Where is my Wallet? Modeling Object Proposal Sets for Egocentric Visual Query Localization

Mengmeng Xu[1,2*] Yanghao Li[1] Cheng-Yang Fu[1] Bernard Ghanem[2] Tao Xiang[1]
Juan-Manuel Pérez-Rúa[1]
[1] Meta AI [2] KAUST, Saudi Arabia
{mengmeng.xu,bernard.ghanem}@kaust.edu.sa, {lyttonhao,chengyangfu,txiang,jmpr}@meta.com

Abstract

This paper deals with the problem of localizing objects in image and video datasets from visual exemplars. In particular, we focus on the challenging problem of egocentric visual query localization. We first identify grave implicit biases in current query-conditioned model design and visual query datasets. Then, we directly tackle such biases at both frame and object set levels. Concretely, our method solves these issues by expanding limited annotations and dynamically dropping object proposals during training. Additionally, we propose a novel transformer-based module that allows for object-proposal set context to be considered while incorporating query information. We name our module Conditioned Contextual Transformer or CocoFormer. Our experiments show the proposed adaptations improve egocentric query detection, leading to a better visual query localization system in both 2D and 3D configurations. Thus, we can improve frame-level detection performance from 26.28% to 31.26% in AP, which correspondingly improves the VQ2D and VQ3D localization scores by significant margins. Our improved context-aware query object detector ranked first and second respectively in the VQ2D and VQ3D tasks in the 2nd Ego4D challenge. In addition to this, we showcase the relevance of our proposed model in the Few-Shot Detection (FSD) task, where we also achieve SOTA results. Our code is available at https://github.com/facebookresearch/vq2d_cvpr.

1. Introduction

The task of Visual Queries Localization can be described as the question, 'when was the last time that I saw X', where X is an object query represented by a visual crop. In the Ego4D [24] setting, this task aims to retrieve objects from an 'episodic memory', supported by the recordings from an

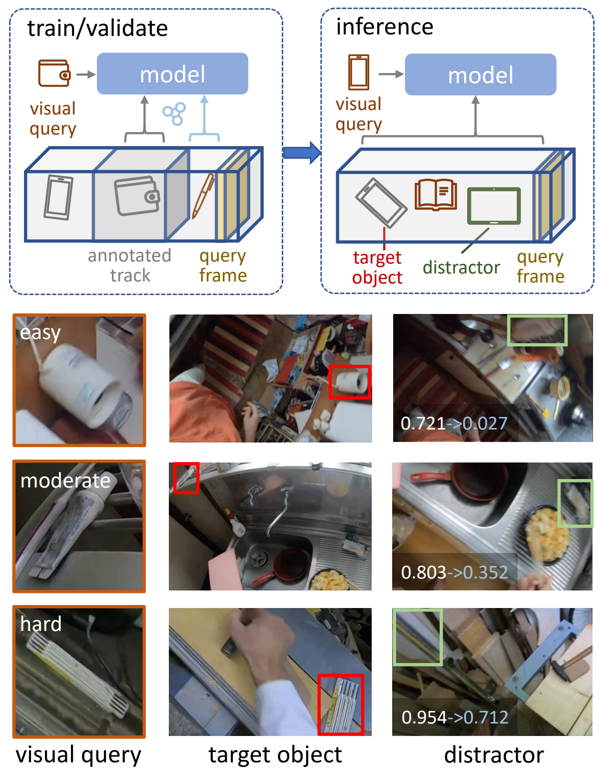

Figure 1. **Existing approaches to visual query localization are biased.** *Top*: only positive frames containing query object are used to train and validate the query object detector, while the model is naturally tested on whole videos, where the query object is mostly absent. Training with background frames alleviates confusion caused by hard negatives during inference and helps predict fewer false positives in negative frames. *Bottom*: visual query, target object, and distractors of three query objects from easy to hard. The confidence scores in white from a biased model get suppressed in our model in cyan. Our method improves the visual query localization system with fewer false positives, even for the very challenging scenario shown in the last row.

*Work done during an internship at Meta AI.

egocentric device, such as VR headsets or AR glasses. A real-world application of such functionality is localizing a user's items via a pre-registered object-centered image of them. A functional visual query localization system will allow users to find their belongings by a short re-play, or via a 3D arrow pointing to their real-world localization.

The current solutions to this problem [24, 64] rely on a so-called 'Siam-detector' that is trained on annotated response tracks but tested on whole video sequences, as shown in Fig. 1 (top, gray arrows). The Siam-detector model design allows the incorporation of query exemplars by independently comparing the query to all object proposals. During inference on a given video, the visual query is fixed, and the detector runs over all the frames in the egocentric video recording.

Although existing methods offer promising results in query object detection performance, it still suffers from *domain* and *task* biases. The *domain bias* appears due to only training with frames with well-posed objects in clear view, while daily egocentric recordings are naturally out of focus, blurry, and undergoing uncommon view angles. On the other hand, *task bias* refers to the issue of the query object always being available during training, while in reality, it is absent during most of the test time. Therefore, baseline models predict false positives on distractors, especially when the query object is out of view. These issues are exacerbated by the current models' independent scoring of each object proposal, as the baseline models learn to give high scores to superficially similar objects while disregarding other proposals to reassess those scores.

In Fig. 1 (bottom), we show the visual query, target object, and distractors for three data samples from easy to hard. The confidence scores in white of the distracting objects are reported from a publicly-available 'Siam-RCNN' model [24]. Due to random viewpoints and the large number of possible object classes that are exhibited in egocentric recordings, the target object is hard to discover and confused with high-confidence false positives. In this work, we show that these biases can be largely tackled by training a conditioned contextual detector with augmented object proposal sets. Our detector is based on a hypernetwork architecture that allows incorporating open-world visual queries, and a transformer-based design that permits context from other proposals to take part in the detection reasoning. We call this model Conditioned Contextual Transformer or CocoFormer. CocoFormer has a conditional projection layer that generates a transformation matrix from the query. This transformation is then applied to the proposal features to create query-conditioned proposal embeddings. These query-aware proposal embeddings are then fed to a set-transformer [33], which effectively allows our model to utilize the global context of the corresponding frame. We show our proposed CocoFormer surpasses Siam-RCNN and is more flexible in different applications as a hyper-network, such as multimodal query object detection and few-shot object detection (FSD).

Our CocoFormer has the increased capability to model objects for visual query localization because it tackles the domain bias by incorporating conditional context, but it still suffers from task bias induced by the current training strategy. To alleviate this, we propose to sample proposal sets from both labeled and unlabeled video frames, which collects data closer to the real distribution. Concretely, we enlarge the positive query-frame training pairs by the natural view-point change of the same object in the view, while we create negative query-frame training pairs that incorporate objects from background frames. Essentially, we sample the proposal set from those frames in a balanced manner. We collect all the objects in a frame as a set and decouple the task by two set-level questions: (1) does the query object exist? (2) what is the most similar object to the query? Since the training bias issue only exists in the first problem, we sample positive/negative *proposal sets* to reduce it. Note that this is an independent process of frame sampling, as the domain bias impairs the understanding of objects in the egocentric view, while task bias implicitly hinders the precision of visual query detection.

Overall, our experiments show diversifying the object query and proposal sets, and leveraging global scene information can evidently improve query object detection. For example, training with unlabeled frames can improve the detection AP from 27.55% to 28.74%, while optimizing the conditional detector's architecture can further push AP to 30.25%. Moreover, the visual query system can be evaluated in both 2D and 3D configurations. In VQ2D, we observe a significant improvement from baseline 0.13 to 0.18 on the leaderboard. In VQ3D, we show consistent improvement across all metrics. In the end, our improved context-aware query object detector ranked first and second respectively in the VQ2D and VQ3D tasks in the 2nd Ego4D challenge, and achieve SOTA in Few-Shot Detection (FSD) on the COCO dataset.

2. Related Work

Object detection. Deep learning models for object detection, either one-stage [5, 32, 36, 55, 77, 79], or two-stage models [7, 17, 22, 50] assume that a large amount of human-annotated data [18, 31, 37] is available. Indeed, extending them to handle novel categories associated with little data is challenging. Effectively handling test-time exemplars on top of that is even more so, and that is exactly what is required for visual query (VQ) tasks which are a form of the few-shot detection task.

Few-shot detection and visual queries. In few-shot detection (FSD), we aim to produce an object detection model that can quickly adapt to novel object categories specified by a small dataset of per-class samples. One example is [28], which encodes each novel class as a re-weighting vector that is applied to object proposal features. Similarly, [46] proposes a query-support hierarchical comparison function in a contrasting learning regime which obtains outstanding results in the COCO FSD benchmark. A related task, focusing on incremental enrollment of novel categories while preserving the performance of old ones is known as incremental few-shot detection (iFSD) [47]. Both FSD and iFSD are similar to the visual query tasks in that detection is conditioned on a small exemplar set (the query in VQ, and the few shots in FSD). All these tasks suffer from an implicit bias in testing time that causes a large amount of false positives [24, 28, 45, 47, 63, 64, 68]. This is, exemplar-based test-time annotations do not provide enough negative signals at both frame and proposal set levels. Additionally, it is naturally difficult for base class training in FSD and VQ to appropriately generalize to novel categories due to the limited data variability in both individual instances and the number of categories [68]. This results in low-accuracy predictions of the novel categories. In fact, these issues are exacerbated in egocentric visual query localization [24]. This is because egocentric video recordings organically capture users' surroundings and their objects from non-conventional points of view. This happens almost in an unintentional way, as daily-life objects are not always at the center of the user's attention they appear in more diverse locations, poses, and under varying levels of motion blur in comparison to third-view recordings. In our paper, we unify FSD and VQ tasks under a single framework, directly tackling these long-lived ailments.

Detecting objects in hyper-networks Conditional neural processes [16] (CNP) are inspired by Gaussian processes' ability to quickly adapt a function to new data in test time. In practice, CNP models learn to change the shape of a predictive function given input-annotation pairs for a target task. They are often leveraged for efficient (amortized training) few-shot learning for classification or regression tasks [16, 67]. On the other hand, hyper-networks [19], a closely-related framework, are often considered in adaptive forms of object detection like few-shot detection [14,20,28, 46, 58], continual object detection [10, 27, 47, 71], and single or multiple object tracking [1,25,34,39,56,57,65,66,69, 74–76, 78, 80, 81]. Hyper-networks predict model parameters of another model, providing an elegant mechanism for exemplar-conditioned model adaptation in [1,28,46,47,68]. Concretely, their most instrumental contributions are the specific architectural designs that are tailored for the respective downstream tasks. These are, for example, the correlation filter in [1], code generators in [47, 68], the reweighting operator in [28], and the hierarchical attention module in [46]. In our work, we embrace the hypernetwork-based design and approach the task from an object proposal set perspective. Thus, we propose to use a task-tailored set model that naturally addresses some of the biases intrinsic to visual query and few-shot detection.

Egocentric visual understanding and visual query. Although computer vision research has largely focused on third-person images and videos, and egocentric vision datasets are not as common across all major areas of research (image and video understanding, object detection, object tracking, *etc.*) [9, 12, 13, 26, 29, 30, 38, 42, 43, 52, 53, 60], recent studies on egocentric vision showcase its own major challenges and often is linked to novel computer vision tasks [8, 11, 15, 24, 35, 44, 51, 54].

One such task, visual query localization is proposed in Ego4D [24]. The idea is to have a super-human memory from an AI agent that can answer questions according to recorded visual experience. Specifically, the VQ task requires spatiotemporal localization of an object specified by a visual query (cropped image of the object). For this task, [24] proposed to use a Siam-RCNN network, which extracts visual features [23] from the query crop and region proposals [50] from the input frame. Subsequent classification based on the inner product of proposals and query features is used to discover the object that is most similar to the query. Our study shows the training protocol of Siam-RCNN is suboptimal due to multiple biases in data sampling, as well as the intrinsic lack of global context modeling that is shown by the aforementioned query-proposal classifier. In this paper, we specifically tackle these biases and demonstrate our effectiveness across several tasks.

3. Method

The visual query (VQ) task takes as input a static image of the query object, and also a video recording from an egocentric view. The expected output is the localization of the object when it is last seen in the video. Specifically, the visual query 2D localization task (VQ2D) requires the output response to be a 2D bounding box for each frame of the temporal segment where the target object last appears. Additionally, the visual query 3D localization task (VQ3D) also requires a 3D displacement vector pointing to the 3D bounding box of the target object from the current camera position. VQ applies when a user asks for the location of an object by showing one image example, such as, 'Where is this [the picture of a wallet]?'.

3.1. Overall solution

We developed a detection + localization pipeline to solve the VQ task. The process is summarized in Fig. 2. Given a visual crop v of a query object o, we detect o based on v in all the frames of the video recording, denoted as f, by a conditional detector. Thus, the detector's inputs are the visual query crop v and the video frame f, and the output is a bounding box with a confidence score (x, y, w, h, c) in

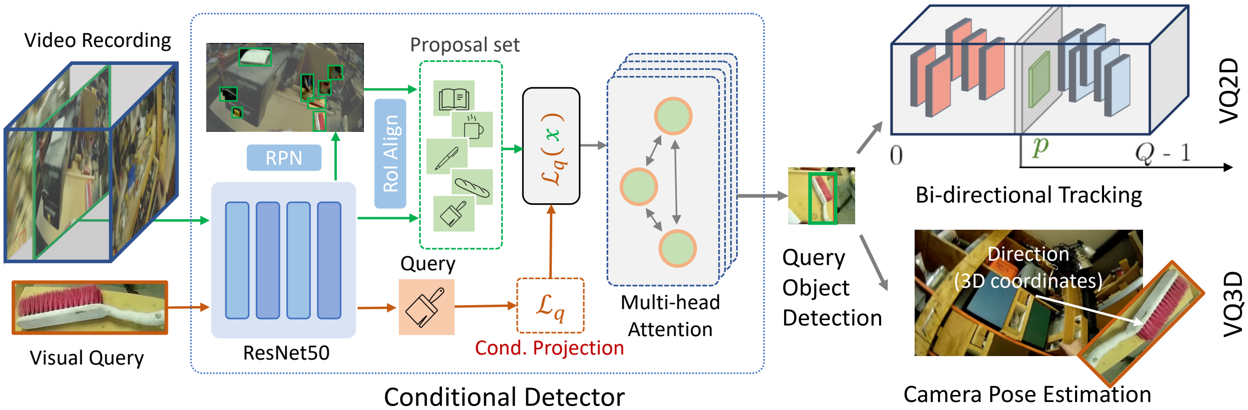

Figure 2. **Our detection + localization pipeline to solve the VQ task.** We firstly **detect** the query object in the video recording conditioned on the visual query, then apply bi-directional tracking from the detection result to **localize** the object in the video (VQ2D), or use camera pose estimation to **localize** the predicted bounding box in real-world coordinates (VQ3D). Specifically, our proposed CocoFormeris trained with our augmented query-frame pairs, and tested on all the frames of video. It has a conditional projection layer that generates unique transformation from the query, and a multi-head attention block on the query-based proposal embedding to exploit global context.

this frame. More specifically, our detector uses Region Proposal Network (RPN) with a ResNet-50 visual backbone to generate bounding box proposals $\{b_1, \cdots, b_N\}$ from the input frame, followed by an RoI-Align operation to extract bounding box features $\{\mathcal{F}(b_1), \cdots, \mathcal{F}(b_N)\}$. On the other side, the visual crop also passes through the same backbone, and generates a unique transformation from the conditional projection layer. The transformation is applied to proposal features, producing query-aware proposal embeddings for the multi-head attention layer, which predicts the target object from the proposal set.

We validate our query object detection result by VQ tasks defined in Ego4D. We strictly follow the episodic memory baseline of both tasks to localize the query object in the video and in the real-world coordinate. For VQ2D, we run a bi-directional tracker from the most recent detection peak, and predict the spatial-temporal bounding box track. For VQ3D, we apply camera pose estimation to the video frames, and output a displacement vector from the camera position to the predicted object [41].

3.2. Conditional Contextual Transformer

Existing methods for VQ2D [24] follow a simple object-query pairwise comparison strategy. A drawback of such formulation is that independent comparisons to the query limits model understanding by forcing a decision on the basis of individual similarity alone, disregarding the whole proposal set. We directly tackle such model bias by incorporating global set-level context in the form of an adapted set-transformer [33] architecture. Our transformer-based model takes the query-proposal feature pairs for the top proposals as a set, and uses self-attention to learn the interaction between the query and all the selected proposals.

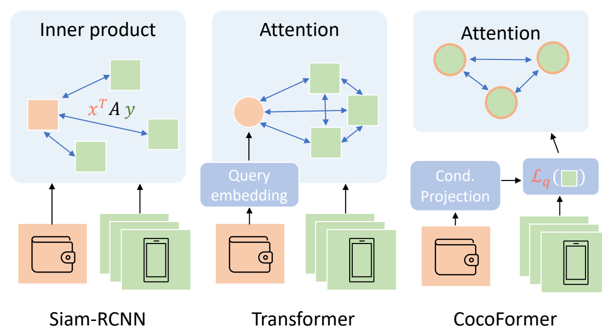

Figure 3. **We compare three model architectures for query object detection** *Left*: The independent comparison in Siam-RCNN leads to the limited capability to model global context. *Mid*: Directly using the attention model can hardly compare each proposal with the visual query. *Right*: CocoFormer applies self-attention on the query-conditioned embeddings, thus the proposal selection is based on the query, and it exploits global context.

We denominate our model Conditional Contextual Transformer or CocoFormer for short, since it provides set-level context and it is conditioned on the provided query. Concretely, our CocoFormer contains a conditional projection layer to generate a query-dependent embedding for each proposal candidate, and a self-attention block to exploit global context from the available proposal set in the camera view. The conditional projection layer $\mathcal{L}_q(x)$ generate a transformation (e.g. $\mathcal{L}_q$) based on the given condition (query) q, then the condition-dependent transformation is apply to the input x in parallel. Our transformation generator is implemented as a 3D tensor with size $(C_{out}, C_{in}, C_{cond})$, which is the dimensions of output, input, and condition, respectively. This tensor firstly operates with the query q on the last dimension, producing a $C_{out} \times C_{in}$ linear transformation. Then the transformation is applied to the input x to get the new embedding. In practise, we take the feature of visual crop as the condition, $q = \mathcal{F}(v)$, and the set of proposal feature as input $x = \{\mathcal{F}(b_1), \cdots, \mathcal{F}(b_N)\}$. Therefore, the output of this layer is the query-bias proposal embedding, i.e., $\{\mathcal{L}_{\mathcal{F}(v)}(\mathcal{F}(b_1)), \cdots, \mathcal{L}_{\mathcal{F}(v)}(\mathcal{F}(b_N))\}$. Similar to meta-train in few-shot learning, only the transformation generation process is optimized in training. A multi-head attention block is used to improve the representation of each query-bias proposal embedding through the attention mechanism several times in parallel. More concretely, key, query, and value are all set to be $\mathcal{L}_{\mathcal{F}(v)}(\mathcal{F}(b_N))$, and the feed-forward network is an FC followed by binary classification.

Fig. 3 compares our CocoFormer with baseline and transformer variants. Baseline Siam-RCNN has a pair-wise inter-production to decide if each proposal is the target object. The independent comparison of all the pairs leads to the limited capability to model global context. Alternatively, directly applying cross-attention or self-attention can enrich contextual awareness, but the model can hardly learn to compare each proposal with the visual query. CocoFormer applies self-attention on the query-conditioned embedding of each proposal candidate, thus it has a better ability to select elements based on the query and exploit global context. Please see Sect. 4.3 for ablation experiment.

3.3. Augmented balanced proposal set

As the sparsely annotated frames are only a small portion of the full Ego4d dataset, we propose to train our CocoFormer with Balanced Proposal Sets (BPS) from extended pseudo annotations on Positive Unlabeled Frames Sampling (P-UFS) and Negative Unlabeled Frames Sampling (N-UFS). We show these augmented proposal sets can minimize the domain and task bias in the training process.

Positive Unlabeled Frames Sampling. The space of all possible object instance appearances is a lot larger than existing annotations in real daily activities. To tackle this data deficiency, it becomes necessary to discover new objects in diverse viewpoints from P-UFS. Concretely, we run a vanilla pre-trained object detector on training videos, then take the resulting high-confidence detections and simulate query instances from them. Afterwards, we run a tracker in bi-direction to obtain different views of the same object, as shown in Fig. 4. This step can efficiently boost the object instances and effectively exploit the full visual information in training videos. Please see Sect. 4.4 for the significance of these extra positive query-frame pairs.

Negative Unlabeled Frames Sampling. Annotated positive frames often contain clear depictions of the object, while unlabeled objects with abnormal viewpoints, motion blur, and loss of focus are often ignored during annotation.

Figure 4. **We augment positive pairs from positive unlabeled frames sampling (P-UFS).** The camera movement in the egocentric video creates different viewpoints, changing object positions, colors, brightness *etc*.

In practice, however, these noisy samples greatly affect and confuse the detector. Thus, we also sample a number of new frames after the target object disappears from the camera view, and name the process Negative Unlabeled Frames Sampling (N-UFS). Those negatives are naturally closer to the data distribution in test time, and we can guarantee that the query object does not exist. Note that N-UFS is applied along standard cross-batch negative sampling from positive frames. It adds objects from random frames that have no annotation. The advantage of training the detector with both positive and negative frames is to be more robust when applied to all the frames in the evaluation process.

Balanced Proposal Set. A strong task bias in the VQ problem is that the training proposal set always contains the target object. This is indeed an issue since, in practice, a majority of frames do not contain the target object. This is the main reason why non-contextual models for VQ tend to predict false positives with large confidence. Clearly, without an appropriate proposal set sampling regime, baseline models will learn a shortcut to the VQ problem. This is, identifying the object that is most similar to the query irrespective of actually being the target. To alleviate this issue, we propose to pose the VQ task as two sequential questions: (1) *does the query object exist in the proposal set* (2) *what is the most similar proposal to the query object*. Since the task bias is mostly in the first question, we train the model by randomly sampling positive/negative proposal sets from a given frame. Concretely, we sample the RPN predictions to collect possible objects in the frame. The positive and negative proposal sets respectively include or exclude the object proposals that overlap with the ground-truth object. In this way, the existence of the query object is independent of the sampled frame, allowing us to effectively minimize the distribution gap between training and testing sets.

4. Experiment

4.1. Dataset, tasks, and evaluation metrics

Dataset. The VQ annotation in the Ego4D dataset includes 433 hours of videos in 54 scenarios, and 22k visual queries of $3k$ object classes with $\sim$8 frames for each query. Only train/val annotations are open to the public, and the test set performance can only be evaluated on a restricted server.

Table 1. **Our method surpasses the ego4d baseline and CVPR winner's method 2022 in the query object detection task.** The best performance is achieved when augmented training pairs and the new architecture are both applied.

method	AP	AP_{50}	AP_{75}	$AR@10$
Siam-RCNN [24]	22.51	44.68	18.24	42.3
Xu. *et al.* † [64]	26.28	49.63	23.91	46.8
Our Method	**31.26**	**57.96**	**28.88**	**47.1**
gain	(+18%)	(+17%)	(+21%)	(+1%)

†Challenge winner in CVPR-2022

Table 2. **In VQ2D, we outperform the baseline and CVPR 2022 winner's method on the validation set and got the first rank on the test set.** Notably, we improve $Succ$ to 48.37, meaning nearly half of the predicted response tracks overlap with the ground truth.

method	tAP_{25}	$stAP_{25}$	$rec\%$	$Succ$
Siam-RCNN [24]	0.20	0.12	32.2	39.8
Xu *et al.* † [64]	0.22	0.15	35.29	43.07
Our Method	**0.27**	**0.20**	**42.34**	**48.37**
gain	(+22%)	(+33%)	(+20%)	(+12%)

†Challenge winner in CVPR-2022

Table 3. **In the VQ3D challenge, we obtain significant improvement over the baseline and get the second rank on the public leaderboard.** Note that we only optimize the detector model, and leave the baseline camera estimation model fixed.

method	$L2 \downarrow$	$angle \downarrow$	$O.Succ \uparrow$	$Succ* \uparrow$
Siam-RCNN [24]	4.64	1.31	0.08	0.49
Our Method	**4.46**	**1.23**	**0.09**	**0.51**
gain	(+3.8%)	(+6.5%)	(+13%)	(+4.1%)

Tasks. We evaluate our method on three tasks: query object detection, visual query 2D localization (VQ2D), and visual query 3D localization (VQ3D). Query object detection requires a detector predicting the bounding box of the matching object in the given image, directly evaluating our detector at a frame level.VQ2D task returns a spatio-temporal location consisting of a set of bounding boxes in contingent frames. VQ3D task gives a 3D displacement vector in the real-world coordinates pointed to the physical location of the query object from the camera at the query time.

Metrics. We followed the metrics in Ego4D for each of the three tasks. We use $\{AP, AP_{50}, AP_{75}, AR@10\}$ for query object detection, $\{tAP_{25}, stAP_{25}, rec\%, Succ\}$ for VQ2D, and $\{L2, angle, O.Succ, Succ*\}$ for VQ3D.

Please see *supplementary material* for the details of metrics, implementation, and the FSD experiment setup.

4.2. Compare with SOTA

Query detection. Tab. 1 compares our method with baseline detectors. Compared to [24], exploiting P-UFS, N-UFS, and BPS improves the detector by reducing the false positive rate, and better representing query objects.

VQ2D localization We run our query detector on the target videos at FPS=5 and use the *Kys* tracker [2] to generate final response tracks. Tab. 2 compares our method with the baseline and previous winner's methods on the validation set. A similar comparison on the test set can be found on the public leaderboard. Promisingly, the success rate has been improved to 48.37, which means nearly half of the predicted response tracks overlap with the ground truth.

VQ3D localization We also follow the VQ3D framework to project the matching object into 3D world coordinates. The test set predictions are summarized in Tab. 3. We observed a consistent improvement over most of the metrics, meaning our improvement can eventually help the 3D prediction. Note that the limited improvement is due to the missing sufficient real-world projections to apply our response tracks. We believe this could be fixed by a more advanced algorithm (*e.g.* [40]) to estimate camera pose in 3D.

4.3. Advances of CocoFormer

CocoFormer is a query-conditional contextual transformer containing a conditional projection layer to transform the proposal candidates based on the query object and multi-head self-attentions to operate the transformations.

Effect of model design. We compare several possible query detection modules in Tab. 4 by applying the exact same training/validation protocol (including our proposed P-UFS, N-UFS, and BPS). The baseline model Siam-RCNN is shown in the first row of Tab. 4. It independently computes the similarity between the query and the candidate objects by the inner product. Siam-RCNN greatly benefits from our improved training strategy as can be seen by comparing to the query detection results reported in Tab. 1. We also compare against vanilla Transformer-based architectures in the second block of Tab. 4. Self-Att means we take the visual query with a learnable embedding as an extra token in a self-attention layer, while Cross-Att means we use each candidate proposal as query, and the visual query as key and value in multi-head attention. Transformer with cross-attention works slightly better than the baseline in detection AP and AP_{50}, but it underperforms in full VQ2D due to inaccurate localization in the strict setting (*e.g.* AP_{75}).The overall performance of CocoFormer is superior to the other designs. An interesting observation from Tab. 4 is that AP_{75} in query detection better correlates to VQ2D scores than other detection metrics. Our full CocoFormer provides clear gains in the strict image-level query detection setting and across the board in VQ2D.

Extension to multimodal queries. We further show that

Table 4. **CocoFormer works better than other methods in visual query localization.** *Top*: Siam-RCNN is a strong baseline with our improved training strategy. *Mid*: Transformer with cross-attention works slightly better than the baseline in detection AP and AP50, but the inaccurate localization ability (e.g. AP75) leads the suboptimal performance in VQ2D. *Bottom*: The overall performance in CocoFormer is generally better than other designs, and the conditional projection layer could localize the target object more tidily in the image, and results in better 2D localization prediction.

model	query operation	visual query detection				visual query 2D localisation			
		AP	AP_{50}	AP_{75}	$AR@10$	tAP_{25}	$stAP_{25}$	$rec\%$	$Succ$
Siam-RCNN	Inner Product	28.74	52.25	27.35	50.1	0.253	0.190	40.29	47.70
Transformer	Self-Attention	17.46	37.61	13.01	39.7	0.148	0.007	26.86	30.18
Transformer	Cross-Attention	29.17	54.87	26.59	46.1	0.212	0.140	37.93	41.51
CocoFormer	Concatenation	**31.96**	**60.07**	28.77	**48.0**	0.257	0.191	42.15	47.81
CocoFormer	Conditional Projection	31.26	57.96	**28.88**	47.1	**0.267**	**0.195**	**42.34**	**48.37**

our CocoFormer is flexible when more object information in the form of text labels is given. Specifically, we take the CLIP [49] embedding of the text prompt as "a photo of a [object]", where the 'object' is replaced by the query's title when it is available. In Tab. 5, we concatenate the text embedding to the query embedding, apply one FC layer for dimension reduction, and feed the FC output to the conditional project layers. Modeling queries with multimodal information result in consistent performance improvement.

Table 5. **Incorporating text in our visual query detector.** Our CocoFormer is flexible enough to readily accept data from extra modalities, and successfully exploit text labels.

Use Text?	Query detection		VQ2D localization		
	AP	AP50	tAP25	stAP25	Succ.
✗	31.26	57.96	0.257	0.195	48.37
✓	**32.65**	**62.64**	**0.269**	**0.201**	**49.03**

Extension to few-shot object detection We adapt our hyper-network to a few-shot object detection framework [46], which proposes hierarchical attention and meta-contrastive learning. Note that we still need their Hierarchical Attention Module to encode spatial information as our input, and apply our CocoFormer for detection. We do base-train of 1-shot, 3-shot, and 5-shot on the MS COCO dataset [37], and report the novel categories average precision in Tab. 6. With our conditional contextual transformer, we can consistently improve the state-of-the-art. Please see the complete table in our *supplementary material*.

4.4. Proposal set sampling from video

Unlabeled Frame Sampling (UFS) reduces domain gap. The model is learned from limited annotated foreground frames by default. Those frames usually have a more clear view, but this is not always true in all the frames. Fig. 6 studies the absolute improvement achieved by P-UFS with different sampling ratios. The number of positive training pairs is extended from 1.2 to 1.4, 1.7, 2.1, 2.6, and 4.2 million, and we observe that 1.7 million (43% extra) pairs give the most gain. The first block in Tab. 7 further studies the proposal sets from N-UFS on Siam-RCNN, where the query object is absent. Although proposal sets from background frames can slightly reduce the detection performance, they help localize the query object slightly better in the VQ2D localization task. Those experiments prove the unlabeled object helps to reduce the domain bias through wider data distribution. Please see *supplementary material* for more comparisons on both Siam-RCNN and CocoFormer.

Table 6. **Assessing model performance (nAP) in Few-Shot Detection.** We show 1-shot, 3-shot, and 5-shot settings on MS COCO. We don't finetune the model on the novel split (*novel ft.*) after base-train. Please see the complete table in our *supplementary material*. † means reproduced result from QA-FewDet [20].

Method	novel ft.	1-shot	3-shot	5-shot
TFA [59]	✓	3.4	6.6	8.3
CoRPN [73]	✓	4.1	-	-
FADI [4]	✓	5.7	-	-
Xiao *et al.* [62]	✓	3.2	6.7	8.1
MPSR [61] †	✓	2.3	5.2	6.7
Zhang *et al.* [72]	✓	4.4	7.2	-
Meta-DETR [70]	✓	7.5	13.5	15.4
DeFRCN [48]	✓	9.3	14.8	16.1
Fan *et al.* [14] †	✗	4.0	5.9	6.9
Meta FRCNN [21]	✗	5.0	-	-
QA-FewDet [20]	✗	5.1	8.6	9.5
FS-DETR [3]	✗	7.0	9.8	10.7
DAnA [6]	✗	11.9	14.0	14.4
hANMCL [46]	✗	12.9	14.4	14.5
ours	✗	**13.3**	**14.7**	**14.8**

Balanced Proposal Set (BPS) reduces task bias. We study the efect of BPS and N-UFS on the VQ task bias in Tab. 7. The last two rows show the seriousness of the task bias issue. BPS samples positive/negative proposal sets during training so that proposals overlapping with the ground truth are randomly removed. This simple strategy impairs the

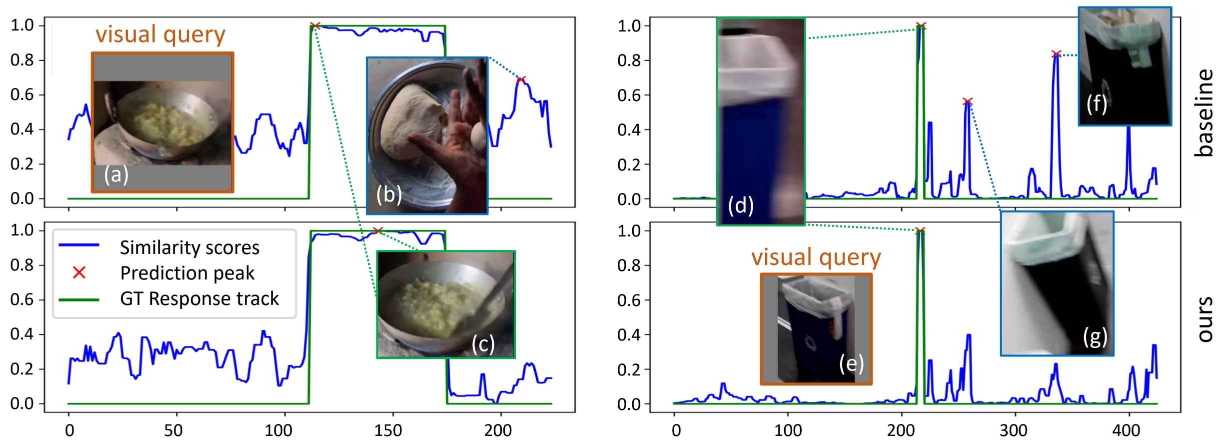

Figure 5. **Visualization of our model *vs*. baseline.** The four plots show the predictions of two videos with baseline (top) and our method (bottom). Each data point on the similarity score curve indicates the confidence (y-axis) of the top-1 detected object at a video frame (x-axis). Although both methods discover the ground truths (c) and (d), the baseline method also reports false positives (b), (f), and (g). The videos can be viewed online at frying-pan video from frame 120184, and blue-bin video from frame 1284.

Table 7. **Dataset biases exist in both frame level and object-set level.** *Top block*: Training Siam-RCNN baseline with N-UPS can help localize the query object slightly better in the video. *Bottom block*: Training our CocoFormer with a BPS can reduce this training bias, and achieve a good VQ2D performance.

N-UFS	BPS	VQ detection AP	AP_{50}	VQ 2D localization tAP_{25}	$stAP_{25}$	$Succ$
✗	✗	26.99	51.12	0.21	0.14	42.45
✓	✗	26.28	49.63	0.22	0.15	43.01
✓	✗	33.28	62.51	0.22	0.15	42.98
✓	✓	31.26	57.96	0.27	0.20	48.37

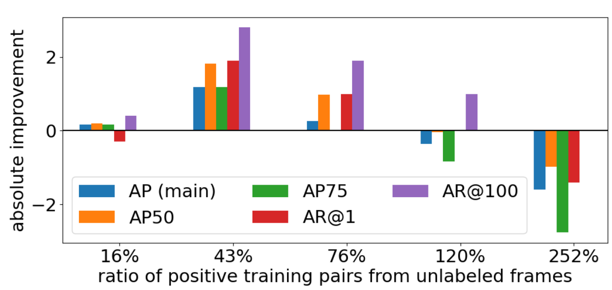

Figure 6. **We extend the positive training pairs and record the absolute improvement on different metrics.** The number of pairs gradually rises from 1.2 million up to 4.2 million. We observe that 1.7 million (43% extra) pairs give the most gains. It is also interesting to see AP75 drops faster when the pseudo pairs increase, as it requires more precise bounding-box predictions.

visual query detector performance (measured on positives frames only) as the proposal set does not always contain positives anymore, but it can evidently reduce VQ task bias, and achieve a much improved VQ2D performance.

5. Qualitative Results

We visualize the predictions of two exemplar videos in Fig. 5. In the frying-pan video, both methods find the query object in different views, but the baseline model also reports a high similarity score to the distractor, as it is also a metal container with food. The blue-bin video is more challenging because the query image crop was captured in a low-lighting condition. Therefore, the black bin detected by the baseline has a more similar visual appearance than the ground truth. Our method is able to model the global context and reports the black bin as negative.

6. Conclusion

We start by tackling Ego4D dataset and task biases of VQ2D, then propose to expand limited annotations and dynamically drop object proposals during training. Moreover, we proposed CocoFormer, a novel transformer-based module that allows for object-proposal set context to be considered while incorporating query information. Our experiments show the proposed adaptations improve egocentric query detection, leading to a better visual query localization system in both 2D and 3D configurations. CocoFormerranked first and second respectively in the VQ2D and VQ3D tasks in the 2nd Ego4D challenge, and achieve SOTA performance in the Few-Shot Detection task.

Acknowledgments. MX and BG are supported by the KAUST Office of Sponsored Research (OSR) through the VCC funding and SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

References

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 3

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *European Conference on Computer Vision*, pages 205–221. Springer, 2020. 6

[3] Adrian Bulat, Ricardo Guerrero, Brais Martinez, and Georgios Tzimiropoulos. Fs-detr: Few-shot detection transformer with prompting and without re-training. *arXiv preprint arXiv:2210.04845*, 2022. 7

[4] Yuhang Cao, Jiaqi Wang, Ying Jin, Tong Wu, Kai Chen, Ziwei Liu, and Dahua Lin. Few-shot object detection via association and discrimination. *Advances in Neural Information Processing Systems*, 34:16570–16581, 2021. 7

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[6] Tung-I Chen, Yueh-Cheng Liu, Hung-Ting Su, Yu-Cheng Chang, Yu-Hsiang Lin, Jia-Fong Yeh, Wen-Chin Chen, and Winston Hsu. Dual-awareness attention for few-shot object detection. *IEEE Transactions on Multimedia*, 2021. 7

[7] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NeurIPS*, 2016. 2

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[10] Keval Doshi and Yasin Yilmaz. Continual learning for anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 254–255, 2020. 3

[11] Peter F Edemekong, Deb L Bomgaars, and Shoshana B Levy. Activities of daily living (adls). 2017. 3

[12] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 3

[13] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 3

[14] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020. 3, 7

[15] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012. 3

[16] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2018. 3

[17] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 2

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 2

[19] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 3

[20] Guangxing Han, Yicheng He, Shiyuan Huang, Jiawei Ma, and Shih-Fu Chang. Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3263–3272, 2021. 3, 7

[21] Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 780–789, 2022. 7

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *TPAMI*, 37(9):1904–1916, 2015. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[24] Ego4D Consortium 2020. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *CVPR*, 2022. 1, 2, 3, 4, 6

[25] Weiming Hu, Qiang Wang, Li Zhang, Luca Bertinetto, and Philip HS Torr. Siammask: A framework for fast online object tracking and segmentation. *T-PAMI*, 2022. 3

[26] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2019. 3

[27] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021. 3

[28] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, 2019. 3

[29] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[30] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, Nov 2016. 3

[31] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *IJCV*, 128(7):1956–1981, 2020. 2

[32] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 2

[33] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*. PMLR, 2019. 2, 4

[34] Peixia Li, Boyu Chen, Wanli Ouyang, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Gradnet: Gradient-guided network for visual object tracking. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 6162–6171, 2019. 3

[35] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018. 3

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*. Springer, 2014. 2, 7

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3

[39] Fan Ma, Mike Zheng Shou, Linchao Zhu, Haoqi Fan, Yilei Xu, Yi Yang, and Zhicheng Yan. Unified transformer tracker for object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8781–8790, 2022. 3

[40] Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Localizing objects in 3d from egocentric videos with visual queries, 2022. 6

[41] Jinjie Mai, Chen Zhao, Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Estimating more camera poses for ego-centric videos is essential for vq3d, 2022. 4

[42] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 3

[43] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018. 3

[44] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *CVPR*, 2020. 3

[45] Alejandro Pardo, Mengmeng Xu, Ali Thabet, Pablo Arbelaez, and Bernard Ghanem. Baod: Budget-aware object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1247–1256, June 2021. 3

[46] Dongwoo Park and Jongmin Lee. Hierarchical attention network for few-shot object detection via meta-contrastive learning. *arXiv preprint arXiv:2208.07039*, 2022. 3, 7

[47] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *CVPR*, 2020. 3

[48] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8681–8690, 2021. 7

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 7

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 3

[51] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 3

[52] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5026–5035, June 2022. 3

[53] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. In *CRCV-TR-12-01*, 2012. 3

[54] Yu-Chuan Su and Kristen Grauman. Detecting engagement in egocentric video. In *European Conference on Computer Vision*, pages 454–471. Springer, 2016. 3

[55] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *CVPR*, 2019. 2

[56] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10860–10869, 2021. 3

[57] Guangting Wang, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. Spm-tracker: Series-parallel matching for real-time visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3643–3652, 2019. 3

[58] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object

detection. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9919–9928, 2020. 3

[59] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *International Conference on Machine Learning (ICML)*, July 2020. 7

[60] Haoqian Wu, Keyu Chen, Haozhe Liu, Mingchen Zhuge, Bing Li, Ruizhi Qiao, Xiujun Shu, Bei Gan, Liangsheng Xu, Bo Ren, Mengmeng Xu, Wentian Zhang, Raghavendra Ramachandra, Chia-Wen Lin, and Bernard Ghanem. Newsnet: A novel dataset for hierarchical temporal segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[61] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *European conference on computer vision*, pages 456–472. Springer, 2020. 7

[62] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *ECCV*, 2020. 7

[63] Mengmeng Xu, Yancheng Bai, and Bernard Ghanem. Missing labels in object detection. In *CVPR workshops*, volume 3, page 5, 2019. 3

[64] Mengmeng Xu, Cheng-Yang Fu, Yanghao Li, Bernard Ghanem, Juan-Manuel Perez-Rua, and Tao Xiang. Negative frames matter in egocentric visual query 2d localization. *arXiv preprint arXiv:2208.01949*, 2022. 2, 3, 6

[65] Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, and Josef Kittler. Joint group feature selection and discriminative filter learning for robust visual object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7950–7960, 2019. 3

[66] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. *arXiv preprint arXiv:2207.07078*, 2022. 3

[67] Zesheng Ye and Lina Yao. Contrastive conditional neural processes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9687–9696, 2022. 3

[68] Li Yin, JM Perez, and Kevin J Liang. Sylph: A hypernetwork framework for incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9035–9045, 2022. 3

[69] Yuechen Yu, Yilei Xiong, Weilin Huang, and Matthew R Scott. Deformable siamese attention networks for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6728–6737, 2020. 3

[70] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. Meta-detr: Few-shot object detection via unified image-level meta-learning. *arXiv preprint arXiv:2103.11731*, 2(6), 2021. 7

[71] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1131–1140, 2020. 3

[72] Weilin Zhang and Yu-Xiong Wang. Hallucination improves few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13008–13017, 2021. 7

[73] Weilin Zhang, Yu-Xiong Wang, and David A Forsyth. Cooperating rpn's improve few-shot object detection. *arXiv preprint arXiv:2011.10142*, 2020. 7

[74] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Robust multi-object tracking by marginal inference. *arXiv preprint arXiv:2208.03727*, 2022. 3

[75] Zelin Zhao, Ze Wu, Yueqing Zhuang, Boxun Li, and Jiaya Jia. Tracking objects as pixel-wise distributions. *arXiv preprint arXiv:2207.05518*, 2022. 3

[76] Linyu Zheng, Ming Tang, Yingying Chen, Guibo Zhu, Jinqiao Wang, and Hanqing Lu. Improving multiple object tracking with single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2453–2462, 2021. 3

[77] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2

[78] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8771–8780, 2022. 3

[79] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, 2019. 2

[80] Ji Zhu, Hua Yang, Nian Liu, Minyoung Kim, Wenjun Zhang, and Ming-Hsuan Yang. Online multi-object tracking with dual matching attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 366–382, 2018. 3

[81] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018. 3