Referring Expression Counting

Siyang Dai[1], Jun Liu[1]*, Ngai-Man Cheung[1]*
[1]Singapore University of Technology and Design
siyang_dai@mymail.sutd.edu.sg, {jun_liu,ngaiman_cheung}@sutd.edu.sg

Abstract

*Existing counting tasks are limited to the class level, which don't account for fine-grained details within the class. In real applications, it often requires in-context or referring human input for counting target objects. Take urban analysis as an example, fine-grained information such as traffic flow in different directions, pedestrians and vehicles waiting or moving at different sides of the junction, is more beneficial. Current settings of both class-specific and class-agnostic counting treat objects of the same class indifferently, which pose limitations in real use cases. To this end, we propose a new task named Referring Expression Counting (**REC**) which aims to count objects with different attributes within the same class. To evaluate the REC task, we create a novel dataset named REC-8K which contains 8011 images and 17122 referring expressions. Experiments on REC-8K show that our proposed method achieves state-of-the-art performance compared with several text-based counting methods and an open-set object detection model. We also outperform prior models on the class agnostic counting (CAC) benchmark [36] for the zero-shot setting, and perform on par with the few-shot methods. Code and dataset is available at https://github.com/sydai/referring-expression-counting.*

1. Introduction

The objective of counting tasks is to predict the number of the target object in an image. The counting task has evolved from class-specific to class-agnostic through the years. The trend clearly indicates an expansion in scope from closed-set to open-set. The next step is to further enhance counting models to handle more fine-grained and in-context queries.

Industries and businesses require specified and customized quantitative analysis in order to create social and economic values. Specifically, such fine-grained analysis can empower transportation industry for traffic monitoring and crowd analysis, retail businesses for customer demographic analysis, warehouses for inventory count, farms for livestock management, government agencies for wildlife monitoring, etc. To this end, we propose a novel counting task named Referring Expression Counting (REC), which aims to count fine-grained and contextual objects within the same class e.g. ”female/male customer in the shop”, ”person in the first/second queue”, etc.

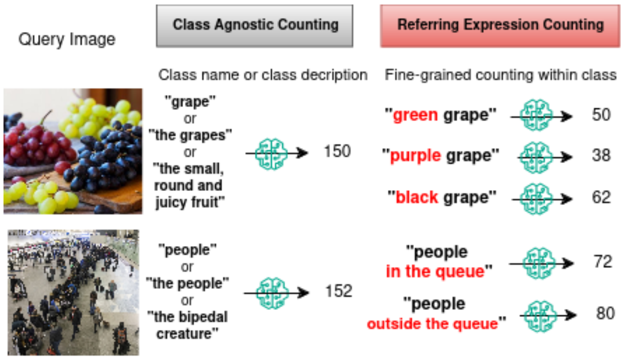


Figure 1. Comparison of class agnostic counting (CAC) task and the proposed Referring Expression Counting (REC) task. CAC takes class name [54] or class in natural language [2] or class description [2, 32] as input and count at class-level. REC aims to count objects in a fine-grained and in-context manner, essentially learning to differentiate attributes of the same-class objects.

In Fig. 1, we illustrate the difference between the proposed REC task and the existing class-agnostic counting (CAC) task. The fundamental difference is that CAC aims to count at class-level while REC aims to count in a fine-grained manner for objects of the same class but with different attributes. Two close works to ours are Teaching CLIP to Count to Ten [32] and CounTX [2]. The former proposes the CountBench, which is obtained by filtering a large-scale image-text dataset. The authors look for captions that contain a number between two and ten, which aligns with the count of a target object in the image, e.g. ”**two** zebras in Cape Town”. The objective is to enhance contrastive models with counting capability but not to count in a fine-grained manner. In CounTX, the authors consider that class names are not natural language and some class names in the original dataset [36] are incorrect or inaccu-

*Corresponding authors

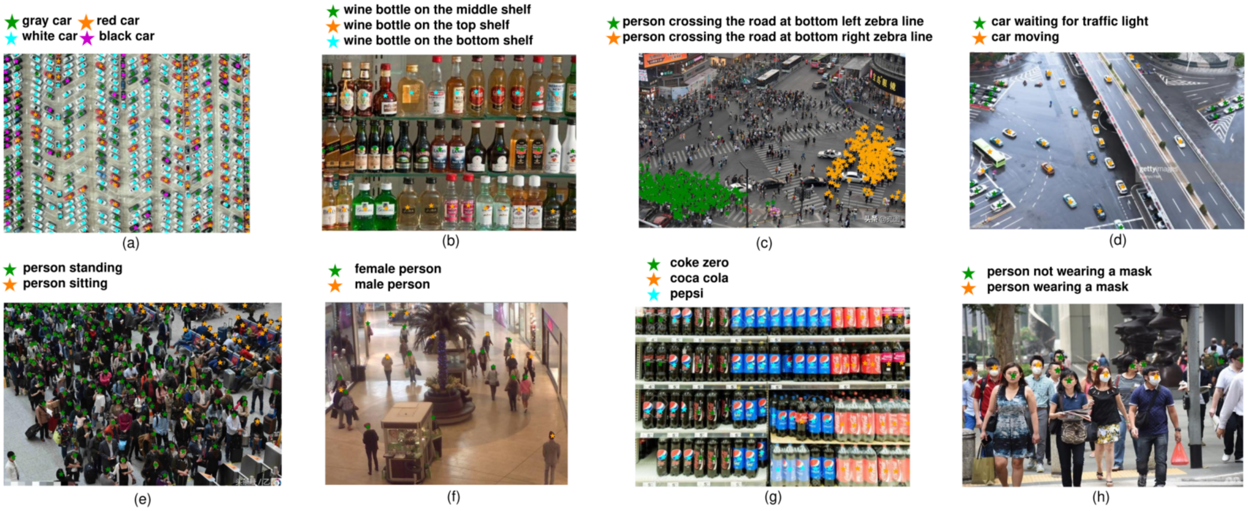

★ gray car ★ red car
★ white car ★ black car
(a)
★ wine bottle on the middle shelf
★ wine bottle on the top shelf
★ wine bottle on the bottom shelf
(b)
★ person crossing the road at bottom left zebra line
★ person crossing the road at bottom right zebra line
(c)
★ car waiting for traffic light
★ car moving
(d)
★ person standing
★ person sitting
(e)
★ female person
★ male person
(f)
★ coke zero
★ coca cola
★ pepsi
(g)
★ person not wearing a mask
★ person wearing a mask
(h)

Figure 2. Samples from REC-8K to illustrate different attributes within the class (better viewed in enlarged version).

rate. Therefore they rewrite the class names by adding ”the” to the beginning of most of the classes and for some inaccurate classes, they provide more descriptions e.g. ”candy pieces” becomes ”the round desserts decorated with colorful candy pieces”. They still aim to tackle the CAC task but not consider different attributes within the class.

In summary, our contributions are as follows: 1. We introduce a new task: Referring Expression Counting (REC), which takes counting to the next level by handling more fine-grained queries. 2. We create a novel benchmark dataset REC-8K which contains 8011 images with 17122 referring expressions covering a wide variety of attribute types. 3. We achieve state-of-the-art performance on the REC-8K benchmark, by leveraging prior knowledge of a vision-language model to align image and text features, and our innovative modules of global-local feature fusion and contrastive learning. We also make it to the top for the zero-shot setting on the CAC benchmark [36], and even match the performance of the top few-shot model.

2. Related work

2.1. Existing counting tasks

Prior to this work, there are two counting tasks: category-specific counting [9, 13–15, 18, 20, 22–24, 29, 31, 40, 43, 44, 53, 57]; and class-agnostic counting by few-shot setting [17, 26, 30, 38, 42, 46, 58, 60] and by zero-shot setting [2, 8, 11, 35, 54, 62]. Category-specific counting models are trained and tested on one category of objects such as people, vehicles etc., but don’t perform well on unseen categories. Class-agnostic counting aims to train a generalist model to count any seen or unseen category of objects. The authors in [36] propose few-shot counting benchmark FSC-147 for class agnostic counting. Their model FamNet predicts the density map by matching the input image with a set of exemplars and adapts novel categories at test time. In [54], zero-shot counting is proposed, which takes the class name as input and trains a conditional VAE to generate semantic embedding prototype for selection of visual exemplars from random image patches. In this paper, we present a new task Referring Expression Counting (REC), which focuses on fine-grained object counting within the class.

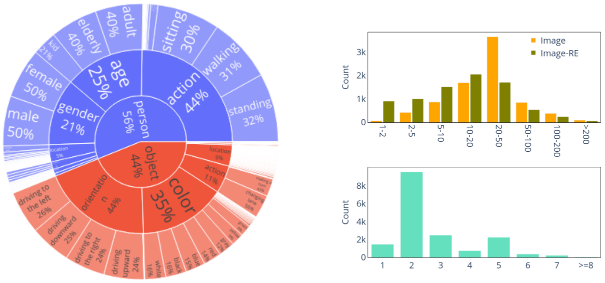
2.2. Language-based counting methods

With recent advances in cross-modality learning [7, 55], text has been involved to take counting tasks to a new level. The methods based on text input are essentially zero-shot and are closely related to our work. In [32], the authors teach CLIP [34] to count up to ten objects by fine-tuning with a contrastive loss between the true and counterfactual prompts besides the original text-image contrastive objective. ZSC [54], CounTX [2] and CLIP-Count [11] embed class names with CLIP’s text encoder for further interaction with the query image features to regress a density map for zero-shot class-agnostic counting. TFOC [62] attempts to use different prompts i.e. point, box and text to query SAM [12], the segmentation model, to count objects in a training-free manner. To tell apart our work from existing works, we discuss the differences between us and [2, 32] in Sec. 1 and illustrate the same in Fig. 1. Again in this work, our input is not the class name or class description but referring expression that enables more fine-grained counting.

2.3. Related referring expression tasks

Referring object detection [16, 19, 21, 33, 41, 45, 47, 52, 59, 61] and referring expression segmentation [27, 28, 51, 56, 63, 64] are closely related to REC. They are aimed for detection or segmentation of target objects based on a referring expression, and requires fine-grained understanding of both text and image. GroundingDino [21] unifies the evaluation of closed-set, open-set and referring expression object detection within a single framework. VLDet [16] directly trains an object detector with image-text pairs by formulating the region-word alignments as a set-matching problem. GRES [19] models the relationship between image regions and regions-words explicitly by cross attention and learn segmentation masks for the referred objects. Authors in [61] extract global-local context features for both text and image by decomposing the text into words and the image into regions, and then compute cosine similarity scores and choose the mask with the highest score. We leverage strong text-image fusion capabilities of GroundingDino to perform reasoning of the referring expression and locate the target objects.

3. Task setting and dataset



(a) Percentages of person & object categories, attribute types and attributes.

(b) Histogram of points per image, per Image-RE pair (up) and words per RE (down).

Figure 3. Dataset REC-8K statistics.

3.1. Referring expression counting task

Counting tasks aim to output the number of a target object in an image. Existing settings either count category-specific objects or class-agnostic objects without considering different attributes within the class. Category-specific counting takes images as the sole input; and class agnostic counting also takes as input the exemplars for the few-shot setting or class names for the zero-shot setting. We believe these settings are not flexible for real-world applications. Counting objects with attributes is more beneficial.

We introduce Referring Expression Counting which takes referring expressions and a query image as inputs and outputs the target object count as well as the location. A referring expression is composed of the **class** and **attribute** of a target object. Unlike class specific datasets [9, 39, 48] and class agnostic dataset [36] which only consider one category of objects per image, REC accepts various referring expressions for objects with different attributes.

3.2. The REC-8K dataset

Counting tasks usually involve dense objects. Referring expression counting additionally demands objects with various attributes. To this end, we create REC-8K, a novel benchmark for evaluation of the REC task.

Data collection and annotation. We collect images suitable for the REC task from various sources. We select from existing datasets: FSC-147 [36], JHU Crowd [39], NWPU [48], VisDrone [65], DETRAC [50], Carpk [9], Mall [5], Crowd Surveillance [57], and also source from photo sharing websites Pixabay and Unsplash to construct REC-8K. Considering application scenarios, we look for images in the domains of crowd analysis, traffic surveillance, retail and warehousing, etc.

We first create novel RE labels based on visible object attributes and annotate the target objects accordingly. Similar to existing counting datasets, we use point annotations for the targets. For object and person category, we draw a dot at the object center and at the center of the human head respectively. If the target object is occluded, we mark the center of the visible area. We use Amazon Mechanical Turk [1] to annotate the dataset. Some sample annotated images are presented in Fig. 2.

Statistics. REC-8K contains 8011 images with a total of 286621 point annotations. The min, average and max number of objects per image is 1, 36 and 1028. Due to the nature of REC setting, we treat an Image-RE pair as a training sample. We have a total of 17122 Image-RE pairs, meaning in average there are 2.13 referring expressions per image. The min, average and max number of target objects per Image-RE pair is 1, 17 and 1004. We show statistics for attribute types separately for **object** and **person** categories due to the distinctive natures. For object, the attribute types include color, location, material, action, variety, size and orientation. For person, we have attribute types clothing, action, accessory, location, age, orientation and gender. We illustrate the distribution of attributes in Fig. 3a and histograms of point annotations and word counts in Fig. 3b. We also provide comparisons with existing counting datasets in Tab. 1.

Dataset splits. Since we treat an Image-RE pair as a sample, complying with our batching strategy in training, we split the dataset into train, val, test sets which contains 10555, 3336, 3231 Image-RE pairs and 4923, 1566, 1522 images respectively. We try to minimize the shared REs between different splits and still maintain a reasonable ratio between train-val and train-test splits. Among the 723, 341, 299 unique REs in train, val, test sets, there are 80 REs shared between train and val sets and 79 REs shared between train and test sets. Despite the shared REs, the images in different splits are unique.

Dataset	Text Anno.	Classes	REs	No. of images	Avg. points
JHU-Crowd [39]	×	1	-	4372	345
NWPU [48]	×	1	-	5109	417
UCF CC 50 [10]	×	1	-	50	1279
CARPK [9]	×	1	-	1448	62
FSC-147 [36]	×	147	-	6135	56
REC-8K (ours)	✓	-	**1182**	**8011**	**36**

Table 1. Comparison of existing counting datasets and REC-8K.

4. Method for REC task

We show the overview of our proposed method in Fig. 4. We first explore the use of open-set object detector as counting model (Sec. 4.1), and select a strong base model for REC. With the prior knowledge from the base model, we create modules that further enhance REC: global-local feature fusion (Sec. 4.2) and contrastive feature learning (Sec. 4.3). We present the final loss function (Sec. 4.4) in the end of this section.

4.1. Object detector as counting model

The mainstream counting models have been based on regression of density maps instead of object detection. However, transformer-based detectors have shown promising results in detection of occluded and small objects [37] thanks to global contextual understanding and positional encoding. Going forward to the open-set detection, state-of-the-art model GroundingDino [21] enables text to image integration at multiple levels by using transformer architecture for both. It also leverages strong capabilities of pretraining with large-scale datasets of transformer-based detector, and end-to-end learning for better generalization. Based on these advantages, we choose GroundingDino as the base model for REC thus name our model GroundingREC.

We adapt GroundingDino to REC by making the following modifications: 1. Instead of bounding box, we take the box center as output to align with the point annotation. In the calculations of both Hungarian matcher cost and final loss, we compute L1 loss for point regressions instead of original GIOU and L1 loss for bounding box. 2. The original GroundingDino detects objects for all nouns in the referring expression (RE), e.g. ”cat on the table” will output bounding boxes for both ”cat” and ”table”. However, in REC, ”on the table” is an attribute for ”cat” and we only need to count cats on the table not the table itself or cats elsewhere in the image. So we employ CLS token as the global semantics for the RE instead of individual text tokens.

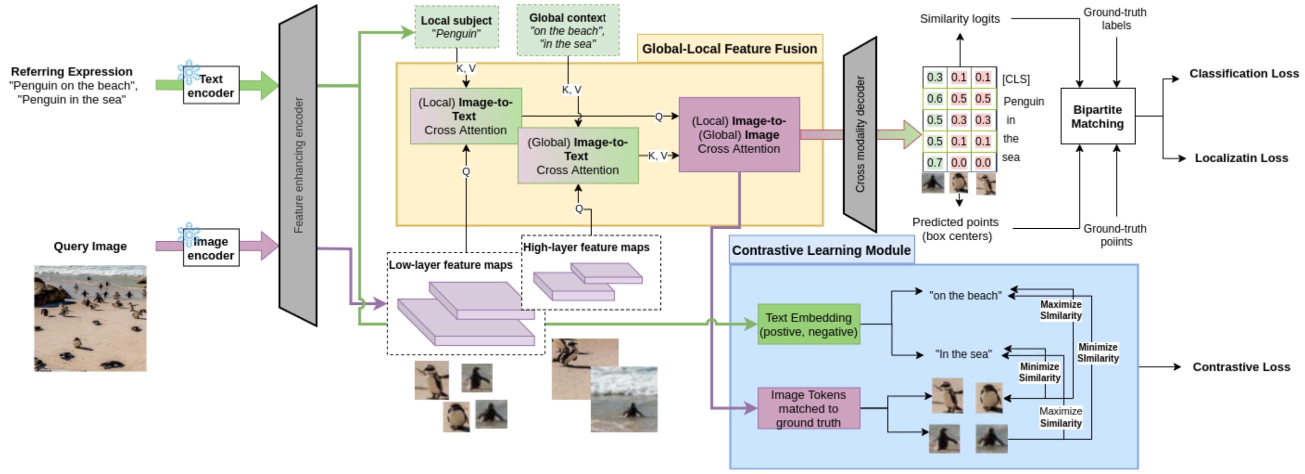
4.2. Global and local feature fusion

We categorize attributes in the RE labels in two high-level categories: local attributes and relational attributes. Local attributes are attributes that can be inferred from a local crop of the specific object in the image, such as color, material, age and gender, etc. Relational attributes are attributes that require understanding of the context, such as location and relative size. We observe that the base model performs better on local attributes than on relational attributes. The hypothesis is that the base model is pre-trained to attend to individual objects in the text input, but we need a global understanding of the image to infer relational attributes. To tackle this, we propose a global-local feature fusion module to enhance the global understanding of the image. The key idea is for the candidate image tokens potentially yielding true detections to be aware of the global context.

The encoder extracts multi-scale image features from different blocks of the image backbone. The encoded image tokens from different layers of feature maps correspond to different sizes of receptive field in the image, with lower layers representing smaller receptive fields and higher layers bigger ones [25]. For REC task, we are looking for target objects that are generally smaller in size. Built on this, we propose global-local feature fusion in two steps as illustrated in the yellow box of Fig. 4 as well as in Algorithm 1. First, we perform cross attention between text and image features, and then between image features at different layers. Specifically, we first split the input referring expression into two parts: the *subject* part which is the object with any local attribute and the *context* part which is the relational attribute. For example, for RE ”red apple in the left bowl”, the *subject* is ”red apple” and the *context* is ”in the left bowl”. We also split the image features into two parts: *lower-layer* image tokens and *higher-layer* image tokens. In order to enhance image features, we use *lower-layer* image features as query to attend to the *subject* part of the text feature, and *higher-layer* image features as query to attend to the *context* part of the text feature. After enhancement in step one, we use the *lower-layer* image features to query the *higher-layer* image features, essentially asking ”what is the global context of this local image feature?”. The global-context informed image tokens are then processed by a cross-modality decoder of GroundingDino to generate the final image features.

Algorithm 1 Global-local feature fusion

```
**Input**: image features $I$, text features $T$, subject mask $M_{subj}$, context mask $M_{ctx}$
**Output**: global-context informed image features $I'$
Split $I$ into lower-layer tokens $I_{low}$ and higher-layer tokens $I_{high}$
# Step 1: cross attention between text and image features
$I_{low} \leftarrow \text{CrossAttention}(I_{low}, T, M_{subj})$
$I_{high} \leftarrow \text{CrossAttention}(I_{high}, T, M_{ctx})$
# Step 2: cross attention between image features
$I_{low} \leftarrow \text{CrossAttention}(I_{low}, I_{high})$
$I' \leftarrow Update(I, I_{low})$
**return** $I'$
```



0.3	0.1	0.1	[CLS]
0.6	0.5	0.5	Penguin
0.5	0.3	0.3	in
0.5	0.1	0.1	the
0.7	0.0	0.0	sea

Figure 4. Overview of the proposed method. The model takes a referring expression and a query image as inputs. After encoder, the image and text embeddings first go through the global-local feature fusion module to yield global-context enhanced image embeddings. Then the contrastive learning module further encourages discriminative feature learning by maximizing the similarity between the image and positive text embedding and minimizing the similarity with negative text embeddings. The final loss function is a weighted sum of the localization, classification and the contrastive loss.

4.3. Contrastive learning

In REC, we want to differentiate the target objects by a given attribute from the other attributes of the same class. For example, in referring expression ”red apple in the left bowl”, we want to differentiate red apples from green apples or apples in other locations. To this end, we propose the contrastive learning module to learn more discriminative features associated with attributes.

In particular, we take multiple RE inputs for different attributes of the same-class object. After feature enhancement encoder of our base model, we obtain the text embeddings of the attributes and image tokens embeddings. By Bipartite Matching, we match the predicted points with the ground truth points, and then identify the image tokens associated with the matched points. For each matched image token k, we take the corresponding RE's attribute as the positive text sample and the other attributes of the same class as negative text samples. By taking mean of the attribute tokens embeddings, we obtain the positive text embedding t_i^+ and the negative ones t_j^-. We then compute s_{ik}, the similarity score between the k-th matched image embedding and the positive text embedding t_i^+, and s_{jk}, the similarity score with the negative text embeddings t_j^-. The contrastive loss is indicated in Eq. (1).

$$\mathcal{L}_{contrast} = -\frac{1}{K}\sum_{k=1}^{K}[\log(s_{ik}) + \sum_{j\neq i}\log(1 - s_{jk})] \tag{1}$$

As illustrated in the blue box of Fig. 4 for the contrastive loss function, we learn more discriminative features for attributes by pushing the positive text embedding t_i^+ closer to the matched image embedding k and pushing the negative text embeddings t_j^- away from it.

4.4. Loss function

With the adaption of DETR-like object detector to REC, we first conduct Bipartite Matching to match the predicted points with the ground truth points. Then we compute the following losses between the matched points.

Localization loss for point regression which is the L1 distance between predicted point $\hat{p}_k$ and ground truth point p_k. In Eq. (2), K is the total number of matched points.

$$\mathcal{L}_{loc} = \frac{1}{K}\sum_{k=1}^{K}\|\hat{p}_k - p_k\| \tag{2}$$

Cross-entropy classification loss for point classification, where y_i is the ground truth label for i-th text token and $\hat{y}_i$ is the class logit between i-th text token and k-th predicted image token. The loss is calculated by taking mean of all scores for N text tokens then averaged over K matched points.

$$\mathcal{L}_{cls} = \frac{1}{K}\sum_{k=1}^{K}\left[-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log\hat{y}_i^k + (1 - y_i)\log(1 - \hat{y}_i^k)\right]\right] \tag{3}$$

Contrastive loss (see Eq. (1)) for image-text alignment as described in Sec. 4.3. The final loss is a weighted sum of all three losses.

$$\mathcal{L} = \mathcal{L}_{loc} + \lambda_1\mathcal{L}_{cls} + \lambda_2\mathcal{L}_{contrast} \tag{4}$$

5. Experiments

5.1. Implementation details

We use GroundingDino [21] as the backbone of our model, which takes as input a referring expression and a query im-

Method	Backbone	Finetuning	Val set					Test set				
			MAE↓	RMSE↓	Prec↑	Rec↑	F1↑	MAE↓	RMSE↓	Prec↑	Rec↑	F1↑
Mean	-	-	14.28	27.75	-	-	-	13.75	25.91	-	-	-
ZSC [54]	ResNet-50	✓	14.84	31.30	-	-	-	14.93	29.72	-	-	-
ZSC [54]	Swin-T	✓	12.96	26.74	-	-	-	13.00	29.07	-	-	-
TFOC [62]	ViT-B	-	16.08	31.61	0.30	0.07	0.12	17.27	32.68	0.23	0.07	0.11
CounTX [2]	ViT-B-16	✓	11.88	27.04	-	-	-	11.84	25.62	-	-	-
GroundingDino [21]	Swin-T	×	11.77	28.6	0.57	0.25	0.34	11.71	26.97	0.59	0.25	0.35
GroundingDino [21]	Swin-T	✓	9.03	21.98	0.56	0.76	0.65	8.88	21.95	0.59	0.76	0.66
GroundingREC (ours)	Swin-T	✓	**6.80**	**18.13**	**0.65**	**0.71**	**0.68**	**6.50**	**19.79**	**0.67**	**0.72**	**0.69**

Table 2. Comparison of Results on Referring Expression Counting benchmark REC-8K.

age and outputs the total count by summing up all positive point predictions. In particular, we use the text encoder and image encoder of GroundingDino to extract the text and image features, and then perform the proposed global-local feature fusion after the feature enhancer. Specifically, we split the image tokens after language-guided query selection of GroundingDino and empirically choose top 10% from the tokens sorted from lowest to highest feature maps as higher-layer tokens and the rest as lower-layer tokens. For model output, we follow the DETR [3] paradigm to output a fixed number of 900 predictions from which we select the positive detections by thresholding the class logits with the following method. We first compute the class logits by dot product each image token with each text token (see score matrix in Fig. 4). Then for each image token, we compare its score with the CLS token by a threshold of 0.25 and scores with rest of the text tokens by a threshold of 0.35. If all scores are higher than the thresholds, we include the corresponding prediction in the final count.

For training, we keep frozen the BERT [6] text encoder and Swin-T [25] image encoder, and only train the proposed global-local feature fusion modules and the cross-modality encoder and decoder of GroundingDino. We use the AdamW optimizer with a learning rate of 1e-5. For Hungarian matching, we set the ratio of localization loss and classification loss to 1:5. For final loss, we set λ_1 and λ_2 to 5 and 0.06 respectively.

5.2. Evaluation metrics

Following the previous works [26, 36, 54], we use the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for evaluation of REC. They are defined as: $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|c_i - \hat{c}_i|$, $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(c_i - \hat{c}_i)^2}$, where c_i is the ground truth count and $\hat{c}_i$ is the predicted count for the i-th Image-RE pair in the test set of size n.

Besides MAE and RMSE, we also consider localization errors for REC, since it's easy to predict objects with a wrong attribute when all objects are of the same class. When the counting error is low for a model, it could be FP and FN predictions cancel each other out. To this end, we propose to use Precision, Recall and F1 score as localization metrics for REC. We first match the predicted points with the target points by Hungarian Matching. Then we calculate the TP, FP and FN based on matched point pairs. Specifically, we find the median of the predicted bounding box area $(w * h)$ and define $\sigma = \frac{\sqrt{w^2+h^2}}{2}$ as the threshold to determine whether a matched predicted point is a TP or FP. Then we compute FP by subtracting TP from the total number of predicted points, and compute FN by subtracting TP from the total number of target points. Finally, we calculate the Precision, Recall and F1 score as: $\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$, $\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$, $\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$.

5.3. Results on REC task

We compare existing language based counting methods, and open-set object detection model GroundingDino [21] with our proposed method on the REC benchmark. In addition, we show some qualitative results of our proposed method. Then we present the ablation study for the proposed global-local feature fusion and contrastive learning modules. Finally, we perform evaluation on the class agnostic counting task and compare with state-of-the-art methods.

Quantitative results. We first compare with language based counting methods including ZSC [54], TFOC [62], CounTX [2]. The fine-tuning of these models are based on the training script provided by the authors. The results are shown in Tab. 2. ZSC [54] uses the class name to generate an exemplar prototype, which is critical for exemplar patch selection and counting result. In REC task, a referring expression is more complex than a class name. Selected exemplars may not fully capture the complex semantics of the referring expression. Especially when the attribute is relational, exemplar-based methods will generally fail due to the lack of global context. The same problem applies to TFOC [62]. TFOC by text prompt essentially leads to finding the most appropriate box prompt for SAM [12], which segments the query image for objects resembling the reference object in the box prompt. For both ZSC and TFOC, the performance is limited by the quality of exemplar and the level of text-image alignment. CounTX

Method	Val set			Test set		
	MAE	RMSE	F1	MAE	RMSE	F1
full model	6.80	18.13	0.68	6.50	19.79	0.69
w/o global-local fusion	7.31	18.77	0.65	7.33	19.71	0.67
w/o contrastive learning	7.25	15.97	0.65	6.92	21.00	0.66

Table 3. Ablation study.

[2] on the other hand enforces the alignment between text and image by using transformer decoder layers. Therefore, CounTX achieves better performance in REC task. However, it's still not as effective as our proposed method in terms of reasoning about the text input and fusing global and local features. We couldn't report localization errors for ZSC and CounTX because their output is density map. For TFOC, the localization metrics are low due to the reason stated above.

We then compare our method with GroundingDino [21] which is a state-of-the-art open-set object detection model that can detect objects of unseen classes. Even without finetuning, GroundingDino has improved counting errors, but the localization performance is still poor. After finetuning, the localization performance is improved further. Lastly, our method outperforms all prior methods with the proposed global-local feature fusion and contrastive learning among different attributes.

Qualitative results. We show good examples in Fig. 5 and bad examples in Fig. 6 to illustrate the strength and weakness of our method. For the good examples in Fig. 5, our method can more easily handle attribute types: color (a, h), simple location (c, e, f) and action (b, d). The success with the color and action types comes from the model's prior knowledge as well as our proposed contrastive learning to differentiate one attribute from another. As shown in (d), for example, the model can differentiate "person standing" from "person sitting" and gives a close count with high TPs. While location attribute is more challenging, our method by fusing global and local features enhances global context learning to locate the target object. The model is able to tell the difference between "on the second top shelf" and other locations in (f).

For the bad examples in Fig. 6, our method fails to count the target object for attribute types of more complex location (a, c, g), negation (f), ambiguous (d) and attribute that involves text (e). The problem is mainly related to text reasoning and alignment of image and text features, which is still an open problem in the field of computer vision. We hope our work can shed some light on this challenge and inspire future research. Another challenging case is the misleading elements such as in image (h), many FPs are being taken into the count: the human figures on the building facade and the reflection of persons in the mirror.

Method	Setting	Val set		Test set	
		MAE	RMSE	MAE	RMSE
FamNet [36]	few-shot	23.75	69.07	22.08	99.54
BMNet+ [38]	few-shot	15.74	58.53	14.62	91.83
CounTR [4]	few-shot	13.13	49.83	11.95	91.23
CACViT [49]	few-shot	10.63	37.95	9.13	48.96
TFOC [62] (box prompt)	few-shot	37.56	113.14	19.95	132.16
ZSC [54]	zero-shot	26.93	88.63	22.09	115.17
CounTX [2]	zero-shot	17.10	65.61	15.88	106.29
TFOC [62] (text prompt)	zero-shot	47.21	127	24.79	137.15
GDino [21] (w/o finetune)	zero-shot	51.11	101.28	54.40	92.36
GDino [21] (w/ finetune)	zero-shot	10.32	55.54	10.82	104.00
GroundingREC (ours)	zero-shot	**10.06**	**58.62**	**10.12**	**107.19**

Table 4. Comparison of state-of-the-art methods and different settings on dataset FSC-147. Note: 1. GDino is short for GroundingDino; 2. TFOC [62] results for val set is obtained by running the code provided by the authors. All the other results are obtained from the original papers.

5.4. Ablation study

In this section, we present the ablation study to analyze the effectiveness of each component in our proposed method. The results are shown in Tab. 3. By removing the contrastive learning, the performance drops especially in the localization metric. It shows that by pushing the image and positive text embedding close to each other and pushing the negative text embeddings away, we actually help the model to learn discriminative features for different attributes. By removing the global-local feature fusion, the performance also drops. This shows the cross attention between global and local features indeed helps with global and relational context learning.

5.5. Results on class agnostic object counting

Besides REC task, we also evaluate our method on class agnostic counting to show our method is generalizable to prior tasks. The results are presented in Tab. 4. We evaluate the benchmark dataset FSC-147 [36] in a zero-shot manner, which means the model is trained on FSC-147 train set and evaluated on novel classes of the val and test set. The few-shot setting in Tab. 4 refers to the known exemplars provided by the dataset during inference.

Tab. 4 shows that our method GroundingREC outperforms the state-of-the-art methods in the zero-shot setting, and is even comparable to the methods in the few-shot setting. We also include our base model GroundingDino in the table for comparison. The similar performance is due to the fact that the text input is simply the class name, which doesn't require global context or discrimination of attributes. Our proposed modules work more effectively in the REC task.

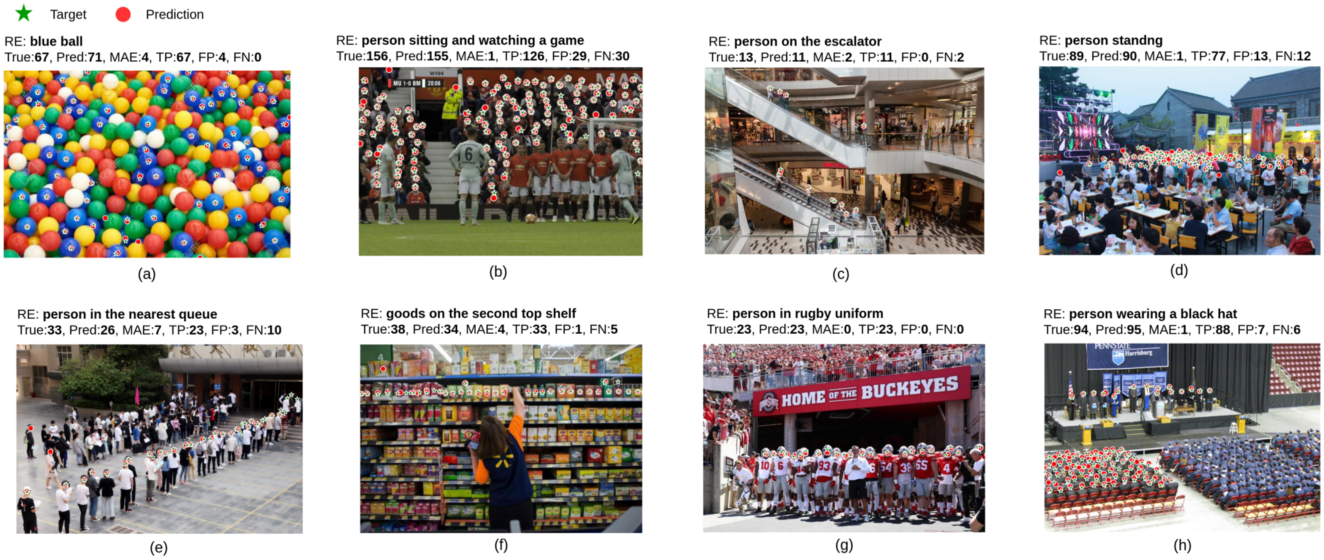

Target
Prediction
RE: blue ball
True:67, Pred:71, MAE:4, TP:67, FP:4, FN:0
(a)
RE: person sitting and watching a game
True:156, Pred:155, MAE:1, TP:126, FP:29, FN:30
(b)
RE: person on the escalator
True:13, Pred:11, MAE:2, TP:11, FP:0, FN:2
(c)
RE: person standng
True:89, Pred:90, MAE:1, TP:77, FP:13, FN:12
(d)
RE: person in the nearest queue
True:33, Pred:26, MAE:7, TP:23, FP:3, FN:10
(e)
RE: goods on the second top shelf
True:38, Pred:34, MAE:4, TP:33, FP:1, FN:5
(f)
RE: person in rugby uniform
True:23, Pred:23, MAE:0, TP:23, FP:0, FN:0
(g)
RE: person wearing a black hat
True:94, Pred:95, MAE:1, TP:88, FP:7, FN:6
(h)

Figure 5. Visualization of results for good examples (better viewed in enlarged version).

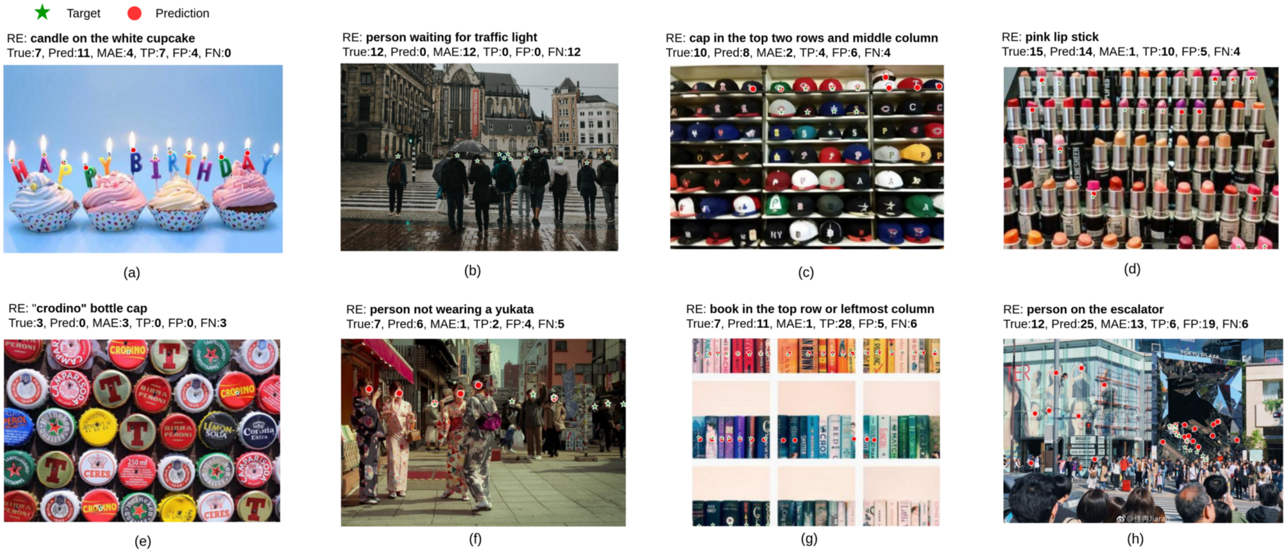

Target
Prediction
RE: candle on the white cupcake
True:7, Pred:11, MAE:4, TP:7, FP:4, FN:0
(a)
RE: person waiting for traffic light
True:12, Pred:0, MAE:12, TP:0, FP:0, FN:12
(b)
RE: cap in the top two rows and middle column
True:10, Pred:8, MAE:2, TP:4, FP:6, FN:4
(c)
RE: pink lip stick
True:15, Pred:14, MAE:1, TP:10, FP:5, FN:4
(d)
RE: "crodino" bottle cap
True:3, Pred:0, MAE:3, TP:0, FP:0, FN:3
(e)
RE: person not wearing a yukata
True:7, Pred:6, MAE:1, TP:2, FP:4, FN:5
(f)
RE: book in the top row or leftmost column
True:7, Pred:11, MAE:1, TP:28, FP:5, FN:6
(g)
RE: person on the escalator
True:12, Pred:25, MAE:13, TP:6, FP:19, FN:6
(h)

Figure 6. Visualization of results for bad examples (better viewed in enlarged version).

6. Conclusion

We propose a new counting task named Referring Expression Counting (REC) which aims to count fine-grained objects with different attributes within the class. To evaluate REC, we create a novel dataset named REC-8K, covering a variety of object classes and attributes types. We also propose a novel model GroundingREC which leverages the prior knowledge from a vision-language model. Through extensive experiments, we show that our model achieves state-of-the-art performance in both referring expression counting and zero-shot counting. By analysis of the experimental results, we show that our model is capable of learning local and relational attributes and discriminative attribute features, powered by the proposed modules. We also point out the limitations of our model in the qualitative analysis, which can be studied and improved in the future work.

Acknowledgement. This work is supported by Economic Development Board of Singapore (EDB) and Singapore Technologies Engineering Ltd (ST Engineering) under the Industry Postgraduate Program (IPP) Project code RS-EDBIP-00031.

References

[1] Amazon Web Services, Inc. Amazon mechanical turk, 2023. [Online; accessed 1-April-2023]. 3

[2] N. Amini-Naieni, K. Amini-Naieni, T. Han, and A. Zisserman. Open-world text-specified object counting. In *British Machine Vision Conference*, 2023. 1, 2, 6, 7

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 6

[4] Liu Chang, Zhong Yujie, Zisserman Andrew, and Xie Weidi. Countr: Transformer-based generalised visual counting. In *British Machine Vision Conference (BMVC)*, 2022. 7

[5] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2467–2474, 2013. 3

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 6

[7] Lin Geng Foo, Hossein Rahmani, and Jun Liu. Ai-generated content (aigc) for various data modalities: A survey, 2023. 2

[8] Michael Hobley and Victor Prisacariu. Learning to count anything: Reference-less class-agnostic counting with weak supervision. *arXiv preprint arXiv:2205.10203*, 2022. 2

[9] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal networks. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017. 2, 3, 4

[10] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2554, 2013. 4

[11] Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. *arXiv preprint arXiv:2305.07304*, 2023. 2

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2, 6

[13] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, 2018. 2

[14] Dongze Lian, Xianing Chen, Jing Li, Weixin Luo, and Shenghua Gao. Locating and counting heads in crowds with a depth prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[15] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. *European Conference on Computer Vision*, 2022. 2

[16] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022. 2, 3

[17] Hui Lin, Xiaopeng Hong, and Yabin Wang. Object counting: You only need to look at one, 2021. 2

[18] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *CVPR*, 2022. 2

[19] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23592–23601, 2023. 2, 3

[20] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2019. 2

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3, 4, 5, 6, 7

[22] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[23] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Estimating people flows to better count them in crowded scenes. In *The European Conference on Computer Vision (ECCV)*, 2020.

[24] Weizhe Liu, Nikita Durasov, and Pascal Fua. Leveraging self-supervision for cross-domain crowd counting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 4, 6

[26] E. Lu, W. Xie, and A. Zisserman. Class-agnostic counting. In *Asian Conference on Computer Vision*, 2018. 2, 6

[27] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1274–1282, 2020. 2

[28] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[29] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6142–6151, 2019. 2

[30] Zhiheng Ma, Xiaopeng Hong, and Shangguan Qinnan. Can sam count anything? an empirical study on sam counting. *arXiv preprint arXiv:2304.10817*, 2023. 2

[31] Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng. Spatial uncertainty-aware semi-supervised crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15549–15559, 2021. 2

[32] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3170–3180, 2023. 1, 2

[33] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306, 2023. 2

[34] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[35] Viresh Ranjan and Minh Hoai Nguyen. Exemplar free class agnostic counting. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 3121–3137, 2022. 2

[36] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3394–3403, 2021. 1, 2, 3, 4, 6, 7

[37] Aref Miri Rekavandi, Shima Rashidi, Farid Boussaid, Stephen Hoefs, Emre Akbas, and Mohammed bennamoun. Transformers in small object detection: A benchmark and survey of state-of-the-art, 2023. 4

[38] Min Shi, Lu Hao, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 7

[39] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *Technical Report*, 2020. 3, 4

[40] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. 2021. 2

[41] Wei Su, Peihan Miao, Huanzhang Dou, Yongjian Fu, and Xi Li. Referring expression comprehension using language adaptive inference, 2023. 2

[42] Khoi Nguyen Thanh Nguyen, Chau Pham and Minh Hoai. Few-shot Object Counting and Detection. In *Proceedings of the European Conference on Computer Vision 2022*, 2022. 2

[43] Jia Wan, Qingzhong Wang, and Antoni B Chan. Kernel-based density map generation for dense object counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[44] Jia Wan, Ziquan Liu, and Antoni B. Chan. A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1974–1983, 2021. 2

[45] Chunlei Wang, Wenquan Feng, Xiangtai Li, Guangliang Cheng, Shuchang Lyu, Binghao Liu, Lijiang Chen, and Qi Zhao. Ov-vg: A benchmark for open-vocabulary visual grounding. *arXiv preprint arXiv:2310.14374*, 2023. 2

[46] Mingjie Wang, Yande Li, Jun Zhou, Graham W. Taylor, and Minglun Gong. Gcnet: Probing self-similarity learning for generalized counting network, 2023. 2

[47] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023. 2

[48] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3, 4

[49] Zhicheng Wang, Liwen Xiao, Zhiguo Cao, and Hao Lu. Vision transformer off-the-shelf: A surprising baseline for few-shot class-agnostic counting, 2023. 7

[50] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 2020. 3

[51] Yixuan Wu, Zhao Zhang, Chi Xie, Feng Zhu, and Rui Zhao. Advancing referring expression segmentation beyond single image. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[52] Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2

[53] Haipeng Xiong, Hao Lu, Chengxin Liu, Liu Liang, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8362–8371, 2019. 2

[54] Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15548–15557, 2023. 1, 2, 6, 7

[55] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023. 2

[56] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 3

[57] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, and Errui Ding. Perspective-guided convolution networks for crowd counting. 2019. 2, 3

[58] Shuo-Diao Yang, Hung-Ting Su, Winston H. Hsu, and Wen-Chin Chen. Class-agnostic few-shot object counting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 870–878, 2021. 2

[59] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 2

[60] Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le. Few-shot object counting with similarity-aware feature enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6315–6324, 2023. 2

[61] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19456–19465, 2023. 2, 3

[62] Mengmi Zhang Zenglin Shi, Ying Sun. Training-free object counting with prompts. In *WACV*, 2024. 2, 6, 7

[63] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*, 2023. 3

[64] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *ICCV*, 2023. 3

[65] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021. 3