

What is the Right Embedding Space for Contrastive Learning in Referring Expression Counting?

Kostas Triaridis^{*1}, Panagiotis Kaliosis^{*1}, E-Ro Nguyen¹, Jingyi Xu¹,
Dimitris Samaras¹, and Hieu Le²

¹ Stony Brook University

² UNC-Charlotte

Abstract. Referring Expression Counting (REC) requires distinguishing visually similar objects described by fine-grained text cues. Existing methods tackle this via image-text contrastive learning of visual features which aims to distinguish visual features corresponding to positive and negative referring expressions. However, this strategy relies heavily on accurate image-text alignment and is constrained by the limited number of available negative expressions. We argue that these limitations can be avoided by performing contrastive learning entirely in the visual embedding space. To this end, we introduce C-REX, a simple but effective supervised contrastive learning framework that learns fine-grained visual representations by contrasting visual tokens within the same image. By shifting the negative samples from incorrect Referring Expressions to visual tokens from the image, C-REX vastly increases the number of potential negatives, providing richer and more stable supervision. This design leads to stronger fine-grained visual discrimination and better generalization across complex open-world counting settings. C-REX serves as a plug-in module that can be seamlessly applied to any existing REC model without architectural changes. We apply it in three different REC architectures and achieve state-of-the-art results improving by up to 28% in MAE and 24.5% in RMSE. Moreover, we show that our framework is versatile and general, and can be applied to other counting tasks like class-agnostic counting, improving the performance of existing models.

Keywords: Referring Expression Counting · Contrastive Learning · Representation Learning

1 Introduction

Referring Expression Counting (REC) [19] generalizes object counting from pre-defined classes to open-world settings, where the targets are specified by free-form referring expressions (REs) that describe both the object category and its

* Equal contribution. Correspondence to kostas@cs.stonybrook.edu

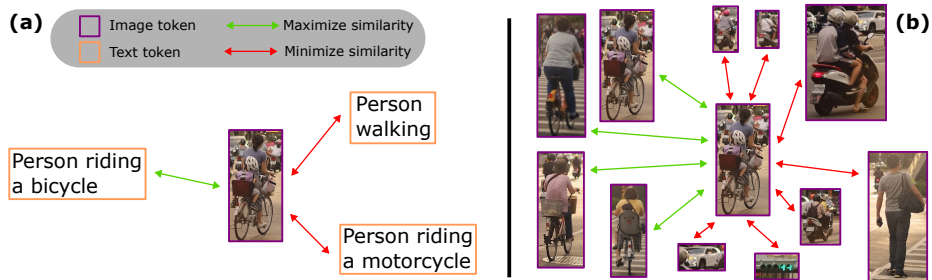


Fig. 1: (a) Standard REC models attempt to distinguish visually similar but contextually distinct objects through image-text contrastive learning. This approach can fall short when the image and text embedding spaces are poorly aligned. (b) We propose operating **solely** in the **image-space** for contrastive learning, directly optimizing relationships among visual embeddings. This provides stronger, more stable supervision, improving fine-grained contextual discrimination between visually similar objects.

fine-grained attributes. Instead of counting all instances of a predefined class, REC enables queries like “*person walking*,” or “*person riding a motorcycle*,” for the input image in Figures 1 and 2. This formulation brings object counting closer to real-world scenarios, where users naturally express counting targets in language rather than selecting from a fixed set of labels. However, this flexibility introduces a key challenge: developing robust, fine-grained visual representations that can precisely separate visually similar objects under diverse expressions.

Previous works [19, 67] address this challenge using image-text contrastive learning (CL). The model is trained to distinguish visual objects associated with a positive prompt (e.g., “*person riding a bicycle*”) from those paired with negative prompts (e.g., “*person riding a motorcycle*” or “*person walking*”), by contrasting the corresponding visual tokens with the positive and negative text tokens directly. However, this approach faces several practical limitations. The negative prompts are sampled from existing textual descriptions for a given image, which severely limits the number of plausible and useful negative samples. As a result, the supervision signal is noisy and unstable as contrastive objectives typically require large batch sizes with numerous negative samples to learn robust fine-grained representations [15, 26, 28, 35, 60]. In addition, the entire contrastive process is anchored in the text embedding space, so its effectiveness strongly depends on the typically poor quality of image-text alignment and the precision of the underlying language representations. This misalignment results into visual embeddings that entangle semantically distinct but visually similar objects, limiting the model’s ability to learn precise fine-grained distinctions.

A straightforward approach to address these misalignment issues is using a contrastive objective that operates in the image space only. Even though image-space CL has been used with great effectiveness in multiple different visual tasks [15, 20, 29, 64], its application in REC has not been explored. To this end, we introduce **C-REX** (Contrastive Learning for Referring Expression Count-



Fig. 2: Existing REC methods [19, 67] rely on image-text contrastive learning approaches that depend on text guidance, which can introduce misalignment and limit fine-grained separation between visually similar instances. **(a)** Each anchor visual token (in blue) is compared to the positive (green) and negative (red) prompt text embeddings. The image and text embedding spaces are not necessarily aligned which can lead to the visual tokens that correspond to those different prompts to be entangled, as they are not separated directly. **(b)** Instead, we propose performing **image-only** contrastive learning, directly optimizing relationships among object embeddings in the visual space. This yields more compact intra-class clusters and clearer inter-class separation, improving discrimination between visually and contextually distinct objects.

ing), a simple supervised contrastive learning framework tailored for REC that shifts the supervision from the text-image space to operate entirely in the image embedding space. Instead of relying on linguistic anchors, C-REX guides visual tokens of objects corresponding to the same RE to be embedded close together and away from the embeddings of objects corresponding to different expressions or classes, as shown on the right panel of Figure 2. This promotes a visual representation space sensitive to fine-grained visual differences that correspond to variation in semantic context. The proposed contrastive objective is more stable as the scarcity of negative samples has been addressed; the model can leverage a large pool of negatives (>100) in the form of visual tokens instead of relying on the limited REs (typically less than 5).

C-REX introduces an intra-image contrastive learning objective that operates entirely in the visual embedding space. Concretely, for each Image-RE pair we match predicted points to the N ground truth points by performing Bipartite Matching. The image tokens responsible for the matched predictions are treated as positive samples in the contrastive objective, whereas all other tokens are treated as negatives. This way, the RE is only used as input to the visual backbone to condition the generated image tokens through cross-attention. Note that, unlike the standard supervised contrastive objective [35], we do not pull negative samples together, since they lack semantic coherence, but only push them

away from the positives. This improves performance as shown in Section 4.5. The simple shift to image-space CL effectively sharpens the visual feature space, improving intra-class separability and enhancing the model’s ability to handle subtle, attribute-level distinctions in open-world counting.

A key strength of our framework lies in its **generality** and **versatility**. C-REX can be plugged into any REC architecture without modification. We validate this by integrating it into GroundingREC [19], CADGD [67], and a detection-based baseline built on GroundingDINO [39]. For GroundingREC and CADGD, we replace their image-text contrastive loss with our image-space objective; for GroundingDINO, we add it alongside the standard counting loss. In all cases, C-REX yields consistent performance gains while preserving the original design. We further verify that combining image-text and image-space contrastive objectives does not improve performance, with image-space supervision alone consistently yielding the best results (Section 4.5). Beyond REC, the same principle of learning fine-grained discriminative features through contrastive alignment in the image-space only, can be applied to other counting settings. We adapt C-REX to **class-agnostic counting** by matching predicted and ground-truth points per class to form positives and treating the remaining samples as negatives. With this simple modification, C-REX also boosts performance for text-based counting models in the class-agnostic setting (Section 4). To summarize, our contributions are the following:

- We demonstrate that counting tasks can benefit from performing contrastive learning in the image (visual) embedding space *only*, rather than in a typically poorly aligned image-text space. We introduce C-REX, a supervised CL framework for REC that operates directly in the visual embedding space.
- Our framework produces fine-grained, context-aware visual representations and achieves state-of-the-art performance across multiple REC architectures, improving by up to 28% in MAE and 24.5% in RMSE.
- We show that C-REX can be applied to any existing REC architecture and can also be deployed in other counting settings that require learning fine-grained discriminative representations, such as class-agnostic counting.

2 Related Work

Early works in object counting focused mostly on class-specific counting [1, 63, 76], addressing challenges in diverse domains such as crowd counting [41, 45, 55, 62, 73] or traffic analysis [11]. More recent work has moved to class-agnostic counting (CAC) [32, 56, 57] where the aim is to generalize counting across various object categories, typically not encountered during training [23, 30, 43, 59, 66, 70]. To enable this, many methods used textual description to describe the classes [5, 8, 13, 34, 40, 77], while others use a small number of annotated examples as visual reference [16, 31, 33, 47, 71] to achieve better results. A further line of work investigates whether the reasoning capabilities of Vision–Language Models (VLMs) can also enhance performance [4, 52]. VLMs are rapidly improving, however in practice, they increasingly rely on external, task-specific components to

use as tools [3]. Counting models specifically have been utilized in this manner [7] and have shown to outperform general use VLMs, outlining the continuing need for dedicated counting models.

Another way the counting task has been approached is in the context of Visual Question Answering (Counting VQA) [2, 14, 25, 61, 78], where a VQA model is tasked with answering a question that requires counting objects in an image. TallyQA [2] focuses on complex counting scenarios where the VQA model has to distinguish between visually similar objects given fine-grained textual descriptions. Similarly, Dai *et al.* [19] introduced REC, where the goal is to count only the subsets of instances of a class in an image that match a given RE. Each RE consists of a **class**, specifying the object category, and an **attribute**, describing its fine-grained characteristics. Direct counting methods like REC differ from counting VQA in that the models are tasked to also localize the predicted objects to ensure more trustworthy predictions. REC is also closely related to Referring Expression Segmentation [44, 46, 65, 68] and Referring Expression Comprehension [53, 58], which focus on segmentation or localization rather than counting.

Early approaches on object counting mostly relied on density map regression [10, 36, 37] or bounding box detection [18, 27] in order to predict the object count. Density-based methods estimate the final count by summing over a density map, being traditionally more accurate in cluttered and densely populated scenes where detection-based methods struggle [6]. However, a critical limitation of density-based methods is the lack of object-level detail which limits their applicability in cases that require localization and explainability [24, 51]. This is especially important in REC, where images contain multiple visually similar instances of the same category, each associated with a different referring expression. In such cases, ensuring that the final count accurately reflects only the correct subset of objects is crucial for reliable and precise counting. Recent approaches, such as CountGD [6] and GeCo [49], leverage state-of-the-art open-set object detectors [39, 75], and reach performance on-par with the best density-based methods [50]. CAD-GD [67] combines density prediction with localization for REC establishing the state-of-the-art for this task.

For REC it is crucial to be able to disambiguate between visually similar objects with subtle contextual differences. For this reason, Dai *et al.* [19] introduce a CL module that contrasts image and RE embeddings, which CAD-GD [67] also employs. Contrastive learning in the image-text space for REC has several disadvantages regarding robustness of learned representations and efficiency (see discussion in Section 3.1) which we address by moving the contrastive supervision to the image space only. CL in the image space only, beyond self-supervised pretraining [15, 26, 28, 60], has also been explored and seen success in other vision tasks such as person re-identification (ReID) [20, 29, 72, 74], where the task is to distinguish between people that are potentially visually similar but correspond to distinct identities. REC faces a similar challenge: models must distinguish between visually similar objects (same class), while discriminating based on the RE context. Our approach in REC differs from those in ReID as the CL samples are visual tokens originating from the same image, instead of different images.

Additionally, we define positive and negative sets for CL on the fly by matching predicted image tokens to ground truth targets conditioned on each RE.

3 Method

For REC it is important to be able to robustly differentiate objects that belong in the same class but have different attributes (REs). To this end we propose a novel contrastive learning module, that learns more robust discriminative features using contrastive supervision on the image space only.

Our goal is to perform contrastive learning in the image embedding space only, as text-image contrastive learning tends to be less stable. In the context of REC, we want our approach to accommodate multiple positive samples, since an image typically contains multiple instances of objects that correspond to the same class-RE pair. To ensure effective representation learning, we aim to bring the embeddings of these instances closer together while maintaining clear separation from objects that correspond to the same class but different RE or a different class altogether. For this reason our proposed loss is based on the supervised contrastive loss [35] that extends the standard contrastive loss to be able to handle multiple positive samples for each anchor. The general formulation of the supervised contrastive objective is as follows: Given a set I of samples (*e.g.* images), for each sample $i \in I$ we get the embeddings z_i , and a set of other samples that correspond to the same “label” as sample i , which we call the set of positives $p \in P(i)$ for i . The loss is calculated as:

$$\mathcal{L}_{\text{sup}} = \sum_{i \in I} -\log \left(\frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{\alpha \in I - \{i\}} \exp(z_i \cdot z_\alpha / \tau)} \right) \quad (1)$$

For our approach, we choose the samples $i \in I$ to correspond to predicted image tokens from our visual backbone. We use the generated image tokens to predict the location of the target objects, as is standard in counting models [6, 19, 67]. We aim to contrast the image tokens that correspond to the objects for a given class and RE (same attributes, semantic context) to the ones that correspond to the same class but different RE or to different classes altogether. To assign labels to our samples $i \in I$ we match them to the N ground truth points via Bipartite Matching, based on their predicted locations. For each Image-RE pair we divide the samples into a positive (I^+) and negative (I^-) class; the image tokens that get matched to ground truth points correspond to the positive class and all other tokens are treated as negatives.

The standard formulation of the supervised contrastive loss pushes samples from the same class closer together. In our case this would mean that samples from the negative class I^- , which may contain samples from a diverse set of class labels, REs, or even tokens corresponding to no class label would be clustered together, which is undesirable and counterproductive. For example, for

the image in Figure 2, given the RE “Person riding a bicycle”, the set of negatives would include tokens that correspond to “people walking”, “cars” or “traffic lights”, which should not be clustered together. To address this, we modify the supervised contrastive objective to only use samples from the positive class as anchors. Based on this modification, our revised supervised contrastive loss can be formulated as follows:

$$\mathcal{L}_{\text{sup}}^* = \underbrace{\sum_{i \in I^+}}_{\text{optimize positives only}} - \log \left(\frac{1}{|I^+| - 1} \sum_{p \in I^+ - \{i\}} \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{\alpha \in I - \{i\}} \exp(z_i \cdot z_\alpha / \tau)} \right) \quad (2)$$

We verify that the motivation for this modification is well-grounded and supported by demonstrating its effectiveness in our ablation study (Section 4.5).

Our final framework, **C-REX**, replaces the image–text contrastive loss used in previous REC methods [19, 67] with our proposed image-space contrastive loss and uses all other counting losses (*e.g.* localization, classification or density loss) unchanged for each model we apply it to.

3.1 Advantages of image-space over image-text CL methods

Previous work on REC [19, 67] also tried to learn more discriminative features by contrasting image embeddings to text embeddings of “candidate” REs. However, their approach has key limitations in the context of contrastive learning. Specifically, their contrastive loss operates on image-text pairs, which presents additional challenges due to the inherent misalignment between image and text embedding spaces [38]. In contrast, our method operates entirely within the image space, ensuring a more stable contrastive signal. Contrastive learning in a language-image setting generally requires more data and larger-scale training to achieve robust representations [17, 54], whereas contrastive learning in the image space is more efficient [48, 75]. This is evident when comparing the performance of text-image models like CLIP [54] to image-specific models like DINOv2 [48] on smaller datasets, where DINOv2 achieves comparable performance to CLIP even when trained on significantly less data [48].

Additionally, a crucial factor in contrastive learning is the presence of a large batch size with numerous negative samples, as this increases the diversity and variability of samples included, thus improving the robustness of the learned representations [15, 26, 28, 35, 60]. However, [19, 67] are limited in that regard, as the number of negative samples is constrained by the number of existing REs for a single class in an image, typically fewer than four. Our contrastive loss overcomes this constraint by leveraging a significantly larger pool of negatives, typically in the hundreds, consisting of image tokens that were not matched to ground truth objects, and so ended up in the negative set of our contrastive framework. These advantages are reflected in our experimental results (Section 4), where

we compare two state-of-the-art models [19, 67] under their original image-text contrastive framework and our proposed C-REX. In both cases, replacing the traditional contrastive objective with C-REX yields consistent improvements across all evaluation metrics.

3.2 Extension to class-agnostic counting

C-REX is designed to be general and versatile, enabling models to distinguish objects with subtle, context-dependent variations, such as those described by REs. This allows us to adapt it to any task that requires multiple image-space object embeddings to be contrasted based on a set of distinguishing attributes. Following this principle, we introduce a simple adaptation for class-agnostic counting, where positive samples are selected based on a target class, rather than an RE. This is less critical in this counting setting where models need to differentiate between objects of different classes, whose image embeddings are already different enough, so we do not expect to see the same amount of improvement as in REC. With this adaptation, our method can improve the performance of the best text-based counting models in the class-agnostic counting task, as shown in Section 4, although the improvement is less significant, as was expected.

4 Experiments

We evaluate C-REX as a general framework that can be applied to existing REC architectures. Specifically, we integrate our proposed framework into two state-of-the-art REC models; GroundingREC [19] and CAD-GD [67], as well as a simple detection-based baseline built on GroundingDINO [39]. For the former two, we replaced their original image-text contrastive objectives with our image-space contrastive loss. For the GroundingDINO baseline, we train using our proposed objective alongside the standard counting objectives used in previous detection-based models [6, 19, 67].

4.1 Implementation Details

For GroundingREC [19], we use the publicly released model checkpoint and evaluation protocol. Since the model and its configuration were reproducible, we directly report their results as our baseline performance. For CAD-GD [67], since no trained checkpoint is publicly available, we retrained the model using the authors’ official open-source implementation. For the required density maps, we followed the same procedure as the authors, established in prior work [37]. More details are provided in the Supplementary.

Training For all models, we use either a Swin-T [42] or Swin-B image encoder and BERT [22] text encoder, both kept frozen during training. We fine-tune only the feature enhancer and cross-modality decoder components. For CAD-GD [67], we follow their training procedure. For the REC-8k dataset we train for 20 epochs and use a learning rate of $1e-5$ that decays by 10x in the 10th

Table 1: Comparison of results on the REC-8K dataset. GrREC and GrDINO correspond to GroundingREC [19] and GroundingDINO [39]. For GroundingREC [19], results are obtained from their released models. For CAD-GD [67], since no pre-trained models are available, we retrain on the REC-8K dataset using their publicly available code. CAD-GD[†] refers to an alternative strategy of selecting positive detections, described by [67], where instead of using a static threshold for all tokens, we select the top N based on the density map estimate. Best result for each backbone in bold.

	Method	Validation set					Test set				
		MAE↓	RMSE↓	Prec↑	Rec↑	F1↑	MAE↓	RMSE↓	Prec↑	Rec↑	F1↑
Swin-T	GrREC	6.80	18.13	0.65	0.71	0.68	6.50	19.79	0.67	0.72	0.69
	+ C-REX	4.74	14.50	0.70	0.71	0.71	4.91	18.87	0.71	0.72	0.72
	CAD-GD	6.33	15.90	0.65	0.69	0.67	6.25	18.91	0.67	0.70	0.69
	+ C-REX	5.98	14.84	0.66	0.70	0.68	5.97	17.37	0.67	0.73	0.70
	CAD-GD [†]	4.65	12.87	0.65	0.69	0.67	4.81	15.02	0.67	0.70	0.69
	+ C-REX	4.70	13.78	0.66	0.70	0.68	4.61	14.71	0.67	0.73	0.70
	GrDINO	5.92	17.09	0.65	0.65	0.65	5.90	19.73	0.68	0.68	0.68
	+ C-REX	4.26	12.89	0.75	0.72	0.73	4.38	17.93	0.76	0.74	0.75
Swin-B	GrREC	5.66	15.24	0.66	0.77	0.71	5.42	18.47	0.71	0.69	0.70
	+ C-REX	4.57	14.22	0.72	0.72	0.72	4.58	18.07	0.74	0.72	0.73
	CAD-GD	5.31	13.68	0.70	0.73	0.72	5.36	16.14	0.73	0.73	0.73
	+ C-REX	5.06	12.73	0.70	0.74	0.72	5.02	15.09	0.72	0.74	0.73
	CAD-GD [†]	4.40	12.20	0.70	0.73	0.72	4.58	13.81	0.73	0.73	0.73
	+ C-REX	4.38	12.43	0.70	0.74	0.72	4.49	13.21	0.72	0.74	0.73
	GrDINO	4.58	14.64	0.70	0.76	0.73	4.64	17.58	0.72	0.76	0.74
	+ C-REX	4.00	14.49	0.78	0.76	0.77	4.20	16.76	0.79	0.77	0.78

epoch and for FSC-147 we train for 30 epochs and use a learning rate of $1e-5$ that decays by 10x in the 15th epoch. For GroundingREC and GroundingDINO we follow Dai *et al.* [19] and do not decay the learning rate during training.

Inference Following prior works [12, 19, 67], we generate 900 output tokens per image and apply the thresholding strategy of Dai *et al.* [19] (0.25 for the CLS token, 0.35 for the remaining tokens) to determine positive detections. For CAD-GD we also evaluate using their alternate proposed strategy (CAD-GD[†]), where, instead of using a static threshold, we select the top N candidate tokens based on the density map estimate as positive detections.

4.2 Datasets & Metrics

REC-8K [19] consists of 8,011 images, each paired with 2.13 REs on average. In total, the dataset contains 17,122 image-RE pairs, where each pair is annotated with the ground-truth points pinpointing the target objects in the image.

FSC-147 [56] is a widely used object counting dataset, which consists of 6,135 images spanning 147 classes, split in a non-overlapping manner. Each image is annotated with three visual exemplars.

Metrics Following prior work [6, 19, 50, 69], we report the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to evaluate and compare counting performance. Moreover, we also report precision, recall and F1 score since counting the correct object instances is also an important factor.

4.3 Results in REC

Quantitative Comparison Table 1 presents our experimental results in the REC-8K dataset, comparing the performance of three detection-based REC architectures—GroundingREC [19], CAD-GD [67], and GroundingDINO [39]—evaluated both in their original configurations and with C-REX applied. All architectures are evaluated using Swin-T and Swin-B backbones. As shown in Table 1, integrating C-REX consistently improves performance across models, backbones and metrics, confirming that replacing the image–text contrastive loss with our image-space objective yields more stable and discriminative representations. C-REX achieves the lowest overall errors and the highest F1 scores (results in **bold**) in most settings, establishing new state-of-the-art performance on REC-8K. We also find that using just C-REX improves performance compared to using both image-text and image-space contrastive losses jointly (Section 4.5).

Figure 4 reports MAE and RMSE across different object count ranges on the REC-8K test set, allowing us to examine model behavior under varying object densities. Across all three REC architectures, integrating C-REX consistently reduces error relative to the corresponding baseline across all count regimes. The improvements are particularly evident in the low- and mid-density ranges (0–20, 21–50, and 51–100 objects), where fine-grained attribute distinctions make the task particularly challenging. These results indicate that the proposed image-space contrastive supervision improves fine-grained instance discrimination and leads to more accurate counting across a wide range of object densities.

C-REX forms positive and negative sets from the fixed number of tokens provided by the visual backbone, so extremely dense scenes can reduce the number of tokens available to serve as negatives. While this could lead to a lower quality supervision signal, we do not observe a negative effect in model performance. Even in high-count regimes, C-REX usually provides a sufficient number of negative image tokens. Additionally, as vision backbone models advance and continue to increase their visual token capacity, the available pool of negatives will organically grow, allowing C-REX to scale more effectively with more expressive backbones than image-text CL methods.

Qualitative Comparison Figure 3 presents qualitative comparisons across the three REC architectures both in their original configurations and with C-REX applied. Each column corresponds to a different RE, covering a diverse set of fine-grained distinctions such as object color (“blue car”), orientation (“car driving to the left”), positional relationships (“nail polish in the top layer”), and actions



Fig. 3: Qualitative comparison of REC predictions across different models. The first row shows the input images and the second row visualizes the ground-truth instance annotations for each referring expression. The remaining rows present predictions from each baseline model (GroundingREC, CAD-GD, and GroundingDINO) followed by their corresponding versions with C-REX applied. Each column corresponds to a different referring expression. Across all examples, models trained with C-REX provide more precise counts and localize a more accurate subset of objects specified by the referring expression. Image contrast is adjusted to better visualize model predictions.

(“person standing”). The first two rows show the input images and ground-truth annotations, while the remaining rows present predictions from each REC model, both in their original baseline form and with C-REX applied.

Across all cases, we observe that integrating C-REX leads to more accurate predictions, with regards to both predicted object count and locations. For instance, in the “car driving to the left” example, all three baseline models

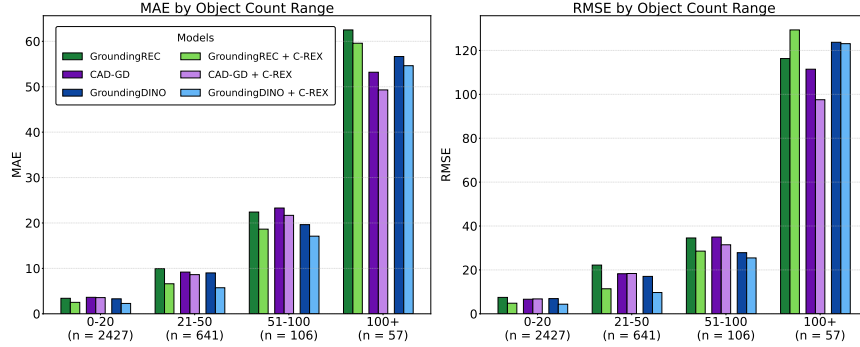


Fig. 4: Performance of each baseline REC model compared to its counterpart trained with our proposed C-REX framework, evaluated across object count ranges on REC-8K. Sample counts are below each bin. In most cases, C-REX yields lower MAE and RMSE compared to the corresponding baseline configuration, demonstrating consistent improvements across low, mid, and high-density conditions.

overcount by including vehicles moving in the opposite direction, whereas their C-REX counterparts more consistently focus on the intended subset. The “blue car” example in the first column also highlights the effect of C-REX: despite the scene containing only two vehicles, all baseline models fail to isolate the correct car, either detecting both or missing it entirely. The distinction between the blue car and the darker neighboring one is subtle yet semantically important, and C-REX consistently identifies the correct one. Moreover, in the “nail polish in the top layer” example, C-REX demonstrates better spatial discriminative ability by concentrating predictions to the correct layer. We provide more examples of both well performing and failure cases in the supplementary.

4.4 Results in Class-Agnostic Counting

To assess the generality of C-REX beyond REC, we adapt it for the class-agnostic counting setting using the FSC-147 benchmark dataset [56]. We apply C-REX to our three detection-based baseline architectures [19, 39, 67] and compare each model’s performance in both its original (vanilla) configuration and with C-REX integrated. Table 2 reports the results, including comparisons with state-of-the-art class-agnostic counting models. In most cases, incorporating C-REX yields consistent improvements across all architectures and both evaluation metrics, indicating that C-REX transfers effectively to class-agnostic counting and benefits performance independent of the underlying model design.

Table 2: Comparison of results on the FSC-147 [56] dataset, using only textual information (class names) as a guide to perform counting. For all models, except CAD-GD and GroundingDino, we report results from their original papers. Best results in bold.

Method	Val set		Test set	
	MAE	RMSE	MAE	RMSE
ZSC [69]	26.93	88.63	22.09	115.17
CounTX [5]	17.10	65.61	15.88	106.29
CountGD [6]	12.14	47.51	14.76	120.42
DAVE [50]	15.48	52.57	14.90	103.42
CAD-GD [67]	13.80	64.74	12.15	99.87
+ C-REX	13.86	61.04	11.90	99.65
GroundingDino	9.71	55.11	10.73	103.79
+ C-REX	9.42	54.74	10.04	102.35
GroundingREC [19]	10.06	58.62	10.12	107.19
+ C-REX	9.66	54.07	12.46	104.53

4.5 Ablation Study

Image-Text vs Image-Space CL We compare the image-text contrastive loss used in previous REC methods [19, 67] to our proposed image-image CL and explore whether they can be complementary (Table 3). Image-space CL consistently outperforms image-text CL, as well as their combination, indicating that contrastive supervision in the image space is more effective for REC. Adding image-space CL on top of image-text CL still improves over the image-text baseline, further supporting that image-text CL is suboptimal for REC.

Selection of positive samples We verify that choosing the positive set by Bipartite Matching outperforms alternative potential selection strategies. We compare it to strategies that select the most similar tokens compared to the text prompt. We show this for different thresholds for the number of most similar tokens: a static threshold, the ground truth count N and $N +$ a small dynamic window to account for the uncertainty of this selection strategy (Table 4a). Our matching-based selection strategy outperforms alternatives.

Supervised CL modification In Table 4b we examine the effectiveness of our proposed modification ($\mathcal{L}_{\text{sup}}^*$) to the supervised contrastive objective. We show that it significantly improves performance compared to both the unmodified version and the vanilla configuration of our GroundingDino baseline. This highlights the importance of our decision to use **only positive** samples as anchors, confirming its crucial role in improving the model’s effectiveness.

Robustness to Referring Expression Paraphrases To evaluate robustness to linguistic variation, we paraphrase the REs in the REC-8K test set using

Table 3: Comparison of contrastive learning strategies: image-text (I-T), image-space (*I*, *ours*), and their combination (I-T & I). Across two architectures, image-space contrastive learning consistently outperforms the standard image-text objective, while combining the two does not yield further improvements. This suggests that contrastive supervision in the image space alone provides the most effective signal for REC.

Method	CL	Val set			Test set		
		MAE↓	RMSE↓	F1↑	MAE↓	RMSE↓	F1↑
GroundingREC	I-T	6.80	18.13	0.68	6.50	19.79	0.69
	I-T & I	5.39	15.39	0.68	5.36	19.17	0.69
	I	4.74	14.50	0.71	4.91	18.87	0.72
CAD-GD	I-T	6.33	15.90	0.67	6.25	18.91	0.69
	I-T & I	6.15	14.56	0.67	6.47	19.04	0.68
	I	5.98	14.84	0.68	5.97	17.37	0.70

Table 4: (a) Ablation on positive sample selection strategy in our proposed contrastive framework. N is the ground truth object count and BM stands for Bipartite Matching. (b) Ablation on the GroundingDino (GrDino) baseline comparing our modified supervised contrastive loss ($\mathcal{L}_{\text{sup}}^*$) with the standard formulation ($\mathcal{L}_{\text{sup}}$) on REC-8k. The proposed modification consistently improves performance across metrics. For this analysis, we select as positives the top- N image tokens most similar to the text prompt.

Selection Strategy	Val set		Test set	
	MAE	RMSE	MAE	RMSE
top 5	7.58	18.34	7.51	20.83
top $N + \sqrt{N}$	6.02	15.79	6.07	18.80
top N	4.86	13.60	5.06	17.53
BM	4.26	12.89	4.38	17.93

(a)

Method	Val set		Test set	
	MAE	RMSE	MAE	RMSE
GrDino	5.92	17.09	5.90	19.73
+ $\mathcal{L}_{\text{sup}}$	5.69	14.98	5.91	19.34
+ $\mathcal{L}_{\text{sup}}^*$	4.86	13.60	5.06	17.53

(b)

Claude [9] and evaluate each model on the paraphrased expressions (Table 5a). C-REX improves performance for GroundingREC and GroundingDINO across all metrics, suggesting that image-space contrastive supervision can improve robustness to text perturbations, while for CAD-GD, the baseline performs better.

4.6 Comparison with off-the-shelf foundation VLM

We evaluate Molmo2-4B [21] in REC by prompting it to “Point to all $\langle RE \rangle$ in the image”, either in a zero-shot setting or in a two-shot setting on the REC-8K test set with examples selected from the same semantic category as the query.

Table 5: (a) Robustness to paraphrased REs on the REC-8K test set. C-REX improves performance for two of the three REC architectures, indicating robustness to linguistic variations. **(b)** Comparison of Molmo2-4B [21] on the REC-8K test set against REC models equipped with C-REX. Molmo performs substantially worse, highlighting the importance of specialized visual representations for REC.

Method	Test set		
	MAE↓	RMSE↓	F1↑
GrREC	9.88	24.77	0.54
GrREC + C-REX	<u>6.22</u>	<u>19.41</u>	<u>0.66</u>
CAD-GD	<u>6.85</u>	<u>18.29</u>	<u>0.68</u>
CAD-GD + C-REX	8.02	20.86	0.67
GrDINO	6.34	21.35	0.67
GrDINO + C-REX	<u>5.66</u>	<u>19.48</u>	<u>0.72</u>

(a)

Method	Test set		
	MAE↓	RMSE↓	F1↑
Molmo 0-shot	10.04	32.18	0.51
Molmo few-shot ($n = 2$)	25.43	42.10	0.39
GroundingREC + C-REX	4.58	18.07	0.73
CAD-GD + C-REX	5.02	15.09	0.73
GroundingDINO + C-REX	4.20	16.76	0.78

(b)

Molmo, despite being partially pre-trained on counting, performs substantially worse than all REC models equipped with C-REX (Table 5b). Interestingly, even prompting with semantically relevant in-context examples leads to worse performance, showing that REC remains challenging for modern MLLMs.

5 Conclusion

In this work, we introduced C-REX, a novel contrastive learning framework designed to tackle Referring Expression Counting (REC) by improving detection-based counting models’ capabilities in distinguishing visually similar but contextually distinct objects. By operating entirely within the image space, C-REX eliminates the misalignment issues inherent in image-text contrastive learning, ensuring a more stable contrastive signal and leveraging a significantly larger pool of negative samples for improved robustness. We demonstrated the effectiveness of C-REX by applying it on three existing REC architectures [19, 39, 67], consistently improving their performance and achieving state-of-the-art results. Beyond REC, we showed that the same framework extends naturally to class-agnostic counting, underscoring its generality and adaptability across related visual counting tasks.

Acknowledgements

This research was partially supported by NSF grants IIS-2123920 and IIS-2212046 and by the CCI startup fund at UNC Charlotte.

References

1. Abousamra, S., Hoai, M., Samaras, D., Chen, C.: Localization in the crowd with topological constraints. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**(2), 872–881 (May 2021)
2. Acharya, M., Kafle, K., Kanan, C.: Tallyqa: Answering complex counting questions. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 33, pp. 8076–8084 (2019)
3. Schick et al., T.: Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* (2023)
4. Alghisi, S., Roccabruna, G., Rizzoli, M., Mousavi, S.M., Riccardi, G.: [de] re] constructing vlms’ reasoning in counting. *arXiv preprint arXiv:2510.19555* (2025)
5. Amini-Naieni, N., Amini-Naieni, K., Han, T., Zisserman, A.: Open-world text-specified object counting. In: *British Machine Vision Conference* (2023)
6. Amini-Naieni, N., Han, T., Zisserman, A.: Countgd: Multi-modal open-world counting. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2024)
7. Amini-Naieni, N., Zisserman, A.: Countgd++: Generalized prompting for open-world counting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 37725–37734 (2026)
8. Amini-Naieni, N., Zisserman, A.: Open-world object counting in videos. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2026)
9. Anthropic: The claude 3 model family: Opus, sonnet, haiku. Tech. rep., Anthropic (2024)
10. Arteta, C., Lempitsky, V., Noble, J.A., Zisserman, A.: Interactive object counting. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *Computer Vision – ECCV 2014*. pp. 504–518. Springer International Publishing, Cham (2014)
11. Asha, C.S., Narasimhadhan, A.V.: Vehicle counting for traffic management system using yolo and correlation filter. In: *2018 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. pp. 1–6 (2018)
12. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*. p. 213–229. Springer-Verlag, Berlin, Heidelberg (2020)
13. Chang, L., Yujie, Z., Andrew, Z., Weidi, X.: Countr: Transformer-based generalised visual counting. In: *British Machine Vision Conference (BMVC)* (2022)
14. Chattopadhyay, P., Vedantam, R., Selvaraju, R.R., Batra, D., Parikh, D.: Counting everyday objects in everyday scenes. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1135–1144 (2017)
15. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *International conference on machine learning*. pp. 1597–1607. PmLR (2020)
16. Chen, X., Huo, S., Jiang, B., Hu, H., Chen, X.: Single domain generalization for few-shot counting via universal representation matching. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4639–4649 (June 2025)
17. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 2818–2829 (2023)

18. Cholakal, H., Sun, G., Khan, F.S., Shao, L.: Object counting and instance segmentation with image-level supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12397–12405 (2019)
19. Dai, S., Liu, J., Cheung, N.M.: Referring expression counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16985–16995 (June 2024)
20. Dai, Z., Wang, G., Yuan, W., Zhu, S., Tan, P.: Cluster contrast for unsupervised person re-identification. In: Proceedings of the Asian conference on computer vision. pp. 1142–1160 (2022)
21. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muenighoff, N., Lo, K., Soldaini, L., Lu, J., Anderson, T., Bransom, E., Ehsani, K., Ngo, H., Chen, Y., Patel, A., Yatskar, M., Callison-Burch, C., Head, A., Hendrix, R., Bastani, F., VanderBilt, E., Lambert, N., Chou, Y., Chheda, A., Sparks, J., Skjonsberg, S., Schmitz, M., Sarnat, A., Bischoff, B., Walsh, P., Newell, C., Wolters, P., Gupta, T., Zeng, K.H., Borchardt, J., Groeneveld, D., Nam, C., Lebrecht, S., Wittlif, C., Schoenick, C., Michel, O., Krishna, R., Weihs, L., Smith, N.A., Hajishirzi, H., Girshick, R., Farhadi, A., Kembhavi, A.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 91–104 (June 2025)
22. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
23. Dumery, C., Etté, N., Fan, A., Li, R., Xu, J., Le, H., Fua, P.: Counting stacked objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19774–19783 (2025)
24. Fu, Q., Min, W., Sheng, W., Peng, C.: Counting dense object of multiple types based on feature enhancement. *Frontiers in Neurorobotics* **18** (2024)
25. Gouthaman, K., Nambiar, A., Srinivas, K.S., Mittal, A.: Linguistically-aware attention for reducing the semantic gap in vision-language tasks. *Pattern Recognition* **112**, 107812 (2021)
26. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
27. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask r-cnn. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2980–2988 (2017)
28. Henaff, O.J., Srinivas, A., De Fauw, J., Razavi, A., Doersch, C., Eslami, S.A., van den Oord, A.: Data-efficient image recognition with contrastive predictive coding. In: International conference on machine learning. pp. 4182–4192. PMLR (2020)
29. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017)
30. Hobley, M., Prisacariu, V.: Abc easy as 123: A blind counter for exemplar-free multi-class class-agnostic counting. In: European Conference on Computer Vision. pp. 304–319. Springer (2024)
31. Huang, Y., Nguyen, D.D., Nguyen, L., Pham, C., Hoai, M.: Count what you want: exemplar identification and few-shot counting of human actions in the wild. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 10057–10065 (2024)

32. Huang, Z., Dai, M., Zhang, Y., Zhang, J., Shan, H.: Point segment and count: A generalized framework for object counting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17067–17076 (2024)
33. Hui, X., Wu, Q., Rahmani, H., Liu, J.: Class-agnostic object counting with text-to-image diffusion model. In: European Conference on Computer Vision. pp. 1–18. Springer (2024)
34. Kang, S., Moon, W., Kim, E., Heo, J.P.: Vlcounter: Text-aware visual representation for zero-shot object counting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2714–2722 (2024)
35. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. *Advances in neural information processing systems* **33**, 18661–18673 (2020)
36. Kong, D., Gray, D., Tao, H.: A viewpoint invariant approach for crowd counting. In: 18th International Conference on Pattern Recognition (ICPR’06). vol. 3, pp. 1187–1190 (2006)
37. Lempitsky, V., Zisserman, A.: Learning to count objects in images. In: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) *Advances in Neural Information Processing Systems*. vol. 23. Curran Associates, Inc. (2010)
38. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* **34**, 9694–9705 (2021)
39. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *European Conference on Computer Vision* (2024)
40. Liu, S., Zhang, P., Zhang, S., Ke, W.: Countse: Soft exemplar open-set object counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21536–21546 (October 2025)
41. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
42. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
43. Lu, E., Xie, W., Zisserman, A.: Class-agnostic counting. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) *Computer Vision – ACCV 2018*. pp. 669–684. Springer International Publishing, Cham (2019)
44. Luo, Z., Wu, Y., Cheng, T., Liu, Y., Xiao, Y., Wang, H., Zhang, X.P., Yang, Y.: Cohd: A counting-aware hierarchical decoding framework for generalized referring expression segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22685–22694 (2025)
45. Meng, H., Hong, X., Lai, Z., Shang, M.: Free lunch enhancements for multi-modal crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14013–14023 (June 2025)
46. Nguyen, E.R., Le, H., Samaras, D., Ryoo, M.: Instance-aware generalized referring expression segmentation. *arXiv preprint arXiv:2411.15087* (2024)
47. Nguyen, T., Pham, C., Nguyen, K., Hoai, M.: Few-shot object counting and detection. In: *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*. p. 348–365. Springer-Verlag, Berlin, Heidelberg (2022)

48. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal* pp. 1–31 (2024)
49. Pelhan, J., Lukežič, A., Zavrtanik, V., Kristan, M.: A novel unified architecture for low-shot counting by detection and segmentation. *Advances in Neural Information Processing Systems* **37**, 66260–66282 (2024)
50. Pelhan, J., Zavrtanik, V., Kristan, M., et al.: Dave-a detect-and-verify paradigm for low-shot counting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 23293–23302 (2024)
51. Perko, R., Klopschitz, M., Almer, A., Roth, P.M.: Critical aspects of person counting and density estimation. *Journal of Imaging* **7** (2021)
52. Pothiraj, A., Stengel-Eskin, E., Cho, J., Bansal, M.: Capture: Evaluating spatial reasoning in vision language models via occluded object counting. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 8001–8010 (2025)
53. Qiao, Y., Deng, C., Wu, Q.: Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia* **23**, 4426–4440 (2020)
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning* (2021)
55. Ranjan, V., Le, H., Hoai, M.: Iterative crowd counting. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (September 2018)
56. Ranjan, V., Sharma, U., Nguyen, T., Hoai, M.: Learning to count everything. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021)
57. Shi, Z., Sun, Y., Zhang, M.: Training-free object counting with prompts. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. pp. 323–331 (2024)
58. Su, W., Miao, P., Dou, H., Li, X.: Scanformer: Referring expression comprehension by iteratively scanning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13449–13458 (2024)
59. Su, Y., Wang, Y., Yao, L., Chau, L.P.: Few-shot class-agnostic counting with occlusion augmentation and localization. In: *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*. pp. 1–5. IEEE (2024)
60. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI* 16. pp. 776–794. Springer (2020)
61. Trott, A., Xiong, C., Socher, R.: Interpretable counting for visual question answering. In: *International Conference on Learning Representations* (2018)
62. Wan, J., Liu, Z., Chan, A.B.: A generalized loss function for crowd counting and localization. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1974–1983 (2021)
63. Wang, B., Liu, H., Samaras, D., Hoai, M.: Distribution matching for crowd counting. In: *Advances in Neural Information Processing Systems* (2020)
64. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L.: Exploring cross-image pixel contrast for semantic segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 7303–7313 (October 2021)

65. Wang, W., Yue, T., Zhang, Y., Guo, L., He, X., Wang, X., Liu, J.: Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12998–13008 (2024)
66. Wang, Y., Yang, B., Wang, X., Liang, C., Chen, J.: Satcount: A scale-aware transformer-based class-agnostic counting framework. *Neural Networks* **172**, 106126 (2024)
67. Wang, Z., Pan, Z., Peng, Z., Cheng, J., Xiao, L., Jiang, W., Cao, Z.: Exploring contextual attribute density in referring expression counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19587–19596 (June 2025)
68. Wu, C., Ji, J., Wang, H., Ma, Y., Huang, Y., Luo, G., Fei, H., Sun, X., Ji, R., et al.: Rg-san: Rule-guided spatial awareness network for end-to-end 3d referring expression segmentation. *Advances in Neural Information Processing Systems* **37**, 110972–110999 (2025)
69. Xu, J., Le, H., Samaras, D.: Zero-shot object counting with language-vision models. *arXiv preprint arXiv:2309.13097* (2023)
70. Xu, J., Le, H., Samaras, D.: Learning to count from pseudo-labeled segmentation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 8754–8763. IEEE (2025)
71. Xu, Y., Song, F., Zhang, H.: Learning spatial similarity distribution for few-shot object counting. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. IJCAI '24 (2024)
72. Yang, B., Chen, J., Ye, M.: Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11069–11079 (October 2023)
73. Yang, M., Li, Z., Zhang, J., Qi, L., Shi, Y.: Taste more, taste better: Diverse data and strong model boost semi-supervised crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24440–24451 (2025)
74. Yuan, Y., Chen, W., Yang, Y., Wang, Z.: In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 354–355 (2020)
75. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: The Eleventh International Conference on Learning Representations (2023)
76. Zhang, Q., Chan, A.B.: Calibration-free multi-view crowd counting. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX. p. 227–244. Springer-Verlag, Berlin, Heidelberg (2022)
77. Zhang, S., Zhou, Q., Ke, W.: Enhancing zero-shot object counting via text-guided local ranking and number-evoked global attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21097–21106 (2025)
78. Zhang, Y., Hare, J., Prügel-Bennett, A.: Learning to count objects in natural images for visual question answering. In: International Conference on Learning Representations (2018)