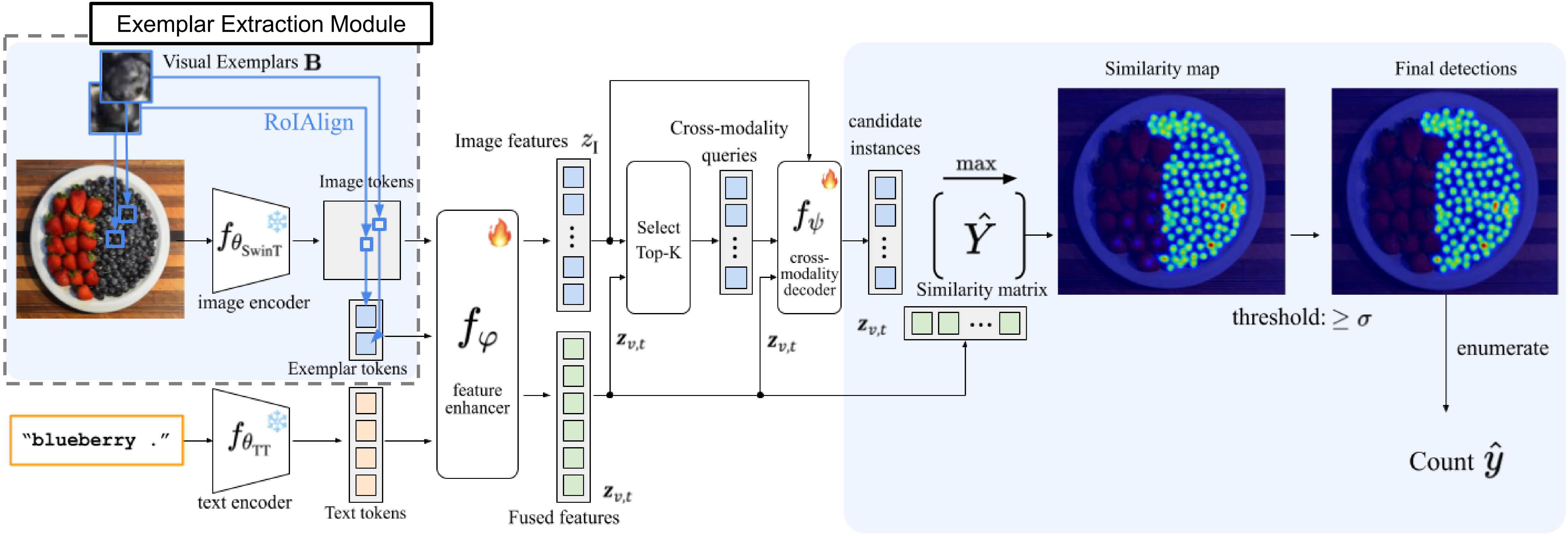


Text Only

tions are plotted and overlaid on top of eac

CountGD Architecture

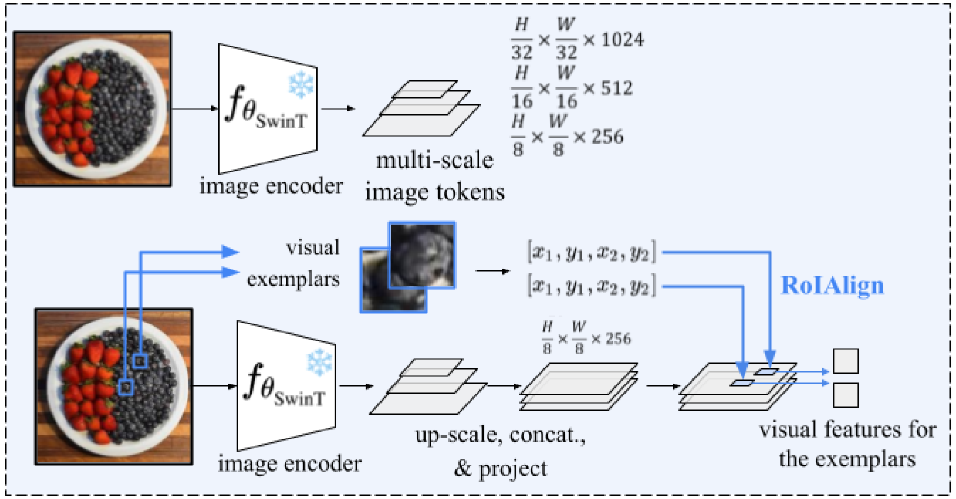
🟦 : Added to GroundingDINO



🟦 : Added to GroundingDINO

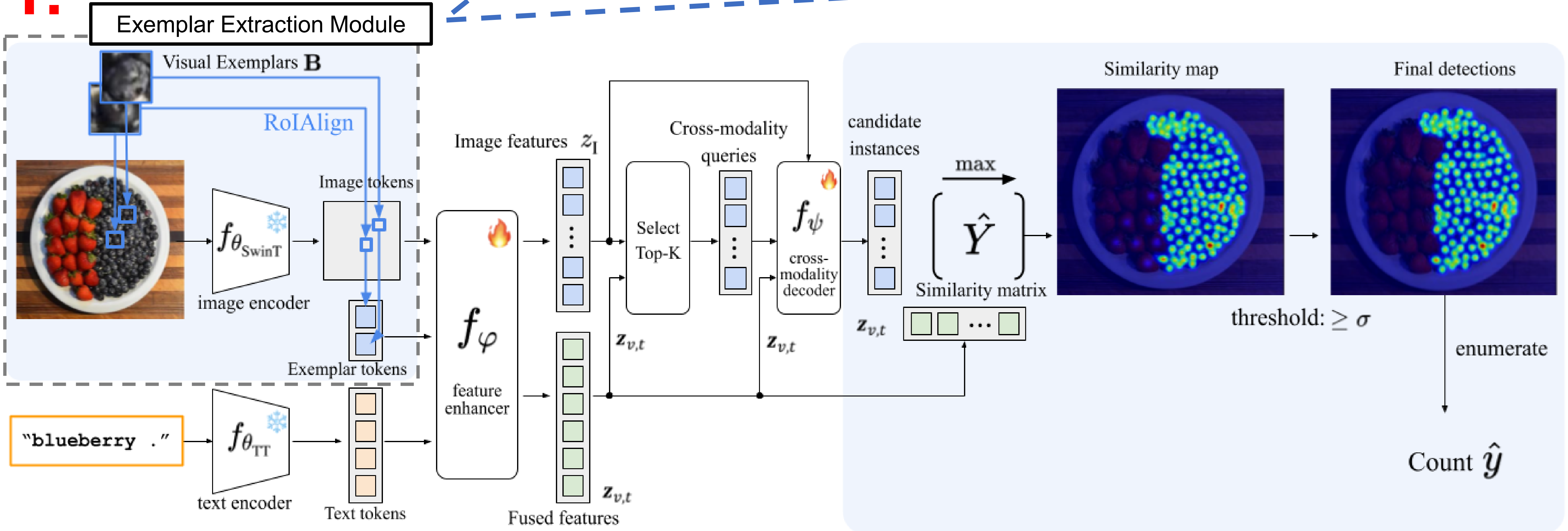
Module to enable inputting visual exemplars into GroundingDINO.

1.



$f_{\theta_{\mathrm{SwinT}}}$
image encoder
multi-scale image tokens
$\frac{H}{32} \times \frac{W}{32} \times 1024$
$\frac{H}{16} \times \frac{W}{16} \times 512$
$\frac{H}{8} \times \frac{W}{8} \times 256$
visual exemplars
$[x_1, y_1, x_2, y_2]$
$[x_1, y_1, x_2, y_2]$
RoIAlign
$f_{\theta_{\mathrm{SwinT}}}$
image encoder
$\frac{H}{8} \times \frac{W}{8} \times 256$
up-scale, concat., & project
visual features for the exemplars

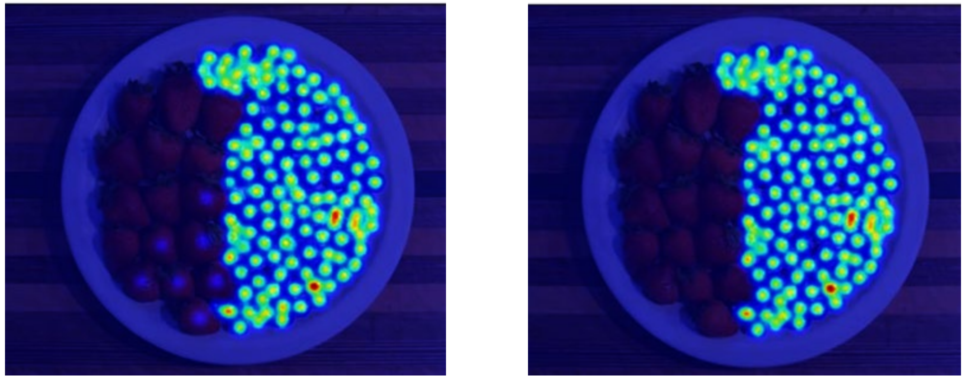
1.

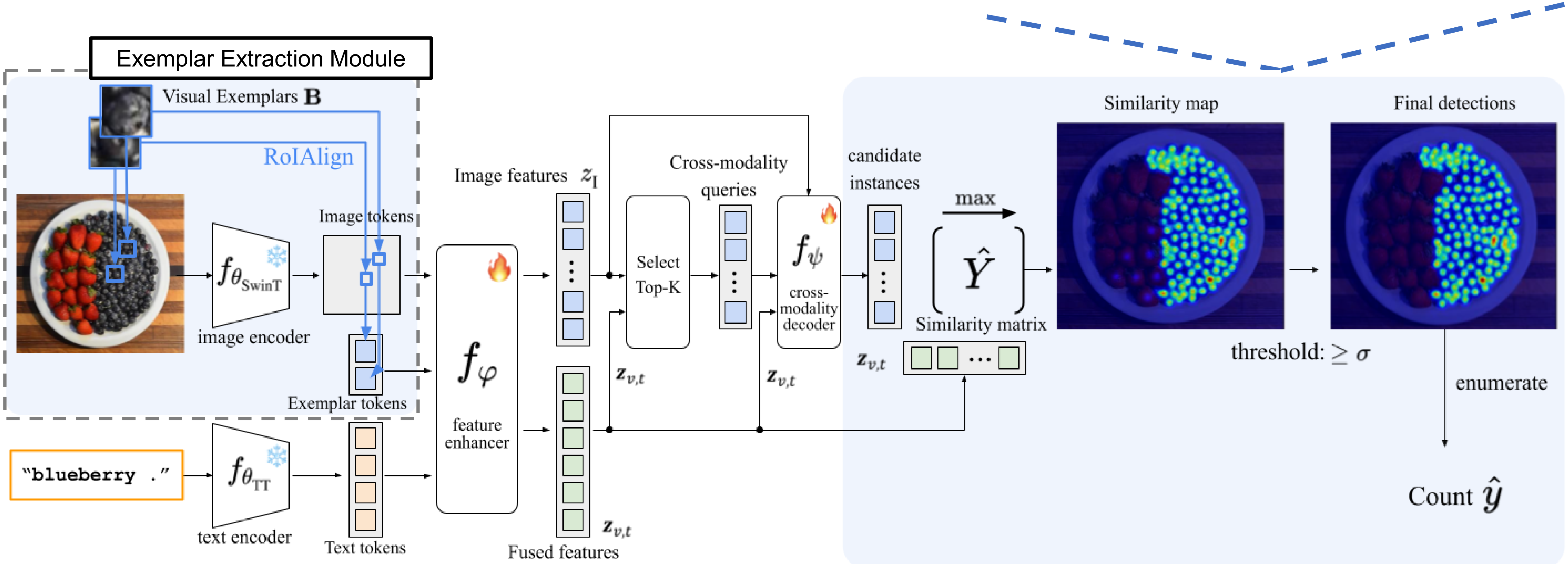


🟦 : Added to GroundingDINO

Getting the final count.

2.

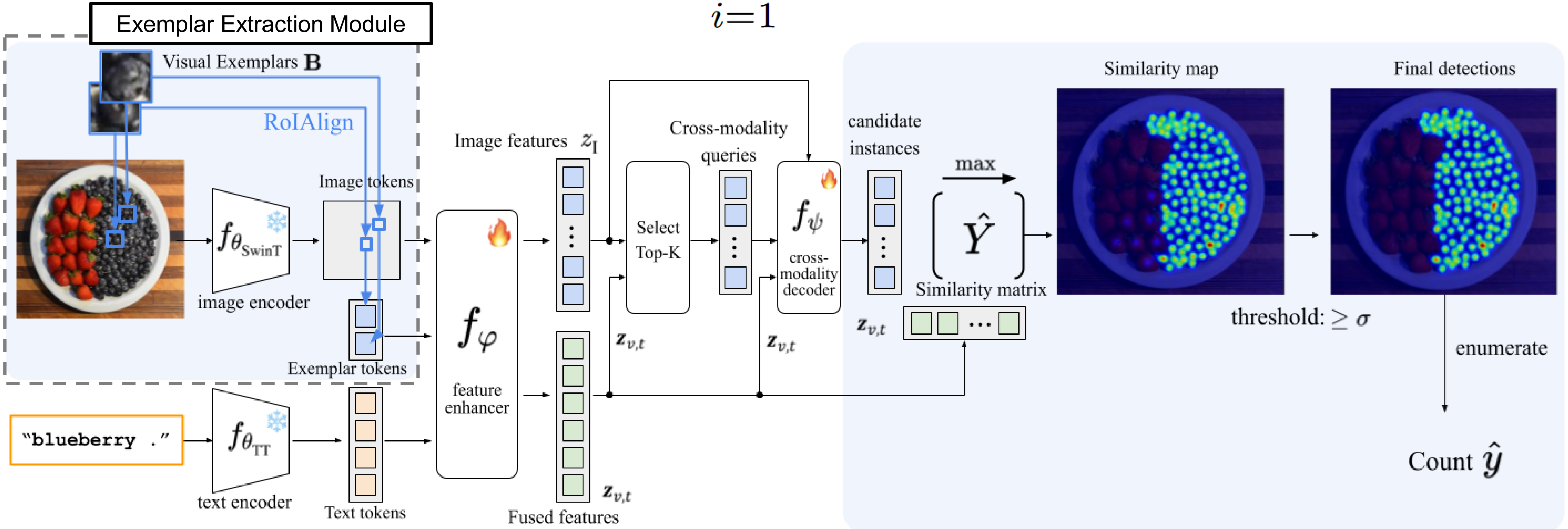




🟦 : Added to GroundingDINO

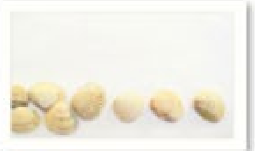
Loss (same as GroundingDINO's but with center points c_i instead of boxes)

$$\lambda_{loc}\sum_{i=1}^{l}|\hat{c}_i - c_i| + \lambda_{cls}\text{FocalLoss}(\hat{\mathbf{Y}}, T)$$



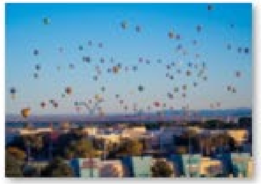
Training – Dataset

- Trained on open-world object counting dataset FSC-147 [1] with text and visual exemplars.
- Text encoder and image encoder frozen during finetuning.


2.jpg

3.jpg

4.jpg


5.jpg

6.jpg

7.jpg

• • •

7705.jpg

21.jpg

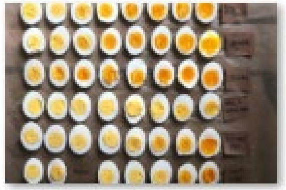
22.jpg

23.jpg

24.jpg

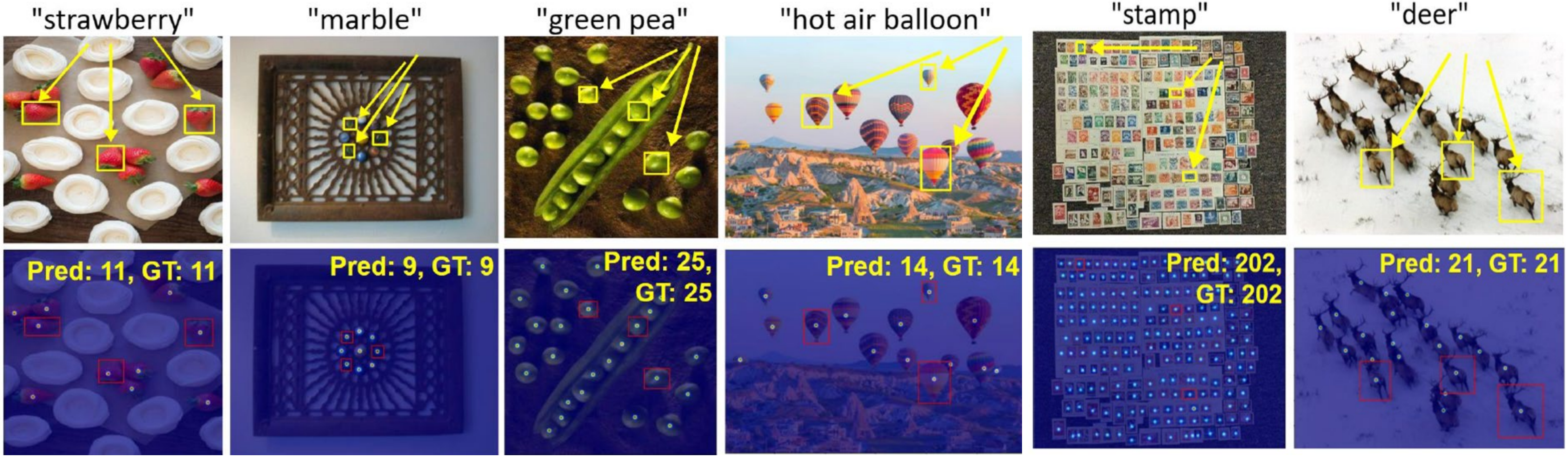
26.jpg

27.jpg


7714.jpg

[1] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proc. CVPR*, 2021.

Results – Qualitative

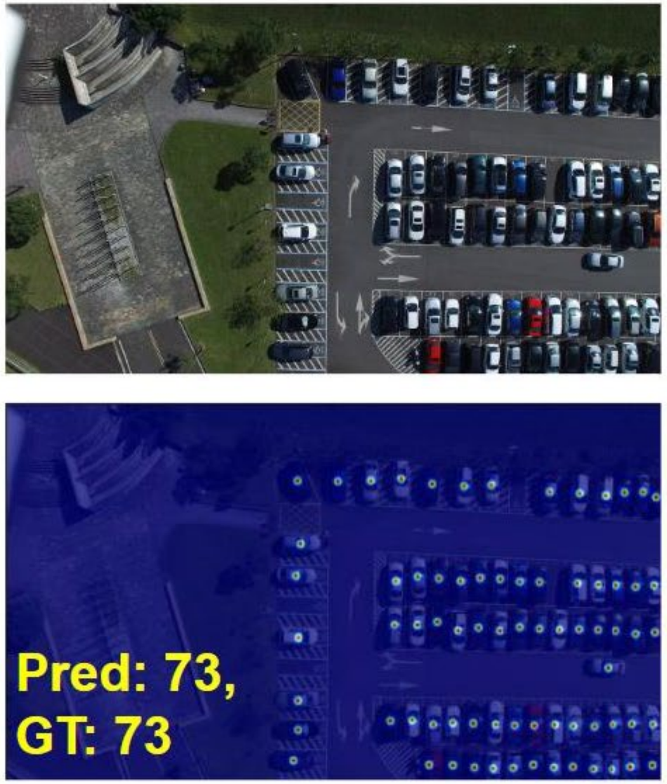


From FSC-147 test set

Dataset 1

Results – Qualitative Continued

"car"

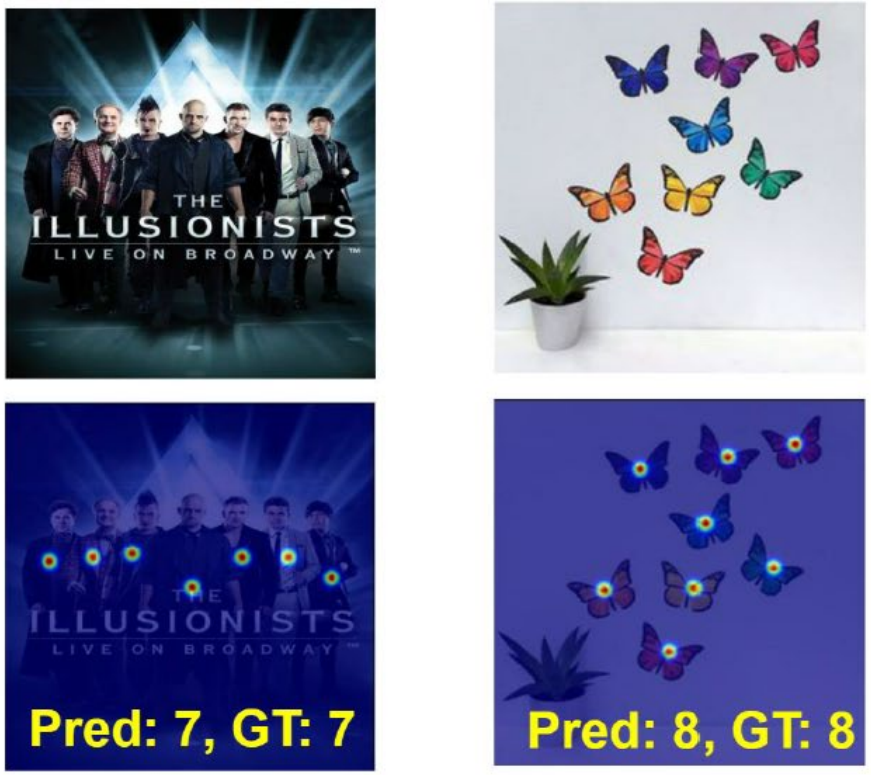


From CARPK

Dataset 2

"the world's greatest **magicians**"

"the beautiful butterfly **wall stickers**"

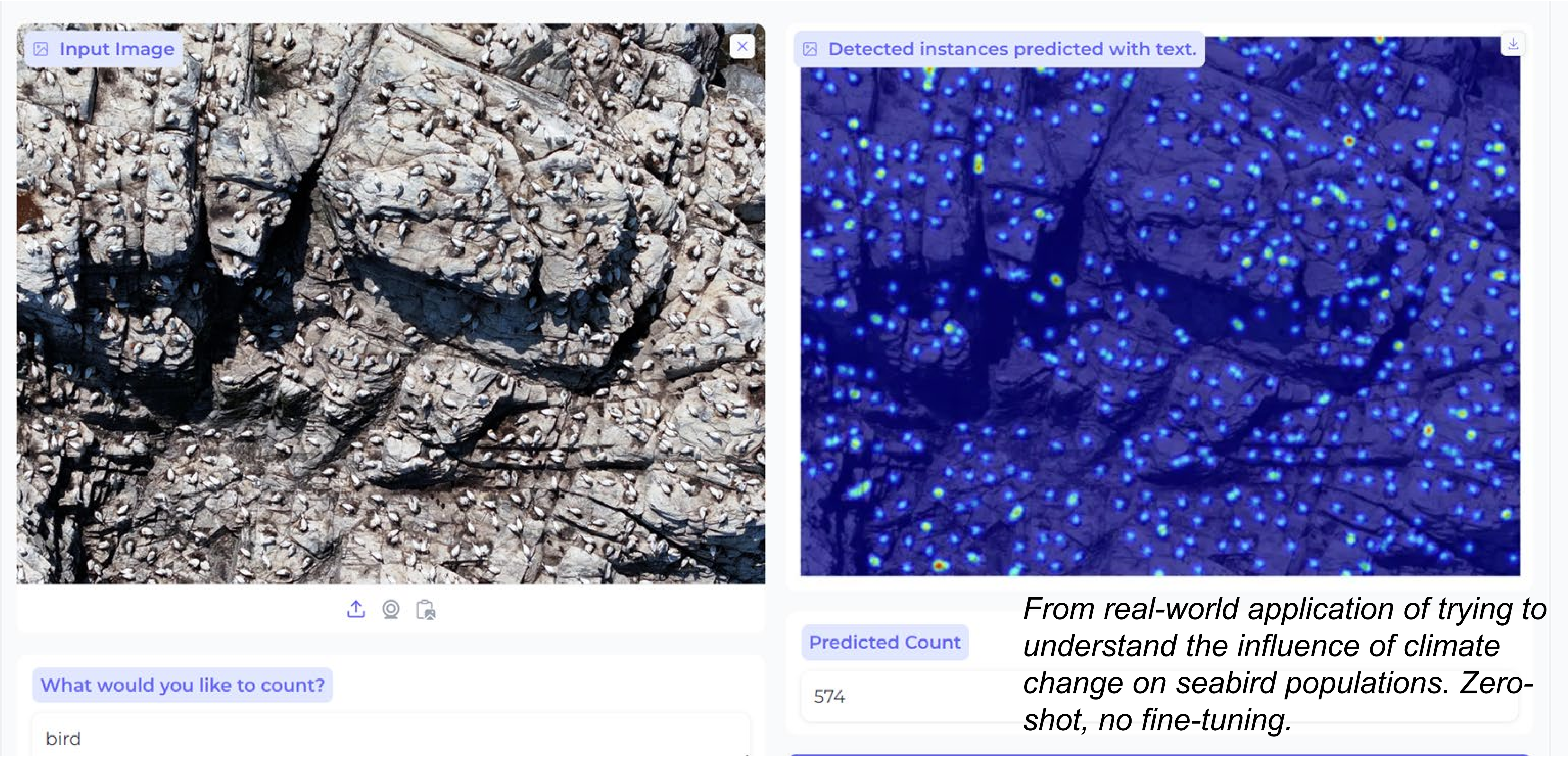


From CountBench

Dataset 3

*****Zero-shot results with no fine-tuning*****

Results – Qualitative Continued



Input Image

Detected instances predicted with text.

What would you like to count?

bird

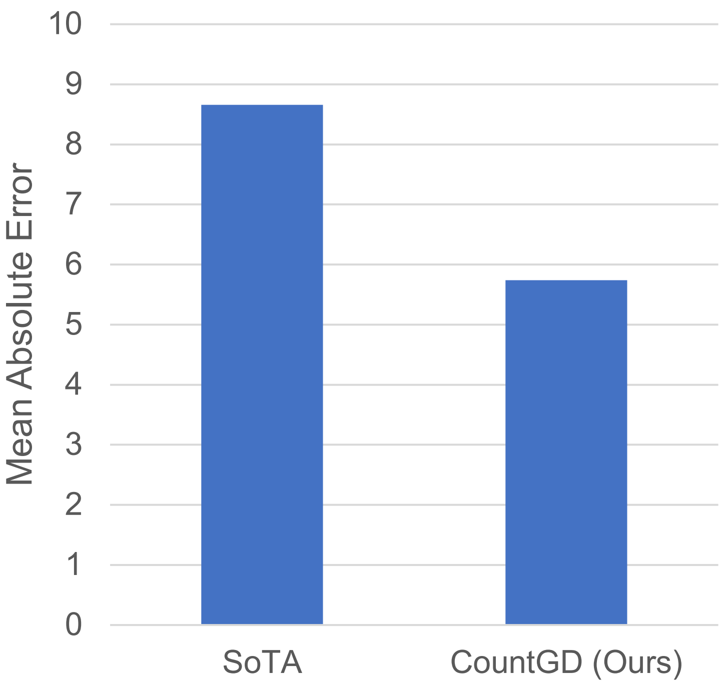
Predicted Count

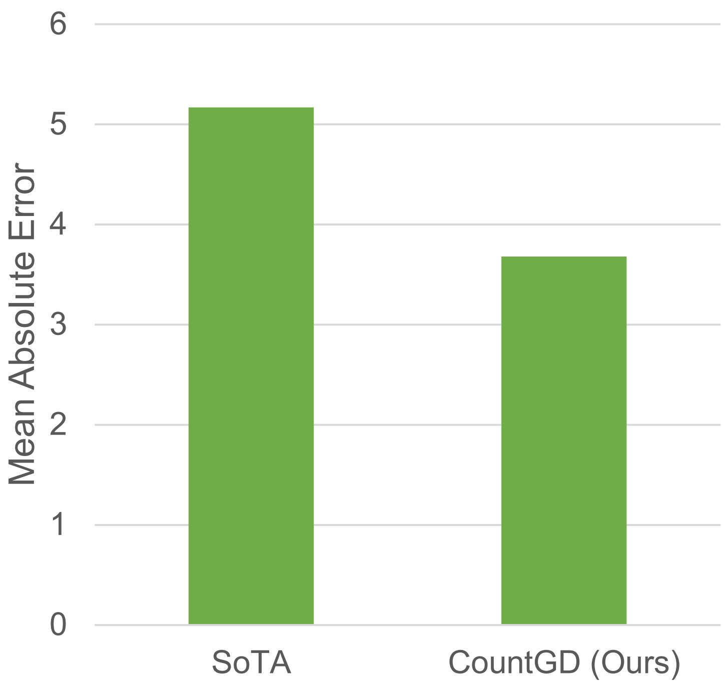
574

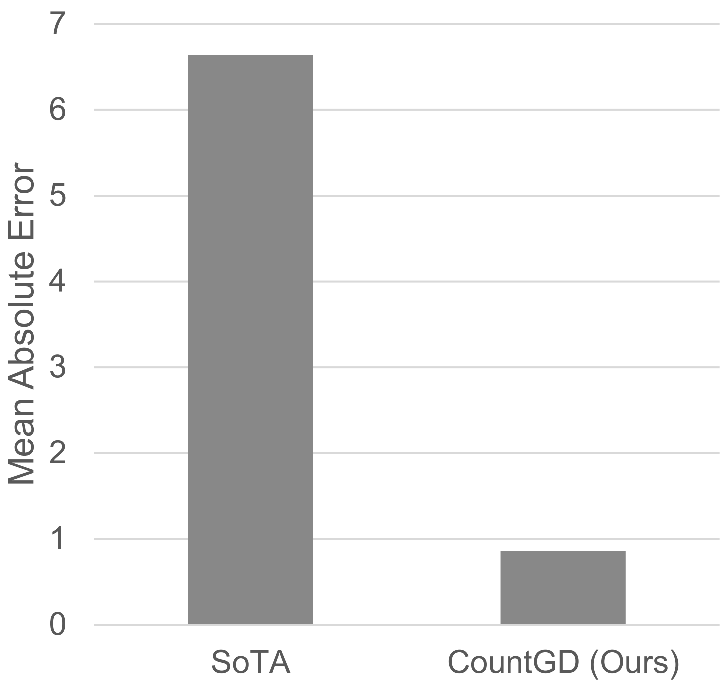
From real-world application of trying to understand the influence of climate change on seabird populations. Zero-shot, no fine-tuning.

Results – Quantitative

- CountGD achieves SOTA for open-world object counting. ***lower is better***.


FSC-147 (Dataset 1)

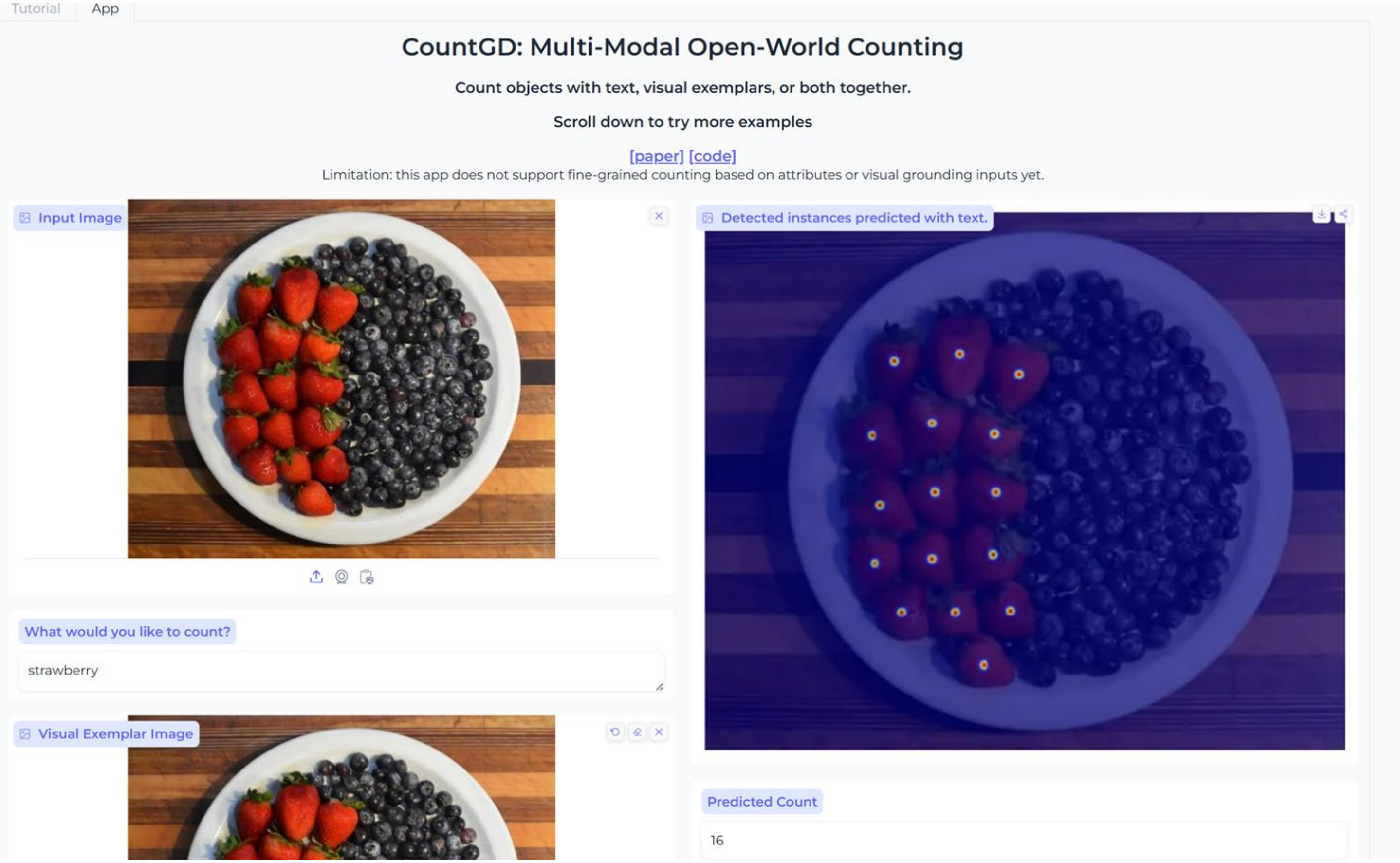

CARPK (Dataset 2)


CountBench (Dataset 3)

App Demo

[CountGD_Multi-Modal_Open-World_Counting - a Hugging Face Space by nikigoli]

[https://www.robots.ox.ac.uk/~vgg/research/countgd/]



CountGD: Multi-Modal Open-World Counting

Count objects with text, visual exemplars, or both together.

Scroll down to try more examples

[paper] [code]

Limitation: this app does not support fine-grained counting based on attributes or visual grounding inputs yet.

Input Image

What would you like to count?

strawberry

Visual Exemplar Image

Detected instances predicted with text.

Predicted Count

16

Thank you!

https://www.robots.ox.ac.uk/~vgg/research/countgd/